SIAM J. APPL. MATH.
Vol. 65, No. 1, pp. 1–23

EPIDEMIOLOGICAL MODELS FOR MUTATING PATHOGENS*

JIA LI†, YICAN ZHOU‡, ZHIEN MA‡, AND JAMES M. HYMAN§

Abstract. We formulate epidemiological models for the transmission of a pathogen that can mutate in the host to create a second infectious mutant strain. The models account for mutation rates that depend on how long the host has been infected. We derive explicit formulas for the reproductive number of the epidemic based on the local stability of the infection-free equilibrium. We analyze the existence and stability of the boundary equilibrium, whose infection components are zero and positive, respectively, and the endemic equilibrium, whose components are all positive. We establish the conditions for global stability of the infection-free and boundary equilibria and local stability of the endemic equilibrium for the case where there is no age structure for the pathogen in the infected population. We show that under certain circumstances, there is a Hopf bifurcation where the endemic equilibrium loses its stability, and periodic solutions appear. We provide examples and numerical simulations to illustrate the Hopf bifurcation.

Key words. epidemic model, pathogen, mutation, infection age, reproductive number, global stability, Hopf bifurcation

AMS subject classifications. 34C32, 34D20, 34D23, 35F99, 92D30

DOI. 10.1137/S0036139903430185

1. Introduction.

One of the biggest challenges in preventing the spread of infectious diseases is the genetic variations of pathogens. Pathogen mutations that circumvent the protective effects of a patient's immune response are common in infectious diseases such as measles [5], hepatitis B [20], HIV [9], West Nile virus [8], and influenza [18, 23, 24, 25].

The generation or selection of mutants that are a reflection of attempts of the pathogen to resist immune attacks of the host and to survive may occur naturally or in response to treatment with antibodies or antiviral drugs. Pathogens frequently alter their antigen expression to escape the immune defense and ensure the persistent infection in a host [10, 19].

There were only a few existing mathematical models accounting for genetic mutations of a pathogen [2, 3, 11, 17, 21], and little has been done to directly model dynamics of mutations which describe the attempts of the pathogen, after its infection in a host, to escape the immune defense of the host. In this paper, we propose an infection-age-structured dynamic model for a pathogen that can mutate into a second infectious strain in the host. The mutation could be the effect of selective immunologic pressure or possibly adaptation to a more efficiently transmitted or a better replicating pathogen resulting from conversion of the original viral pathogen.

The model formulation for the origin of the pathogen strain is based on a susceptible-infective-recovered (SIR) model with variable infection ages and is governed by partial differential equations (PDEs). The dynamics of the mutant are based on

*Received by the editors June 13, 2003; accepted for publication (in revised form) March 10, 2004; published electronically September 24, 2004. This work was partially supported by Department of Energy contract W-7405-ENG-36 and Applied Mathematical Sciences Program KC-07-01-01.
http://www.siam.org/journals/siap/65-1/43018.html

†Department of Mathematical Sciences, University of Alabama, Huntsville, AL 35899 (li@math.uah.edu).

‡Department of Applied Mathematics, Xian Jiaotong University, Xian, China (zhouyc@mail.xjtu.edu.cn, zhma@xjtu.edu.cn).

§Energy Program Theoretical Division, MS-B284, Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545 (hyman@lanl.gov).

an ordinary differential equation (ODE) SIR model. We characterize the threshold conditions of the model epidemic with an explicit formula for the reproductive number of infection, which determines the stability of the infection-free equilibrium. We analyze the stability of boundary equilibria of the model, where some, but not all, of the infection components are zero. We then investigate the existence and stability of the endemic equilibrium, whose components are all positive. We obtain explicit formulas for the endemic equilibrium and the characteristic equation of this equilibrium, which determines its stability. We then consider the special case where the rate at which the pathogen converts to its mutant and the transmission rate of the original pathogen are both independent of infection age. In this simplified situation, the model equations reduce to a system of ODEs. We obtain global stability of the infection-free equilibrium and a unique boundary equilibrium. We show that under certain conditions, the unique endemic equilibrium may undergo a Hopf bifurcation resulting in a periodic solution. We provide examples and numerical simulations to illustrate the stability change of the endemic equilibrium and the Hopf bifurcation.

2. Model formulation. We base our SIR model on the spread of a pathogen that can mutate in the host to create a second, cocirculating, mutant strain. We assume that after a certain period of infection, the original strain, referred to as Strain 1, is selected against in the intrahost selection process and is converted to a mutant, referred to as Strain 2, such that a proportion of the individuals infected by Strain 1 are then carrying Strain 2. Let $S(t)$ be the susceptibles and $i(t,\tau)$ the distribution of infectives infected by Strain 1 with infection stage, or time since infection, τ, such that $\int_{\tau_1}^{\tau_2} i(t,\tau)d\tau$ is the total number of infectives with infection ages between τ_1 and τ_2 [1, 7, 13, 14, 22]. Let $J(t)$ be the infectives infected by Strain 2 and $R(t)$ the group of individuals who are recovered and immune to both strains. We further assume that the genetic difference between the two strains, or the drift of the mutation, is relatively small so that there is perfect cross-immunity; that is, once an individual is recovered from infection by one of the two strains, the individual is immune to both strains.

The dynamics of the transmission in this model are governed by the system

$$
\begin{aligned}
\frac{dS(t)}{dt} &= \mu(S^0 - S(t)) - \left(\int_0^\infty \beta_1(\tau) i(t,\tau)d\tau + \beta_2 J(t)\right) S(t),\\
\frac{\partial i(t,\tau)}{\partial t} + \frac{\partial i(t,\tau)}{\partial \tau} &= -(\mu+\gamma_1) i(t,\tau) - \kappa(\tau) i(t,\tau),\\
i(t,0) &= S(t)\int_0^\infty \beta_1(\tau) i(t,\tau)d\tau,\\
i(0,\tau) &= \psi(\tau),\\
\frac{dJ(t)}{dt} &= \beta_2 J(t) S(t) - (\mu+\gamma_2) J(t) + \int_0^\infty \kappa(\tau) i(t,\tau)d\tau,\\
\frac{dR(t)}{dt} &= \gamma_1 \int_0^\infty i(t,\tau)d\tau + \gamma_2 J(t) - \mu R(t),
\end{aligned}
\tag{2.1}
$$

where μS^0 is the input flow into the susceptible population, μ is the total removal rate which accounts for both natural death and people moving in and out of the susceptible population, γ_1 and γ_2 are the recovery rates from the infection, $\beta_1(\tau)$ and β_2 are the transmission rates of Strain 1 and Strain 2, respectively, $\kappa(\tau)$ is the mutation rate, or the rate at which Strain 1 is converted to Strain 2, and $\psi(\tau)$ is the initial distribution of infectives infected by Strain 1.

3. Thresholds of the epidemic. Assume that the initial distribution of the infectives is zero. Then $E_0 := (S^0, 0, 0)$ is the infection-free equilibrium. As is well known, its stability determines the thresholds of the epidemic [6, 7, 13, 15, 16]. We investigate the local stability of E_0 as follows.

Since the dynamics of $R(t)$ do not affect the evolution of S, i, and J, we omit the equation for $R(t)$ when studying the growth of the epidemic. Linearizing system (2.1) about E_0, by defining the perturbation variables $x(t) = S(t) - S^0$, $y(t,\tau) = i(t,\tau)$, $z(t) = J(t)$, we obtain the system

$$
\begin{aligned}
&\frac{dx(t)}{dt} = -\mu x(t) - \left(\int_0^\infty \beta_1(\tau) y(t,\tau)\, d\tau + \beta_2 z(t)\right) S^0, \\
&\begin{cases} \dfrac{\partial y(t,\tau)}{\partial t} + \dfrac{\partial y(t,\tau)}{\partial \tau} &= -(\mu+\gamma_1) y(t,\tau) - \kappa(\tau) y(t,\tau), \\ y(t,0) &= S^0 \displaystyle\int_0^\infty \beta_1(\tau) y(t,\tau)\, d\tau, \end{cases} \qquad (3.1) \\
&\frac{dz(t)}{dt} = \beta_2 z(t) S^0 - (\mu+\gamma_2) z(t) + \int_0^\infty \kappa(\tau) y(t,\tau)\, d\tau.
\end{aligned}
$$

Let $x(t) = x_0 e^{\rho t}$, $y(t,\tau) = p(\tau) e^{\rho(t-\tau)}$, and $z(t) = z_0 e^{\rho t}$, where x_0, $p(\tau)$, z_0, and ρ are to be determined. Substituting them into (3.1), we obtain the equations

$$\rho x_0 = -\mu x_0 - S^0 \int_0^\infty \beta_1(\tau) p(\tau) e^{-\rho\tau}\, d\tau - \beta_2 S^0 z_0, \tag{3.2}$$

$$\frac{dp(\tau)}{d\tau} = -(\mu+\gamma_1) p(\tau) - \kappa(\tau) p(\tau), \tag{3.3}$$

$$p(0) = S^0 \int_0^\infty \beta_1(\tau) p(\tau) e^{-\rho\tau}\, d\tau, \tag{3.4}$$

$$\rho z_0 = (\beta_2 S^0 - \mu - \gamma_2) z_0 + \int_0^\infty \kappa(\tau) p(\tau) e^{-\rho\tau}\, d\tau \tag{3.5}$$

for $p(\tau) \not\equiv 0$, $x_0 \neq 0$, $z_0 \neq 0$, and ρ.

Equations (3.3) and (3.4) are decoupled from (3.2) and (3.5). Integrating (3.3) from 0 to τ gives

$$p(\tau) = p(0) e^{-(\mu+\gamma_1)\tau - \Delta(\tau)}, \tag{3.6}$$

where $\Delta(\tau) := \int_0^\tau \kappa(v)\, dv$. Substituting (3.6) into (3.4) yields the characteristic equation

$$p(0) = S^0 p(0) \int_0^\infty \beta_1(\tau) e^{-\rho\tau} e^{-(\mu+\gamma_1)\tau - \Delta(\tau)}\, d\tau. \tag{3.7}$$

Defining

$$C(\rho) = S^0 \int_0^\infty \beta_1(\tau) e^{-\rho\tau} e^{-(\mu+\gamma_1)\tau - \Delta(\tau)}\, d\tau,$$

we note that (3.7) has a nonzero solution $p(0)$ if and only if there exists ρ such that $C(\rho) = 1$.

We first consider the case where ρ is a real number. Since

$$C'(\rho) = -S^0 \int_0^\infty \tau \beta_1(\tau) e^{-\rho\tau} e^{-(\mu+\gamma_1)\tau - \Delta(\tau)}\, d\tau < 0,$$

$C(\rho)$ is a decreasing function of ρ. Noticing $\lim_{\rho\to-\infty} C(\rho) = \infty$ and $\lim_{\rho\to\infty} C(\rho) = 0$, if we define the number

$$R_1 := C(0) = S^0 \int_0^\infty \beta_1(\tau) e^{-(\mu+\gamma_1)\tau - \Delta(\tau)} d\tau, \tag{3.8}$$

then there exists a unique real solution ρ to the equation $C(\rho) = 1$, which is negative if $R_1 < 1$ and positive if $R_1 > 1$.

If $\rho := \rho_1 + i\rho_2$ is a complex number, where $i = \sqrt{-1}$, then by separating the real and imaginary parts of $C(\rho) = 1$, the real part ρ_1 satisfies

$$1 = S^0 \int_0^\infty \beta_1(\tau) e^{-\rho_1\tau} e^{-(\mu+\gamma_1)\tau - \Delta(\tau)} \cos(\rho_2\tau) d\tau. \tag{3.9}$$

However, since

$$S^0 \int_0^\infty \beta_1(\tau) e^{-\rho_1\tau} e^{-(\mu+\gamma_1)\tau - \Delta(\tau)} \cos(\rho_2\tau) d\tau \le S^0 \int_0^\infty \beta_1(\tau) e^{-\rho_1\tau} e^{-(\mu+\gamma_1)\tau - \Delta(\tau)} d\tau,$$

solution ρ_1 to (3.9) must be negative if $R_1 < 1$. That is, equation $C(\rho) = 1$ can have solutions with negative real part only if $R_1 < 1$.

The solution ρ of $C(\rho) = 1$ can be used to determine $p(\tau)$. The initial values, x_0 and z_0, can now be defined from (3.2) and (3.5). The number R_1 defined in (3.8) is a threshold value for Strain 1 because if $R_1 > 1$ the epidemic for Strain 1 grows, while if $R_1 < 1$ it delays. It is also the number of secondary infective cases generated by infection of Strain 1. We refer to R_1 as the reproductive number for Strain 1.

If initially no one is infected with Strain 1, i.e., $i(t, \tau) = 0$, then $p(\tau) = 0$ for all τ. Equations (3.2) and (3.5) can be reduced to

$$\begin{aligned} \rho x_0 &= -\mu x_0 - \beta_2 S^0 z_0, \\ \rho z_0 &= (\beta_2 S^0 - \mu - \gamma_2) z_0, \end{aligned} \tag{3.10}$$

and they determine threshold conditions for Strain 2. Define

$$R_2 := \frac{\beta_2 S^0}{\mu + \gamma_2}. \tag{3.11}$$

All solutions ρ of system (3.10) are negative if and only if $R_2 < 1$. Therefore, R_2 is a threshold value for Strain 2 and is the number of secondary infective cases generated by infection of Strain 2. We refer to R_2 as the reproductive number of Strain 2.

The thresholds for the epidemic can be summarized as follows.

THEOREM 3.1. *Define the reproductive number, R_0, of infection in the total population by*

$$R_0 := \max\{R_1, R_2\},$$

that is,

$$R_0 = \max\left\{ S^0 \int_0^\infty \beta_1(\tau) e^{-(\mu+\gamma_1)\tau - \Delta(\tau)} d\tau, \quad \frac{\beta_2 S^0}{\mu + \gamma_2} \right\}.$$

Then the infection-free equilibrium E_0 is asymptotically stable if $R_0 < 1$ and is unstable if $R_0 > 1$.

4. Boundary equilibrium. Cocirculating strains of the pathogen compete with each other to infect the susceptible population. When only one strain is present, the solution is on the boundary of the feasibility solution space and we call the stationary solution a boundary equilibrium.

An equilibrium of system (2.1), $(S, i(\tau), J)$, satisfies the system

$$\mu(S^0 - S) - \left(\int_0^\infty \beta_1(\tau) i(\tau) d\tau + \beta_2 J\right) S = 0, \tag{4.1a}$$

$$\frac{di(\tau)}{d\tau} = -(\mu + \gamma_1) i(\tau) - \kappa(\tau) i(\tau), \tag{4.1b}$$

$$i(0) = S \int_0^\infty \beta_1(\tau) i(\tau) d\tau, \tag{4.1c}$$

$$\beta_2 JS - (\mu + \gamma_2) J + \int_0^\infty \kappa(\tau) i(\tau) d\tau = 0. \tag{4.1d}$$

It follows from (4.1d) that if $J = 0$, then $i(\tau) = 0$ for all τ. That is, there does not exist a boundary equilibrium with $i(\tau) \geq 0$ and $J = 0$, and the only boundary equilibrium has $i(\tau) = 0$ for all τ and $J \neq 0$. We denote it as $E_1 := (S_1, i_1(\tau), J_1)$.

Solving (4.1a) and (4.1d), we have

$$S_1 = \frac{\mu + \gamma_2}{\beta_2}, \quad J_1 = \frac{\mu}{\beta_2}\left(\frac{S^0 \beta_2}{\mu + \gamma_2} - 1\right) = \frac{\mu}{\beta_2}(R_2 - 1). \tag{4.2}$$

Thus the boundary equilibrium E_1 exists if and only if $R_2 > 1$.

To study stability of this boundary equilibrium, we linearize system (2.1) about E_1 by letting $x(t) = S(t) - S_1$, $y(t) = J(t) - J_1$, $z(t, \tau) = i(t, \tau)$, and we obtain the system

$$\begin{aligned}
&\frac{dx(t)}{dt} = -\mu x(t) - \beta_2 J_1 x(t) - \beta_2 S_1 y(t) - S_1 \int_0^\infty \beta_1(\tau) z(t, \tau) d\tau, \\
&\frac{dy(t)}{dt} = \beta_2 J_1 x(t) - (\mu + \gamma_2) y(t) + \beta_2 S_1 y(t) + \int_0^\infty \kappa(\tau) z(t, \tau) d\tau, \\
&\begin{cases} \dfrac{\partial z(t, \tau)}{\partial t} + \dfrac{\partial z(t, \tau)}{\partial \tau} = -(\mu + \gamma_1) z(t, \tau) - \kappa(\tau) z(t, \tau), \\ z(t, 0) = S_1 \displaystyle\int_0^\infty \beta_1(\tau) z(t, \tau) d\tau. \end{cases}
\end{aligned} \tag{4.3}$$

Using the same approach as in section 3, we first derive the characteristic equation for E_1,

$$1 = S_1 \int_0^\infty \beta_1(\tau) e^{-\rho\tau} e^{-(\mu + \gamma_1)\tau - \Delta(\tau)} d\tau, \tag{4.4}$$

and define

$$R_b := S_1 \int_0^\infty \beta_1(\tau) e^{-(\mu + \gamma_1)\tau - \Delta(\tau)} d\tau.$$

If $R_b < 1$, then $\lim_{t \to \infty} z(t, \tau) = 0$.

Next we locate the eigenvalues of the following matrix from system (4.3):

$$\begin{bmatrix} -\mu - \beta_2 J_1 & -\beta_2 S_1 \\ \beta_2 J_1 & -(\mu + \gamma_2 - \beta_2 S_1) \end{bmatrix} = \begin{bmatrix} -\mu - \beta_2 J_1 & -\beta_2 S_1 \\ \beta_2 J_1 & 0 \end{bmatrix}.$$

The trace and determinant of this matrix are negative and positive, respectively. Therefore, its eigenvalues both have negative real part.

In summary we have the following.

THEOREM 4.1. *The unique boundary equilibrium*

$$E_1 = (S_1, i_1(\tau), J_1) = \left(\frac{\mu + \gamma_2}{\beta_2},\ 0,\ \frac{\mu}{\beta_2} \left(\frac{S^0 \beta_2}{\mu + \gamma_2} - 1 \right) \right)$$

exists if and only if $R_2 > 1$. *It is locally asymptotically stable if*

$$R_b = \frac{\mu + \gamma_2}{\beta_2} \int_0^\infty \beta_1(\tau) e^{-(\mu+\gamma_1)\tau - \Delta(\tau)} d\tau < 1$$

and is unstable if $R_b > 1$.

If $R_2 > 1$, then $S^0 > (\mu + \gamma_2)/\beta_2 := \tilde{S}_1$. Notice that R_b can be rewritten as $R_b = \tilde{S}_1/S^0 R_1 = R_1/R_2$. When the boundary equilibrium E_1 exists, $R_2 > 1$, and hence $S^0 > \tilde{S}_1$ and $R_b < R_1$. If $R_2 > 1 > R_1$, then $R_b < 1$, which implies that the boundary equilibrium E_1 is asymptotically stable. In the situation where $R_2 > 1$ and $R_1 > 1$, the infection-free equilibrium is unstable and the two strains cannot both die out. If $R_2 > R_1 > 1$, then $R_b < 1$ and the boundary equilibrium E_1 exists and is asymptotically stable. In the last possible case, if $R_1 > R_2 > 1$, then although the boundary equilibrium E_1 exists, it is unstable. This situation may lead to the existence and stability of an endemic equilibrium or other dynamical features of system (2.1).

5. Endemic equilibrium. The cocirculating strains of the pathogen can coexist. The stationary coexistence solution is an endemic equilibrium whose components are all positive.

5.1. Existence of the endemic equilibrium. Let $E^* := (S^*, i^*(\tau), J^*)$ be an endemic equilibrium of system (2.1). It follows from (4.1b) that

$$i^*(\tau) = i^*(0) e^{-(\mu+\gamma_1)\tau - \Delta(\tau)}.$$

By substituting this into (4.1c), we arrive at the equation

$$i^*(0) = i^*(0) S^* \int_0^\infty \beta_1(\tau) e^{-(\mu+\gamma_1)\tau - \Delta(\tau)} d\tau = i^*(0) \frac{S^* R_1}{S^0}. \tag{5.1}$$

Equation (5.1) has a solution $i^*(0) > 0$ if and only if

$$S^* = \frac{S^0}{R_1}. \tag{5.2}$$

It follows from (4.1c) that

$$i^*(0) = S^* W_1,$$

where we define $W_1 := \int_0^\infty \beta_1(\tau) i^*(\tau) d\tau$. Then

$$i^*(\tau) = S^* W_1 e^{-(\mu+\gamma_1)\tau - \Delta(\tau)}. \tag{5.3}$$

Define

$$W_2 := \int_0^\infty \kappa(\tau) i^*(\tau) d\tau = S^* W_1 \int_0^\infty \kappa(\tau) e^{-(\mu+\gamma_1)\tau - \Delta(\tau)} d\tau = S^* W_1 K, \tag{5.4}$$

where

$$K := \int_0^\infty \kappa(\tau) e^{-(\mu+\gamma_1)\tau - \Delta(\tau)} d\tau.$$

The equilibrium equations (4.1a) and (4.1d) can be expressed as

$$\mu S^0 = (\mu + W_1 + \beta_2 J^*)\, S^*, \tag{5.5}$$

$$W_2 = ((\mu + \gamma_2) - \beta_2 S^*)\, J^*. \tag{5.6}$$

Substituting (5.2) into (5.6) yields

$$(\mu + \gamma_2)\left(1 - \frac{\beta_2 S^*}{\mu + \gamma_2}\right) J^* = (\mu + \gamma_2)\left(1 - \frac{R_2}{R_1}\right) J^* = W_2. \tag{5.7}$$

Since $W_2 > 0$, there exists a positive solution J^* of (5.7) if and only if

$$\frac{R_2}{R_1} < 1.$$

Suppose $R_2 < R_1$. Then solving (5.7) for J^* yields

$$J^* = \frac{W_2}{(\mu + \gamma_2)\left(1 - \frac{R_2}{R_1}\right)}. \tag{5.8}$$

Substituting (5.8) into (5.5) gives

$$\mu + W_1 + \beta_2 \frac{W_2}{(\mu + \gamma_2)\left(1 - \frac{R_2}{R_1}\right)} = \frac{\mu S^0}{S^*} = \mu R_1. \tag{5.9}$$

We then substitute (5.4) into (5.9) to obtain

$$W_1 + \frac{\beta_2 S^0 K}{(\mu + \gamma_2)(R_1 - R_2)} W_1 = \mu(R_1 - 1), \tag{5.10}$$

which implies that $W_1 > 0$ if $R_1 > 1$.

Solving (5.10) for W_1 yields

$$W_1 = \frac{\mu\,(R_1 - 1)\,(R_1 - R_2)\,(\mu + \gamma_2)}{((\mu + \gamma_2)\,(R_1 - R_2) + \beta_2 K S^0)}. \tag{5.11}$$

W_2 can be determined by substituting (5.11) into (5.4). Finally, substituting W_2 and W_1 into (5.3) and (5.8), we obtain the expression for the unique positive endemic equilibrium.

THEOREM 5.1. *If $R_1 > 1$ and $R_1 > R_2$, then there exists a unique endemic equilibrium $E^* = (S^*, i^*(\tau), J^*)$ given by*

$$S^* = \frac{S^0}{R_1}, \quad i^*(\tau) = \frac{S^0 W_1}{R_1} e^{-(\mu+\gamma_1)\tau - \int_0^\tau \kappa(v)dv}, \quad J^* = \frac{K S^0 W_1}{(\mu + \gamma_2)\,(R_1 - R_2)}, \tag{5.12}$$

where W_1 is defined in (5.11).

5.2. Stability of the endemic equilibrium. We investigate the local stability of the endemic equilibrium, E^*, by linearizing system (2.1) about E^*. Let $x(t) = S(t) - S^*$, $y(t,\tau) = i(t,\tau) - i^*(\tau)$, and $z(t) = J(t) - J^*$. The linearization results in the perturbation equations

$$
\begin{aligned}
&\frac{dx(t)}{dt} = -\left(\mu + W_1 + \beta_2 J^*\right) x(t) - \beta_2 S^* z(t) - S^* \int_0^\infty \beta_1(\tau) y(t,\tau) d\tau, \\
&\begin{cases} \dfrac{\partial y(t,\tau)}{\partial t} + \dfrac{\partial y(t,\tau)}{\partial \tau} &= -(\mu+\gamma_1) y(t,\tau) - \kappa(\tau) y(t,\tau), \\ y(t,0) &= S^* \displaystyle\int_0^\infty \beta_1(\tau) y(t,\tau) d\tau + W_1 x(t), \end{cases} \qquad (5.13) \\
&\frac{dz(t)}{dt} = \beta_2 J^* x(t) - (\mu+\gamma_2) z(t) + \beta_2 S^* z(t) + \int_0^\infty \kappa(\tau) y(t,\tau) d\tau.
\end{aligned}
$$

Suppose $x = x_0 e^{\rho t}$, $y = \hat{y}(\tau) e^{\rho(t-\tau)}$, and $z = z_0 e^{\rho t}$. Substituting these variables into system (5.13) and solving for $\hat{y}(\tau)$, with initial condition $\hat{y}(0)$, leads to the system

$$
\begin{aligned}
&(\rho + \mu + W_1 + \beta_2 J^*) x_0 + \beta_2 S^* z_0 + S^* \int_0^\infty \beta_1(\tau) \hat{y}(\tau) e^{-\rho\tau} d\tau = 0, \\
&-\beta_2 J^* x_0 + (\rho + \mu + \gamma_2 - \beta_2 S^*) z_0 - \int_0^\infty \kappa(\tau) \hat{y}(\tau) e^{-\rho\tau} d\tau = 0, \qquad (5.14) \\
&\hat{y}(\tau) = \left(S^* \int_0^\infty \beta_1(\tau) \hat{y}(\tau) e^{-\rho\tau} d\tau + W_1 x_0 \right) e^{-(\mu+\gamma_1)\tau - \Delta(\tau)}.
\end{aligned}
$$

We simplify these notations by defining the functions

$$
\begin{aligned}
H(\rho) &:= \int_0^\infty \beta_1(\tau) \hat{y}(\tau) e^{-\rho\tau} d\tau, & Q(\rho) &:= \int_0^\infty \kappa(\tau) \hat{y}(\tau) e^{-\rho\tau} d\tau, \\
P_1(\rho) &:= \int_0^\infty \beta_1(\tau) e^{-\rho\tau} e^{-(\mu+\gamma_1)\tau - \Delta(\tau)} d\tau, & P_2(\rho) &:= \int_0^\infty \kappa(\tau) e^{-\rho\tau} e^{-(\mu+\gamma_1)\tau - \Delta(\tau)} d\tau.
\end{aligned}
$$

Multiplying $\hat{y}(\tau)$ in (5.14) by $\beta_1(\tau)e^{-\rho\tau}$ and $\kappa(\tau)e^{-\rho\tau}$, respectively, and then integrating from 0 to ∞ yields

$$
H(\rho) = \frac{W_1 P_1(\rho)}{1 - S^* P_1(\rho)} x_0 \qquad (5.15)
$$

and

$$
Q(\rho) = \left(S^* H(\rho) + W_1 x_0\right) P_2(\rho) = \left(\frac{S^* W_1 P_1(\rho)}{1 - S^* P_1(\rho)} + W_1 \right) P_2(\rho) x_0. \qquad (5.16)
$$

Substituting (5.15) and (5.16) into system (5.14), we obtain the characteristic equation

$$
\left(\rho + \mu + \beta_2 J^* + \frac{W_1}{1 - S^* P_1(\rho)} \right) (\rho + \mu + \gamma_2 - \beta_2 S^*) + \left(\beta_2 J^* + \frac{W_1 P_2(\rho)}{1 - S^* P_1(\rho)} \right) \beta_2 S^* = 0 \qquad (5.17)
$$

and arrive at the following result.

THEOREM 5.2. *The endemic equilibrium, given in* (5.12)*, is locally asymptotically stable if all roots, ρ, of the characteristic equation* (5.17) *have negative real part.*

The results obtained for the two-strain SIR model (2.1) are summarized in Table 1. The stability of the endemic equilibrium is not listed because it requires knowledge of the roots of the characteristic equation (5.17) and we have not established the explicit criterion.

TABLE 1

The existence conditions for the boundary and endemic equilibria, E_1 and E^*, and stability conditions for the infection-free and boundary equilibria, E_0 and E_1. These conditions are based on the relations between the two reproductive numbers, R_1 and R_2, for the two strains.

	$R_1 < 1$, $R_2 < 1$	$R_2 < 1 < R_1$	$R_1 < 1 < R_2$	$1 < R_1 < R_2$	$1 < R_2 < R_1$
E_0	stable	unstable	unstable	unstable	unstable
E_1	does not exist	does not exist	stable	stable	unstable
E^*	does not exist	exists	does not exist	does not exists	exists

6. Constant mutation rate. Because (5.17) is a transcendental equation, it is difficult to determine when all the roots of the characteristic equation have negative real part and, hence, whether the endemic equilibrium is stable. To gain insight into the transmission dynamics of the disease governed by system (2.1), we consider the special case where the mutation rate from Strain 1 to Strain 2 is constant and where the infection rate of Strain 1 is independent of the infection stages. We define these constant rates as $\kappa(\tau) := k$ and $\beta_1(\tau) := \beta_1$.

Let the total infectives be $I(t) := \int_0^\infty i(t,\tau)d\tau$. Integrating the equation for $i(t,\tau)$ in (2.1) with respect to τ and using the initial condition $i(t,0)$ reduces the system of PDEs to the system of ODEs,

$$\frac{dS}{dt} = \mu(S^0 - S) - \beta_1 IS - \beta_2 JS, \tag{6.1a}$$

$$\frac{dI}{dt} = \beta_1 SI - (\mu + \gamma_1 + k)I, \tag{6.1b}$$

$$\frac{dJ}{dt} = \beta_2 SJ - (\mu + \gamma_2)J + kI. \tag{6.1c}$$

The reproductive numbers of Strains 1 and 2, R_1 and R_2, for system (6.1) are

$$R_1 = \frac{S^0\beta_1}{\mu + \gamma_1 + k}, \qquad R_2 = \frac{S^0\beta_2}{\mu + \gamma_2}. \tag{6.2}$$

The only boundary equilibrium with $I = 0$ and $J > 0$ exists if $R_2 > 1$ and it has the same expression as in section 4. This boundary equilibrium is stable if $R_1 < R_2$ and is unstable if $R_1 > R_2$.

We now establish existence and local stability of the endemic equilibrium of system (6.1).

For $\kappa(\tau) = k$, the term K defined in (5.4) becomes

$$K = \frac{k}{\mu + \gamma_1 + k}. \tag{6.3}$$

Substituting (6.2) and (6.3) into (5.12), we obtain the endemic equilibrium, $E^* = (S^*, I^*, J^*)$, with

$$\begin{aligned} S^* &= \frac{\mu + \gamma_1 + k}{\beta_1}, \\ I^* &= \frac{\mu\big(S^0\beta_1 - (\mu + \gamma_1 + k)\big)\big(\beta_1(\mu + \gamma_2) - \beta_2(\mu + \gamma_1 + k)\big)}{\beta_1\big(\beta_1(\mu + \gamma_2) - \beta_2(\mu + \gamma_1)\big)(\mu + \gamma_1 + k)}, \\ J^* &= \frac{\mu k\big(S^0\beta_1 - (\mu + \gamma_1 + k)\big)}{\big(\beta_1(\mu + \gamma_2) - \beta_2(\mu + \gamma_1)\big)(\mu + \gamma_1 + k)}. \end{aligned} \tag{6.4}$$

By solving (6.1) for an endemic equilibrium, we have the equivalent solution

$$\begin{aligned} S^* &= \frac{S^0}{R_1}, \\ I^* &= \frac{\mu(\mu+\gamma_1+k)(R_1-1)(R_1-R_2)}{\beta_1\big(kR_1+(\mu+\gamma_1)(R_1-R_2)\big)}, \\ J^* &= \frac{\mu k S^0(R_1-1)}{(\mu+\gamma_2)\big(kR_1+(\mu+\gamma_1)(R_1-R_2)\big)}. \end{aligned}$$

Hence E^* exists if and only if $R_1 > 1$ and $R_1 > R_2$.

Based on $\mu + \beta_2 J^* = \mu S^0/S^* - \beta_1 I^*$, the characteristic equation for system (6.1) has the form

(6.5)
$$\left(\rho + \mu R_1 + \frac{\gamma_1+\mu+k}{\rho}\beta_1 I^*\right)(\rho+\mu+\gamma_2-\beta_2 S^*) + \left(\beta_2 J^* + \frac{k}{\rho}\beta_1 I^*\right)\beta_2 S^* = 0.$$

This can be expressed as

$$\rho^3 + a_1\rho^2 + a_2\rho + a_3 = 0,$$

where

$$\begin{aligned} a_1 &:= \mu R_1 + \mu + \gamma_2 - \beta_2 S^* = \mu\frac{S^0}{S^*} + k\frac{I^*}{J^*}, \\ a_2 &:= \mu R_1(\mu+\gamma_2-\beta_2 S^*) + (\mu+\gamma_1+k)\beta_1 I^* + \beta_2^2 J^* S^* \\ &= \beta_1^2 S^* I^* + \beta_2^2 S^* J^* + \mu\frac{S^0}{S^*}k\frac{I^*}{J^*}, \\ a_3 &:= \left((\mu+\gamma_2)\beta_1 - \beta_2(\mu+\gamma_1)\right)(\mu+\gamma_1+k)I^* = \beta_1 S^* k\frac{I^*}{J^*}\left(\beta_1 I^* + \beta_2 J^*\right). \end{aligned}$$

Since $a_1 > 0$ and $a_3 > 0$, it follows from the Routh–Hurwitz criterion that all characteristic roots of (6.5) have negative real part if and only if $a_1 a_2 > a_3$.

A straightforward calculation yields

(6.6)
$$\begin{aligned} a_1 a_2 - a_3 &= \mu\frac{S^0}{S^*}\left(\beta_1^2 S^* I^* + \beta_2^2 S^* J^* + \mu\frac{S^0}{S^*}k\frac{I^*}{J^*}\right) + \mu\frac{S^0}{S^*}\left(k\frac{I^*}{J^*}\right)^2 + kS^* I^*\beta_2(\beta_2-\beta_1) \\ &= \frac{\mu(R_1-1)(R_1-R_2)}{(\sigma_1+k)R_1-\sigma_1 R_2}\left(\mu R_1(\sigma_1+k)^2 + \frac{\mu\sigma_2 R_2^2 k}{R_1-R_2}\right. \\ &\qquad\qquad \left. - \frac{k\sigma_2 R_2\big(R_1(\sigma_1+k)-R_2\sigma_2\big)}{R_1^2}\right) \\ &\quad + \frac{\mu(R_1-R_2)}{R_1^2}(\mu\sigma_2 R_1^3 + \sigma_2^2(R_1-R_2)R_1) \\ &= \frac{\mu(R_1-1)(R_1-R_2)}{R_1^2\big((\sigma_1+k)R_1-\sigma_1 R_2\big)}(c_2k^2 + c_1 k + c_0), \end{aligned}$$

where

$$\sigma_1 := \mu + \gamma_1, \quad \sigma_2 := \mu + \gamma_2,$$

$$c_0 := \mu\sigma_1^2 R_1^3 + \sigma_2 R_1\big(\mu R_1^2 + \sigma_2(R_1 - R_2)\big)\frac{\sigma_1(R_1 - R_2)}{R_1 - 1},$$

$$c_1 := 2\mu\sigma_1 R_1^3 + \frac{\mu\sigma_2 R_1^2 R_2^2}{R_1 - R_2} + \frac{\sigma_2 R_1^2\big(\mu R_1^2 + \sigma_2(R_1 - R_2)\big)}{R_1 - 1} - \sigma_2 R_2(\sigma_1 R_1 - \sigma_2 R_2),$$

$$c_2 := \mu R_1^3 - \sigma_2 R_1 R_2.$$

Hence all roots of (6.5) have negative real part if $c_2k^2 + c_1k + c_0 > 0$, and at least one of the roots of (6.5) has positive real part if $c_2k^2 + c_1k + c_0 < 0$.

We summarize the results in the following theorem.

THEOREM 6.1. *When the mutation rate is constant, the dynamical behavior of epidemic model* (6.1) *can be described as one of the following cases:*

1. *If we define* $R_0 := \max\{R_1, R_2\}$ *and* $R_0 < 1$, *then the infection-free equilibrium,* $E_0 := (S^0, 0, 0)$, *is the only equilibrium and is locally asymptotically stable. If* $R_0 > 1$, *then* E_0 *is unstable.*
2. *If* $R_1 < 1 < R_2$, *or* $1 < R_1 < R_2$, *the only boundary equilibrium, given by*

$$E_1 := \left(\tilde{S}, 0, \tilde{J}\right) = \left(S^0 R_2,\ 0,\ \frac{\mu S^0}{\sigma_2 R_2}(R_2 - 1)\right), \tag{6.7}$$

   exists and is locally asymptotically stable. In this case, the endemic equilibrium, E^*, *does not exist.*
3. *If* $R_2 < 1 < R_1$, *the endemic equilibrium,* E^*, *exists and is the only nontrivial equilibrium. It is locally asymptotically stable if* $c_2k^2 + c_1k + c_0 > 0$ *and unstable if* $c_2k^2 + c_1k + c_0 < 0$.
4. *If* $1 < R_2 < R_1$, *the boundary equilibrium,* E_1, *exists but is unstable. The endemic equilibrium,* E^*, *exists and is locally asymptotically stable if* $c_2k^2 + c_1k + c_0 > 0$ *and unstable if* $c_2k^2 + c_1k + c_0 < 0$.

6.1. The global stability of the equilibria. In this section we establish that when the infection-free equilibrium and the boundary equilibrium of system (6.1) are locally asymptotically stable, they are globally stable.

THEOREM 6.2.

1. *If the infection-free equilibrium,* E_0, *is locally asymptotically stable, then it is globally stable; that is,* E_0 *is globally asymptotically stable if* $R_0 < 1$.
2. *If* $R_1 < 1 < R_2$, *the only boundary equilibrium,* E_1, *given in* (6.7), *is globally asymptotically stable.*

Proof. It follows from (6.1b) that

$$I(t) = I(0)e^{\int_0^t \beta_1 S(\tau)d\tau - (\mu+\gamma_1+k)t}$$

for all $t \geq 0$. Hence, the hyperplane $I = 0$ is invariant for system (6.1).

If $R_1 < 1$, we can further show that the hyperplane attracts all solutions started in the first octant, $S \geq 0$, $I \geq 0$, $J \geq 0$. That is, $\lim_{t\to\infty} I(t) = 0$. It can be seen from (6.1a) that $dS/dt \leq \mu(S^0 - S)$ and hence $S(t) \leq S^0 + S(0)e^{-\mu t}$ and from (6.1b) that

$$\begin{aligned} I(t) &\leq I(0)e^{\int_0^t \beta_1\left(S^0 + S(0)e^{-\mu\tau}\right)d\tau - (\mu+\gamma_1+k)t} = I(0)e^{(\mu+\gamma_1+k)(R_1-1)t + \frac{\beta_1 S(0)}{\mu}\left(1-e^{-\mu t}\right)} \\ &\leq I(0)e^{\frac{\beta_1 S(0)}{\mu}}e^{(\mu+\gamma_1+k)(R_1-1)t} \to 0 \end{aligned}$$

as $t \to \infty$.

Based on the attractiveness of the hyperplane $I = 0$, to prove the global asymptotic stability of the infection-free equilibrium E_0 or the boundary equilibrium E_1 in the first octant, it suffices to show that these two equilibria are globally asymptotically stable in the hyperplane $I = 0$.

We first show that all the solutions of (6.1) in the hyperplane $I = 0$ approach E_0 if $R_0 < 1$. We use the Lyapunov function V_1 defined by

$$V_1(S, J) := J + S - S^0 - S^0 \ln \frac{S}{S^0}$$

for system (6.1). Along the trajectories of system (6.1) in the hyperplane $I = 0$ we have

$$\begin{aligned}
\left.\frac{dV_1}{dt}\right|_{(6.1)} &= (\beta_2 S - (\mu + \gamma_2))\, J + \frac{S - S^0}{S}\left(\mu(S^0 - S) - \beta_2 JS\right) \\
&= -\frac{\mu(S - S^0)^2}{S} + \left(S^0\beta_2 - (\mu + \gamma_2)\right) J \\
&= -\frac{\mu(S - S^0)^2}{S} + J(R_2 - 1)\sigma_2 < 0
\end{aligned}$$

if $R_2 < 1$. Hence it follows from Lyapunov stability theory that E_0 is globally asymptotically stable.

We next assume $R_1 < 1 < R_2$ and show the global stability of the boundary equilibrium $E_1 = (\tilde{S}, 0, \tilde{J})$. We use

$$V_2(S, J) = J - \tilde{J} - \tilde{J} \ln \frac{J}{\tilde{J}} + S - \tilde{S} - \tilde{S} \ln \frac{S}{\tilde{S}}$$

as a Lyapunov functions for system (6.1). In the hyperplane $I = 0$,

$$\begin{aligned}
\left.\frac{dV_2}{dt}\right|_{(6.1)} &= (\beta_2 S - (\mu + \gamma_2))\,(J - \tilde{J}) + \frac{\mu(S^0 - S)(S - \tilde{S})}{S} - (S - \tilde{S})\beta_2 J \\
&= \beta_2(S - \tilde{S})(J - \tilde{J}) + \frac{\mu(S^0 - S)(S - \tilde{S})}{S} - (S - \tilde{S})\beta_2 J \\
&= -\beta_2(S - \tilde{S})\tilde{J} + \frac{\mu(S^0 - S)(S - \tilde{S})}{S} \\
&= -\frac{S - \tilde{S}}{S}\left(S\beta_2\tilde{J} - \mu(S^0 - S)\right) = -\frac{S - \tilde{S}}{S}\left((\beta_2\tilde{J} + \mu)S - \mu S^0\right) \\
&= -\frac{S - \tilde{S}}{S}\left(\frac{\mu S^0}{\tilde{S}}S - \mu S^0\right) = -\frac{\mu S^0(S - \tilde{S})^2}{\tilde{S}S} \le 0.
\end{aligned}$$

The maximum invariant subset of the set $\left\{(S, I, J) \;\middle|\; \frac{dV}{dt} = 0\right\}$ in the hyperplane $I = 0$ contains only E_1. Then it follows from the LaSalle invariance principle that E_1 is globally asymptotically stable on the hyperplane $I = 0$. □

Note that we have not been able to prove the global stability of the boundary equilibrium E_1 for the case $1 < R_1 < R_2$.

6.2. Hopf bifurcation near the endemic equilibrium. We know from Theorem 6.1 that if $R_2 < 1 < R_1$ or $1 < R_2 < R_1$, the boundary equilibrium either does not exist or is unstable, and the positive endemic equilibrium is asymptotically stable if $c_2k^2 + c_1k + c_0 > 0$ and is unstable if $c_2k^2 + c_1k + c_0 < 0$. We now show that as the

endemic equilibrium loses stability, periodic solutions can bifurcate from the endemic equilibrium.

To investigate the bifurcation and to simplify the mathematical analysis, we study the bifurcation in terms of the mutation rate k and the two basic reproductive numbers R_1 and R_2 and assume that individuals infected by the two strains have the same recovery rate $\gamma_1 = \gamma_2 := \gamma$, and hence $\sigma_1 = \sigma_2 := \sigma$. Under these assumptions, and after some tedious algebraic manipulations, (6.6) becomes

(6.8)
$$a_1a_2 - a_3 = \mu(R_1 - 1)(R_1 - R_2)\Big(\mu - \frac{\sigma R_2}{R_1^2}\Big)k + \sigma\mu^2 R_1(2R_1 - R_2 - 1) + \frac{\mu\sigma^2(R_1 - R_2)^2}{R_1}.$$

All terms in (6.8) are positive except $\mu - \sigma R_2/R_1^2$. If $\mu R_1^2 \geq \sigma R_2$, then $a_1a_2 > a_3$. It follows from the Routh–Hurwitz criterion that the endemic equilibrium E^* is locally asymptotically stable.

Suppose $\mu R_1^2 < \sigma R_2$. We define a critical number k_0 as

$$k_0 = \frac{\sigma\mu R_1^3(2R_1 - R_2 - 1) + \sigma^2(R_1 - R_2)^2 R_1}{(R_1 - 1)(R_1 - R_2)(\sigma R_2 - \mu R_1^2)} \tag{6.9}$$

such that E^* is locally asymptotically stable if $k < k_0$ and is unstable if $k > k_0$. For $k = k_0$, the characteristic equation (6.5) for the linearization of system (6.1) has two pure imaginary roots. The parameter k can be used as a bifurcation parameter such that as k passes through k_0, a Hopf bifurcation occurs and a periodic solution bifurcates from the endemic equilibrium.

The reproductive numbers R_1 and R_2 can also be used as bifurcation parameters. Rewrite $a_1a_2 - a_3$ as a quadratic function of $R_1 - R_2$:

$$a_1a_2 - a_3 = \mu d_2(R_1 - R_2)^2 + \mu d_1(R_1 - R_2) + \mu d_0,$$

where

$$\begin{aligned} d_0 &:= \sigma\mu R_1(R_1 - 1), \\ d_1 &:= (R_1 - 1)(\mu - \sigma/R_1)k + \sigma\mu R_1, \\ d_2 &:= \sigma^2/R_1 + \sigma(R_1 - 1)k/R_1^2. \end{aligned}$$

Fixing R_1 and then solving the equation $d_2(R_1 - R_2)^2 + d_1(R_1 - R_2) + d_0 = 0$ for R_2 yields the two solutions

$$R_2^+ = R_1 + \frac{d_1 + \sqrt{d_1^2 - 4d_2d_0}}{2d_2}, \qquad R_2^- = R_1 + \frac{d_1 - \sqrt{d_1^2 - 4d_2d_0}}{2d_2}.$$

For $R_1 > 1$, $d_2 > 0$ and $d_0 > 0$. If $R_1 > R_2$ and $d_1^2 < 4d_2d_0$, the inequality $a_1a_2 - a_3 > 0$ always holds. The endemic equilibrium, E^*, is locally asymptotically stable. If $d_1^2 > 4d_2d_0$, E^* is locally asymptotically stable provided $0 < R_2 < R_2^-$ or $R_2^+ < R_2 < R_1$ and is unstable provided $R_2^- < R_2 < R_2^+$. As R_2 passes through R_2^- or R_2^+, periodic solutions bifurcate from the endemic equilibrium.

The dynamics of system (6.1) are summarized, based on R_1 and R_2, in Figure 1. We divide the R_1-R_2 plane into five regions. In Region I, $R_1 < 1$ and $R_2 < 1$. The infection-free equilibrium, E_0, is the only equilibrium and is globally asymptotically stable. In both Regions II and III, the boundary equilibrium, E_1, is globally asymptotically stable, whereas the endemic equilibrium, E^*, does not exist in Region II

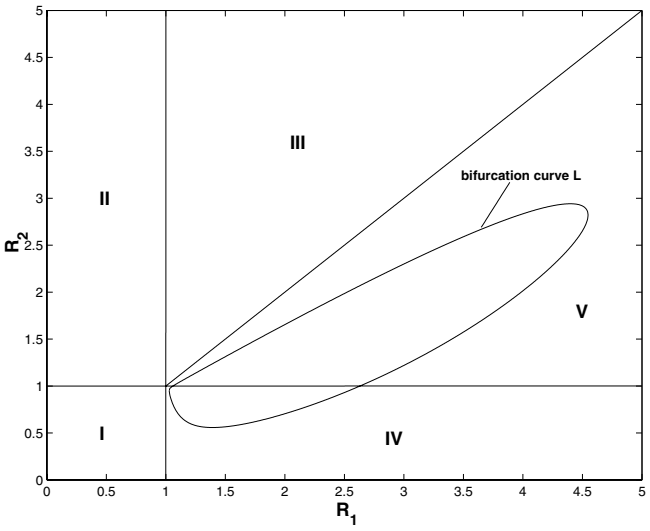


FIG. 1. *Schematic illustrations of dynamical behavior of system* (6.1) *based on the reproductive numbers,* R_1 *and* R_2*. The infection-free equilibrium,* E_0*, is the only equilibrium in Region* I *and is globally asymptotically stable. The boundary equilibrium,* E_1*, exists in Regions* II, III, *and* V*. It is globally asymptotically stable in both Regions* II *and* III *but is unstable in Region* V*. The endemic equilibrium,* E^**, exists in Regions* III, IV, *and* V*. It is unstable in Region* III *and in the interior of the region enclosed by the bifurcation curve* L*. It is locally asymptotically stable in the complement of the region enclosed by curve* L *in* IV *and* V*. For a fixed* R_1 *in the interval of the projection of curve* L *on the* R_1*-axis, as* R_2 *crosses through curve* L*, periodic solutions are bifurcated.*

and exists but is unstable in Region III. While E^* exists in both Regions IV and V and is the only nontrivial equilibrium in Region IV, and E_1 exists but is unstable in Region V, the stability of E^* is determined by the closed bifurcation curve L in these two regions. E^* is unstable and a Hopf bifurcation takes place in the interior of the region enclosed by L. E^* is asymptotically stable elsewhere in Regions IV and V.

We illustrate these results by examples using k, or R_1 and R_2, as bifurcation parameters.

Example 6.1. We use k as a bifurcation parameter and let $\sigma_1 = \sigma_2 = 1/2$, $\mu = 1/100$, $R_1 = 3$, and $R_2 = 2$. System (6.1) becomes

$$\begin{aligned} \frac{dS}{dt} &= \frac{1}{100}(S^0 - S) - \left(\frac{3+6k}{2S^0}I + \frac{1}{S^0}J\right)S, \\ \frac{dI}{dt} &= \frac{3+6k}{2S^0}SI - \frac{1+2k}{2}I, \\ \frac{dJ}{dt} &= \frac{1}{S_0}SJ - \frac{1}{2}J + kI \end{aligned} \tag{6.10}$$

and has the endemic equilibrium

$$E^* = \left(\frac{S^0}{3}, \frac{S^0}{75(1+6k)}, \frac{2kS^0}{25(1+6k)}\right) = (S^*, I^*, J^*).$$

The linearization of system (6.10) at E^* has the characteristic equation

$$f(\rho) = \rho^3 + \frac{59}{300}\rho^2 + \left(\frac{k}{150} + \frac{3}{200}\right)\rho + \frac{k}{300} + \frac{1}{600} = 0. \tag{6.11}$$

The critical number k_0 defined in (6.9) can be determined as $k_0 = 33/52$. Then, if $k < 33/52$, all roots of (6.11) have negative real part, and hence the endemic equilibrium of (6.10) is stable. If $k > 33/52$, there exist two roots with positive real part, and hence the endemic equilibrium of system (6.10) is unstable. For $k = 33/52$, (6.11) has a negative real root and two pure imaginary conjugates:

$$\rho_1 = -\frac{59}{300}, \quad \rho_2 = \frac{\sqrt{13}}{26}i, \quad \rho_3 = -\frac{\sqrt{13}}{26}i.$$

For k greater than, but near $33/52$, (6.11) has a negative real root $\rho_1(k)$ and a pair of complex conjugates $\rho_2(k) = \bar{\rho}_3(k) := \xi(k) + i\eta(k)$. Substituting the complex conjugates into (6.11) and then separating the real and imaginary parts yields the equations for $\xi(k)$ and $\eta(k)$:

$$\begin{aligned} &\xi^3 - 3\xi\eta^2 + \frac{59}{300}\xi^2 - \frac{59}{300}\eta^2 + \frac{k}{150}\xi + \frac{3}{200}\xi + \frac{k}{300} + \frac{1}{600} = 0, \\ &3\xi^2\eta - \eta^3 + \frac{59}{150}\xi\eta + \frac{k}{150}\eta + \frac{3}{200}\eta = 0. \end{aligned} \tag{6.12}$$

By differentiating (6.12) with respect to k, we have

$$\begin{aligned} &\left(3\xi^2 - 3\eta^2 + \frac{59}{150}\xi + \frac{k}{150} + \frac{3}{200}\right)\frac{d\xi}{dk} - \left(6\xi\eta + \frac{59}{150}\eta\right)\frac{d\eta}{dk} + \frac{1}{150}\xi + \frac{1}{300} = 0, \\ &\left(6\xi\eta + \frac{59}{150}\eta\right)\frac{d\xi}{dk} + \left(3\xi^2 - 3\eta^2 + \frac{59}{150}\xi + \frac{k}{150} + \frac{3}{200}\right)\frac{d\eta}{dk} + \frac{1}{150}\eta = 0. \end{aligned} \tag{6.13}$$

Solving (6.13) for $d\xi/dk$ and substituting $k = 33/52$, $\xi = 0$, and $\eta = \sqrt{13}/26$ into the expression of $d\xi/dk$ yields $d\xi/dk = 169/9679 > 0$. Therefore, system (6.10) undergoes a Hopf bifurcation and a periodic solution is bifurcated near $k = 33/52$.

To determine the bifurcation direction, we first discuss the stability of the endemic equilibrium of system (6.10) as $k = 33/52$. Let $x_1 = S - S^*$, $y_1 = I - I^*$, and $z_1 = J - J^*$ to transform the endemic equilibrium to the origin of a new system. Using the linear transformation

$$\begin{aligned} x_1 &= \frac{3125\sqrt{13}}{767}y - \frac{125}{6}z, \\ y_1 &= x + z, \\ z_1 &= \frac{639}{236}x - \frac{1125\sqrt{13}}{3068}y - \frac{539}{39}z, \end{aligned}$$

and rescaling $t = 2\sqrt{13}\,\hat{t}$, we transform the resulting system into

$$\begin{aligned} \frac{dx}{d\hat{t}} &\approx y + 319.95\left(S^0\right)^{-1}xy - 453.74\left(S^0\right)^{-1}xz - 6.93\left(S^0\right)^{-1}y^2 \\ &\quad + 243.20\left(S^0\right)^{-1}yz - 330.96\left(S^0\right)^{-1}z^2, \\ \frac{dy}{d\hat{t}} &\approx -x + 13.55\left(S^0\right)^{-1}xy - 19.22\left(S^0\right)^{-1}xz + 19.36\left(S^0\right)^{-1}y^2 \\ &\quad + 228.07\left(S^0\right)^{-1}yz - 362.37\left(S^0\right)^{-1}z^2, \\ \frac{dz}{d\hat{t}} &\approx -1.42z + 40.63\left(S^0\right)^{-1}xy - 57.62\left(S^0\right)^{-1}xz + 6.93\left(S^0\right)^{-1}y^2 \\ &\quad + 117.39\left(S^0\right)^{-1}yz - 180.40\left(S^0\right)^{-1}z^2. \end{aligned} \tag{6.14}$$

The nonlinear terms of the right-hand side of system (6.14) are quadratic and satisfy the existence conditions of the center manifold theorem [4, 12]. Hence, there exists a manifold $z = h(x, y)$ of system (6.14) which can be expanded as

$$z = h_{20}x^2 + h_{11}xy + h_{02}y^2 + o(r^2), \quad r = \sqrt{x^2 + y^2}, \tag{6.15}$$

where $o(r^2)$ denotes higher order terms and h_{ij} are to be determined.

Substituting (6.15) into system (6.14), we obtain

$$h_{20} = 8.38\left(S^0\right)^{-1}, \quad h_{11} = 11.89\left(S^0\right)^{-1}, \quad h_{02} = -3.50\left(S^0\right)^{-1}.$$

Substituting (6.15) with these h_{ij} again into the first two equations of system (6.14), we have the following equations on the center manifold:

(6.16)

$$\begin{aligned}
\frac{dx}{d\hat{t}} &= y - 6.93\left(S^0\right)^{-1} y^2 + 319.95\left(S^0\right)^{-1} xy - 3802.35\left(S^0\right)^{-2} x^3 - 3357.04\left(S^0\right)^{-2} x^2 y \\
&\quad + 4479.62\left(S^0\right)^{-2} xy^2 - 851.17\left(S^0\right)^{-2} y^3 + o(r^3), \\
\frac{dy}{d\hat{t}} &= -x + 13.55\left(S^0\right)^{-1} xy + 19.36\left(S^0\right)^{-1} y^2 - 161.06\left(S^0\right)^{-2} x^3 \\
&\quad + 1682.70\left(S^0\right)^{-2} x^2 y + 2779.06\left(S^0\right)^{-2} xy^2 - 798.25\left(S^0\right)^{-2} y^3 + o(r^3).
\end{aligned}$$

Consider the function

$$\begin{aligned}
V(x, y) = x^2 + y^2 &- 239.10\left(S^0\right)^{-1} x^3 - 38.71\left(S^0\right)^{-1} xy^2 + 4.42\left(S^0\right)^{-1} y^3 \\
&+ 59133.80\left(S^0\right)^{-2} x^4 - 6381.38\left(S^0\right)^{-2} x^3 y - 151.46\left(S^0\right)^{-2} xy^3 \\
&- 2462.13\left(S^0\right)^{-2} y^4.
\end{aligned}$$

It is positive definite in a small neighborhood of the origin. Along the trajectories of system (6.16),

$$\left.\frac{dV(x, y)}{dt}\right|_{(6.16)} = -1223.33\left(S^0\right)^{-2}(x^2 + y^2)^2 + o(r^4) < 0.$$

Therefore, V is a Lyapunov function for system (6.16) and the trivial solution of system (6.16) is asymptotically stable. It follows from the reducible principle of the center manifold theorem that the trivial solution of system (6.14), and hence the endemic equilibrium of system (6.10), is asymptotically stable for $k = 33/52$. Since the endemic equilibrium is unstable for $k > 33/52$, it follows from the Hopf bifurcation theorem that there exists a stable periodic solution in the neighborhood of the endemic equilibrium of system (6.10).

We illustrate the stable endemic equilibrium ($k < k_0$) and the stable periodic solutions ($k > k_0$) in Figures 2 and 3. In Figure 2, $k = 0.135 < k_0 = 0.6346$, and the endemic equilibrium $E^* = (3.3363, 0.0074, 0.0059)$ is asymptotically stable. In Figure 3, $k = 0.9846 > k_0 = 0.6346$, and the endemic equilibrium E^* is unstable. The solutions quickly converge to the stable periodic solution.

Example 6.2. In this example, we use R_1 and R_2 as bifurcation parameters. Let

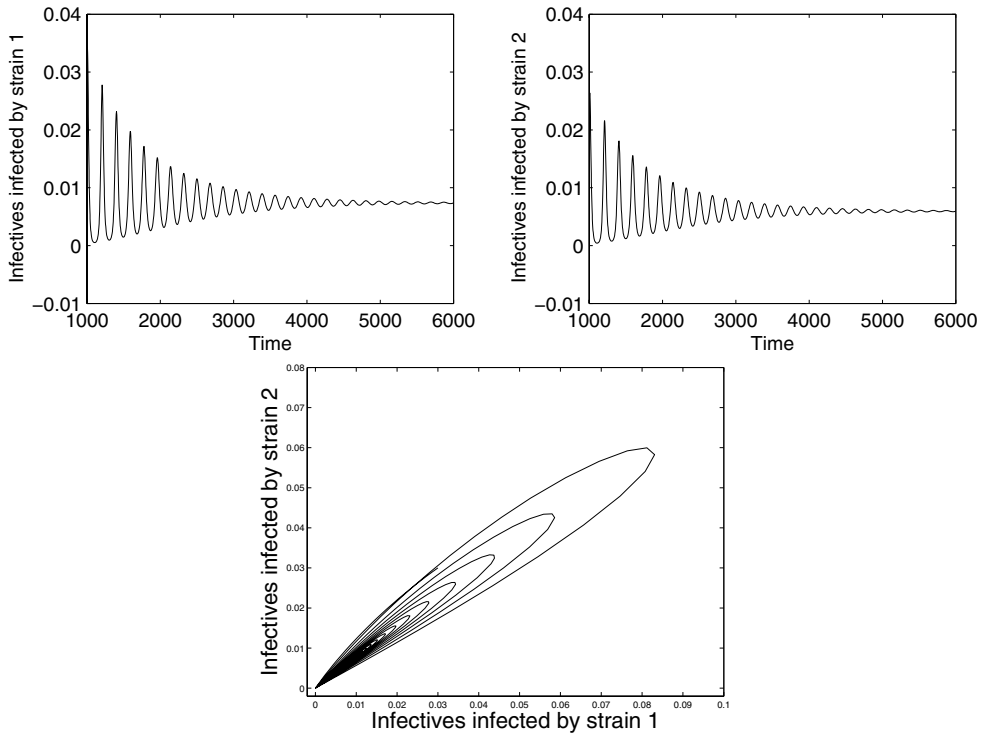


FIG. 2. *The solutions of system* (6.1) *for* $\mu = 0.01$, $\gamma_1 = \gamma_2 = 0.49$, $R_1 = 3$, *and* $R_2 = 2$. *The mutation rate* $k = 0.135$ *is used as a bifurcation parameter and is less than the critical value* $k_0 = 33/52$. *The endemic equilibrium* $(3.3363, 0.0074, 0.0059)$ *is asymptotically stable. The top two figures are the solutions of* I *and* J *versus time* t. *The bottom figure is the projected* I-J *phase plane of the phase space.*

$\sigma_1 = \sigma_2 = 1/10$, $\mu = 1/100$, and $k = 9/10$ in system (6.1), so that we have

$$\begin{aligned} \frac{dS}{dt} &= \frac{1}{100}(S^0 - S) - \left(\frac{R_1}{S^0}I + \frac{R_2}{10S^0}J\right)S, \\ \frac{dS}{dt} &= \frac{R_1}{S^0}SI - I, \\ \frac{dS}{dt} &= \frac{R_2}{10S_0}SJ - \frac{1}{10}J + \frac{9}{10}I. \end{aligned} \tag{6.17}$$

In region $D := \{(R_1, R_2) \mid R_1 > R_2,\ R_1 > 1\}$, the endemic equilibrium of system (6.17) is given by

$$E^* = \left(\frac{S^0}{R_1}, \frac{S^0(R_1 - 1)(R_1 - R_2)}{10(10R_1 - R_2)R_1}, \frac{9S^0(R_1 - 1)}{10(10R_1 - R_2)}\right).$$

The characteristic equation of the linearization of system (6.17) at E^* is

$$f(\rho) = \rho^3 + a_1\rho^2 + a_2\rho + a_3 = 0, \tag{6.18}$$

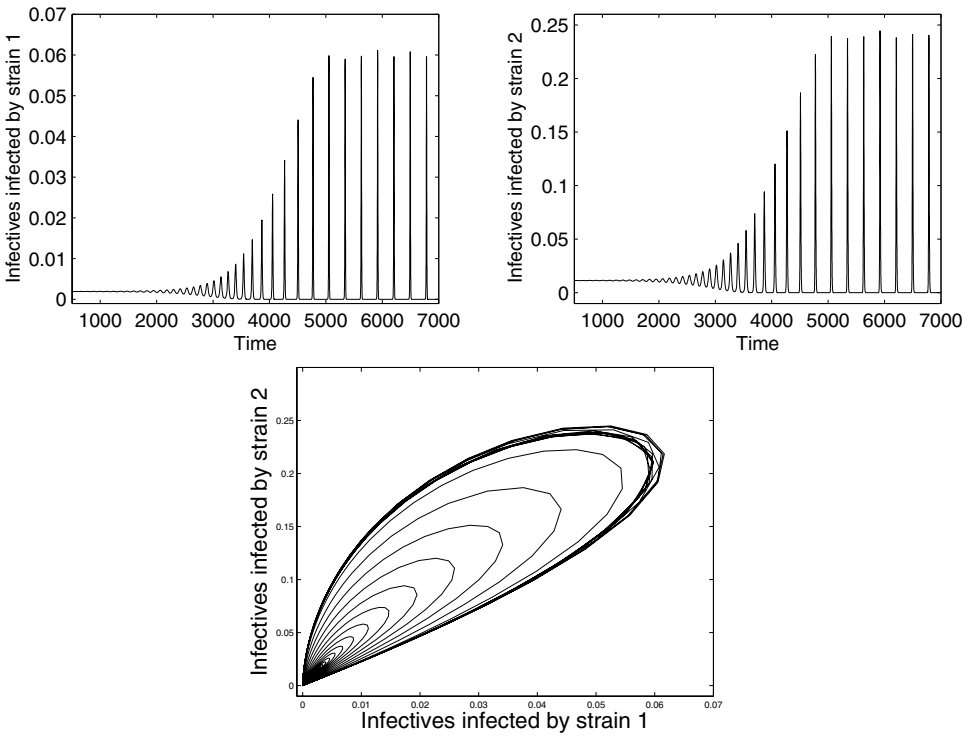


FIG. 3. *All parameters are the same as those in Figure* 2*, except the mutation rate* $k = 0.9846$ *is greater than the critical value* $k_0 = 33/52$*. The endemic equilibrium is unstable and a stable periodic solution is bifurcated from the endemic equilibrium. The top two figures show how the solutions with initial values near the unstable endemic equilibrium rapidly converge to the stable periodic solution. This can be also seen in the bottom figure of the I-J phase plane.*

where

$$a_1 = \frac{R_1^2 + 10R_1 - 10R_2}{100R_1},$$
$$a_2 = \frac{11R_1^2 - 10R_1R_2 - 10R_1 + 9R_2}{1000R_1},$$
$$a_3 = \frac{(R1-1)(R1-R2)}{1000R_1}.$$

E^* is asymptotically stable if

$$a_1a_2 - a_3 = \frac{100R_2^2R_1 - 90R_2^2 - 101R_2R_1^2 + 90R_1R_2 - 10R_1^3R_2 + 11R_1^4}{100000R_1^2} > 0. \tag{6.19}$$

Define function $H(R_2)$ as the numerator in (6.19). Then

$$H(R_2) = (100R_1 - 90)\,R_2^2 - \left(10R_1^3 + 101R_1^2 - 90R_1\right)R_2 + 11R_1^4.$$

The two zeros $R_2^{(1)} < R_2^{(2)}$, for R_1 and R_2, are in D, if

$$Q(R_1) := \left(10R_1^3 + 101R_1^2 - 90R_1\right)^2 - 44R_1^4\,(100R_1 - 90) > 0.$$

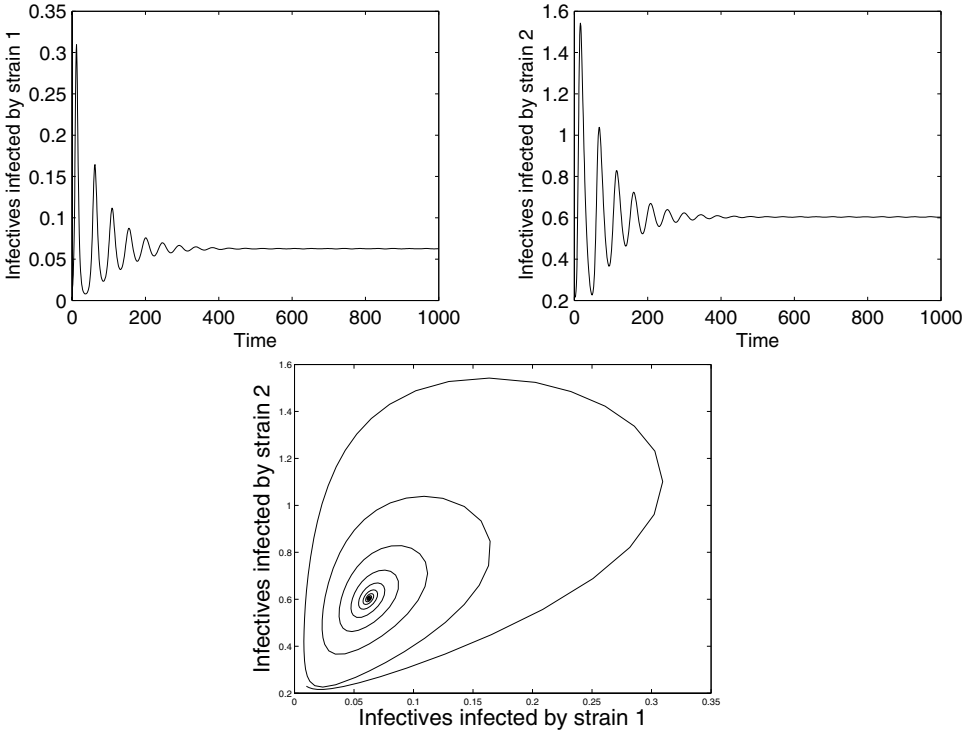


FIG. 4. *The reproductive numbers R_1 and R_2 are used as bifurcation parameters. The parameters $R_1 = 3$, $R_2^{(1)} = 1.5$, and $R_2 = 0.2$ are chosen so that $R_2 < R_2^{(1)}$. Then R_1 and R_2 are in Region* IV *in Figure* 1. *The endemic equilibrium exists and is asymptotically stable.*

Numerical computations verify that $Q(R_1)$ has two zeros, $R_1^{(1)} < R_1^{(2)}$, in the intervals $(1.01, 1.04)$ and $(4.50, 4.60)$, respectively. If $R_1 < R_1^{(1)}$ or $R_1 > R_1^{(2)}$, then $Q(R_1) < 0$, and if $R_1^{(1)} < R_1 < R_1^{(2)}$, then $Q(R_1) > 0$.

Suppose $R_1 < R_1^{(1)}$ or $R_1 > R_1^{(2)}$. Then $Q(R_1) < 0$ and $H(R_2)$ is always positive. If $R_1^{(1)} < R_1 < R_1^{(2)}$, then $Q(R_1) > 0$ and there are two zeros of $H(R_1)$, $R_2^{(1)} < R_2^{(2)}$ in D. If, moreover, $R_2 < R_2^{(1)}$ or $R_2 > R_2^{(2)}$, then $H(R_2) > 0$. Hence, in either case, $H(R_2) > 0$ and E^* is asymptotically stable. However, if $R_1^{(1)} < R_1 < R_1^{(2)}$ but $R_2^{(1)} < R_2 < R_2^{(2)}$, then $H(R_2) < 0$, for R_2 in D, and the endemic equilibrium is unstable.

For each R_1 in the interval $(R_1^{(1)}, R_1^{(2)})$, E^* changes its stability as R_2 increases from 0 to R_1. E^* is stable for R_2 in $(0, R_2^{(1)})$, unstable for R_2 in $(R_2^{(1)}, R_2^{(2)})$, and stable again for R_2 in $(R_2^{(2)}, R_1)$. At $R_2 = R_2^{(1)}$ or $R_2 = R_2^{(2)}$, the roots of characteristic equation (6.18) are imaginary indicating the existence of a periodic solution by Hopf bifurcation theory.

In numerical simulations, we fix $R_1 = 3$. The two roots of $H(R_2) = 0$ are $R_2^{(1)} = 3/2$ and $R_2^{(2)} = 99/35$. The characteristic roots of (6.18), with $R_2^{(1)} = 3/2$, are $\rho = -2/25$, $\rho = \sqrt{5}/20i$, and $\rho = -\sqrt{5}/20i$. The characteristic roots of (6.18), for $R_2^{(2)} = 99/35$, are $\rho = -1/28$, $\rho = \sqrt{2}/25i$, and $\rho = -\sqrt{2}/25i$.

In Figure 4, $\beta_1 = 0.3$ and $\beta_2 = 0.002$, and $R_2 = 0.2 < R_2^{(1)}$. The endemic equilib-

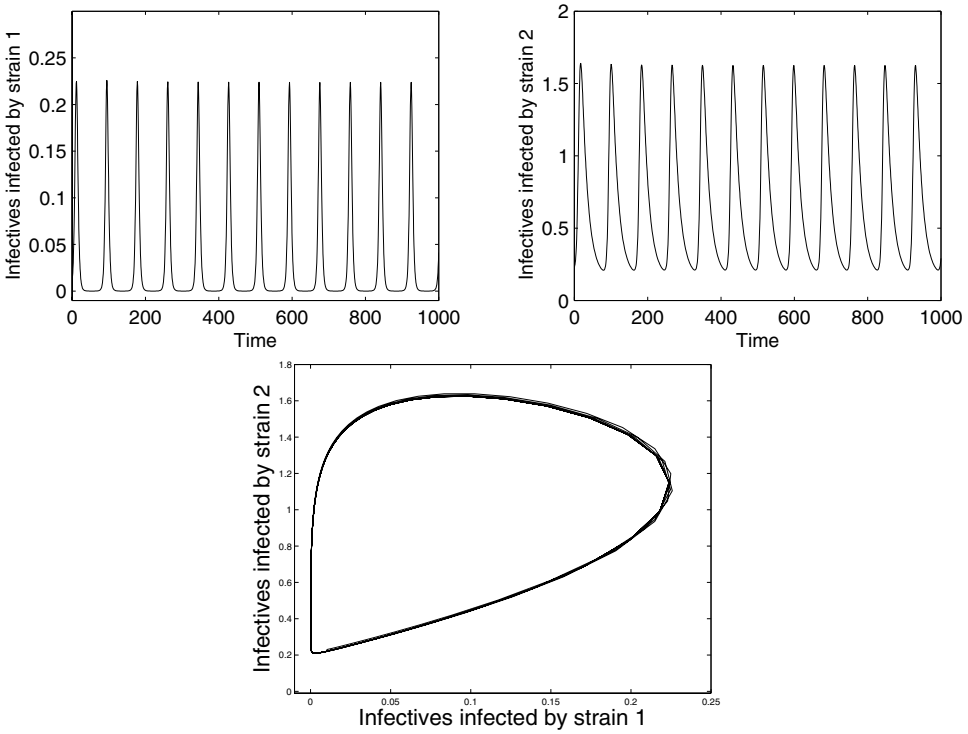


FIG. 5. *The parameters are chosen as in Figure* 4*, except* $R_2 = 2$ *by increasing* β_2 *to* 0.02 *whereas* $\beta_2 = 0.002$ *in Figure* 4*. Then* $R_2^{(1)} < R_2 < R_2^{(2)} = 2.829$*, and* R_1 *and* R_2 *are in the interior of the region enclosed by the bifurcation curve L, in Figure* 1*. The endemic equilibrium loses its stability. A periodic solution is bifurcated and is asymptotically stable.*

rium $E^* = (3.3348, 0.02627, 0.6032)$ is locally asymptotically stable, as is shown. We then increase β_2 to 0.02 so that $R_2 = 2$, which is between $R_2^{(1)}$ and $R_2^{(2)}$. The endemic equilibrium loses its stability and a periodic solution is bifurcated from the endemic equilibrium, as is shown in Figure 5. We continue increasing β_2 to 0.0286 such that $R_2 = 2.8571 > R_2^{(2)}$. The periodic solution disappears and the endemic equilibrium, $E^* = (3.3358, 0.0035, 0.6621)$, regains its stability, as is shown in Figure 6.

7. Concluding remarks. One of the challenges in modeling the spread of infectious diseases is to understand and predict the spread of competing strains of the same pathogen. After a strain of a pathogen infects a host, the mutation can be caused by an attempt of a pathogen to evade the immune defense of the host, the effect of selective immunologic pressure, or possibly adaptation to a more efficiently transmitted or better replicating pathogen.

We have formulated a simple compartmental mathematical model for the competition, mutation, and spread of a pathogen and its mutant strain. The model accounts for a continuous infection-age structure for the original pathogen, and the mutation rate of the pathogen depends on how long the host has been infected.

We model the transmission dynamics of pathogens by a system of partial differential-integral equations. We established conditions for the existence and stability of the infection-free equilibrium, the boundary equilibrium, and the endemic equilibrium. We derived formulas for the reproductive numbers, R_1 and R_2, for the two strains

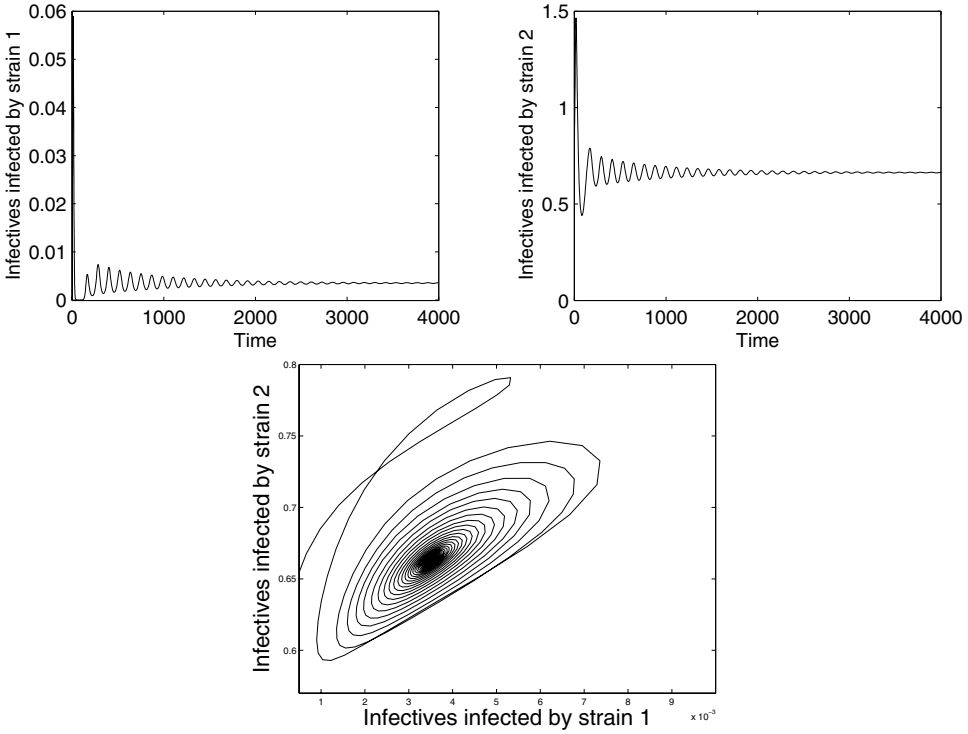


FIG. 6. *The parameters are chosen as in Figure* 4 *except* $\beta_2 = 0.0286$ *so that* $R_2 = 2.857 > R_2^{(2)}$. *Then* R_1 *and* R_2 *are in Region* V *and above the region enclosed by the bifurcation curve* L. *The endemic equilibrium regains its stability.*

based on the local stability of the infection-free equilibrium. We established the conditions for existence of the boundary equilibrium, E_1, where only one strain of the pathogen is in circulation, and the endemic equilibrium, E^*, where both the strain and its mutant are in circulation. We obtained stability conditions for E_1. These conditions, listed in Table 1, are expressed in terms of the two reproductive numbers. We investigated the stability of E^* and derived the characteristic equation of the linearization about E^*. The roots of this transcendental equation determine the stability of E^*.

To gain insight into transmission dynamics of the diseases with mutating strains, we simplified the model to make it more analytically tractable. By assuming the pathogen mutates with a constant rate, the PDE system is reduced into a system of ODEs. For pathogens with a constant mutation rate, we extended the local stability results for the infection-free and boundary equilibria of the ODE system, to prove that if $R_0 < 1$, E_0 is not only locally but also globally asymptotically stable. We also proved that if $R_1 < 1 < R_2$, then E_1 is globally asymptotically stable.

We established explicit conditions for the stability of the endemic equilibrium E^* when the mutation rate is constant. Furthermore, we identified the regions for the parameters where E^* loses its stability and periodic solutions bifurcate from E^*. For the special case where the two strains have the same recovery rate, we proved Hopf bifurcations using either the mutation rate, k, or the reproductive numbers, R_1 and R_2, as bifurcation parameters.

For the case where $R_1 = 3 > R_2 = 2 > 1$, we used k as a bifurcation parameter and identified regions where E_0 and E_1 are both unstable. In Example 6.1, we established a critical value, k_0, such that if $k < k_0$, the endemic equilibrium is asymptotically stable, and if $k > k_0$, the endemic equilibrium is unstable and periodic solutions appear through a Hopf bifurcation. We presented numerical simulations to illustrate that if both reproductive numbers exceed the threshold value, then the mutant cannot completely wipe out the original pathogen strain. We also showed that if the mutation rate is below the critical value, k_0, the two strains can coexist and eventually stay at a constant steady state level. On the other hand, if the mutation rate is above the critical value, k_0, there can be sustained periodic oscillations of the two pathogen strains. This phenomenon may furnish us with an interpretation of periodic appearance of pathogen strains of some diseases, such as influenza, and can provide useful guidance for disease intervention programs. Note that in this example we fixed R_2. Since R_2 is a function of the mutation rate, k, as we vary k we must also adjust the infection rate β_2 in the bifurcation analysis.

We also used R_1 and R_2 as bifurcation parameters, while fixing other parameters, including the mutation rate. Figure 1 illustrates the regions in the R_1-R_2 plane where the equilibria have different dynamics. We identified a closed bifurcation curve, L, for $R_1^{(1)} < R_1 < R_1^{(2)}$, where if R_1 and R_2 are within the curve, the endemic solution is periodic. We showed that for R_1 in the interval $(R_1^{(1)}, R_1^{(2)})$, as R_2 increases and passes through curve L, the stable steady state equilibrium changes its stability and becomes unstable. As R_2 continues to increase and passes through curve L the second time, the steady state equilibrium regains its stability. That is, the curve L identifies the parameter values where the solution undergoes a Hopf bifurcation.

Example 6.2 illustrates the Hopf bifurcation for $R_1^{(1)} < R_1 < R_1^{(2)}$. In Figure 4, (R_1, R_2) is outside the region enclosed by L with R_2 below L. In this case, the endemic equilibrium is asymptotically stable and the two strains eventually coexist at a steady state level with $I^* = 0.0035$. Figure 5 shows how when (R_1, R_2) is within the L the endemic solutions are periodic. In Figure 6, (R_1, R_2) are again outside L, but R_2 is above L. Once again, the two strains can coexist, but the steady state level $I^* = 0.02627$ is much higher than in Figure 4 because the mutant in the latter case has a larger reproductive number.

These examples illustrate the wide range of behavior that can exist when a pathogen mutates in the host to create a second infectious mutant strain. The explicit formulas for the reproductive numbers and the detailed analysis for the existence and stability of the boundary equilibrium can provide insight into the complexity of these epidemics. For the simplified cases where the mutation rate is not infection-age dependent, we were able to establish conditions for the global stability of the infection-free and boundary equilibria. Our analysis of the situation where the steady state equilibrium loses its stability through a Hopf bifurcation, and periodic solutions appear, may also help in understanding similar transitions in epidemics with mutating pathogens.

Acknowledgment. The authors thank two anonymous referees for their valuable comments and suggestions.

REFERENCES

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

[2] V. Andreasen, S. A. Levin, and J. Lin, *A model of influenza A drift evolution*, Z. Angew. Math. Mech., 76 (1996), pp. 421–424.
[3] V. Andreasen, J. Lin, and S. A. Levin, *The dynamics of cocirculating influenza strains conferring partial cross-immunity*, J. Math. Biol., 35 (1997), pp. 825–842.
[4] J. Carr, *Applications of Center Manifold Theory*, Springer-Verlag, New York, 1981.
[5] R. Cattaneo, A. Schmid, D. Eschle, K. Baczko, V. ter Meulen, and M. A. Billeter, *Biased hypermutation and other genetic changes in defective measles viruses in human brain infections*, Cell, 55 (1988), pp. 255–265.
[6] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz, *On the definition and computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations*, J. Math. Biol., 28 (1990), pp. 365–382.
[7] O. Diekmann and J. A. P. Heesterbeek, *Mathematical Epidemiology of Infectious Diseases*, John Wiley, New York, 2000.
[8] G. D. Ebel, A. P. Dupuis, II, K. Ngo, D. Nicholas, E. Kauffman, S. A. Jones, D. Young, J. Maffei, P. Y. Shi, K. Bernard, and L. D. Kramer, *Partial genetic characterization of West Nile virus strains, New York State,* 2000, Emerg. Infec. Dis., 7 (2001), pp. 650–653.
[9] J. J. Eron, P. L. Vernazza, D. M. Johnston, F. Seillier-Moiseiwitsch, T. M. Alcorn, S. A. Fiscus, and M. S. Cohen, *Resistance of HIV-1 to antiretroviral agents in blood and seminal plasma: Implications for transmission*, AIDS, 15 (1998), pp. 181–189.
[10] G. Francois, M. Kew, P. Van Damme, M. J. Mphahlele, and A. Meheus, *Mutant hepatitis B viruses: A matter of academic interest only or a problem with far-reaching implications*, Vaccine, 19 (2001), pp. 3799–3815.
[11] M. Girvan, D. S. Callaway, M. E. J. Newman, and S. H. Strogatz, *Simple model of epidemics with pathogen mutation*, Phys. Rev. E (3), 65 (2002), 031915.
[12] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.
[13] H. W. Hethcote, *The mathematics of infectious diseases*, SIAM Rev., 42 (2000), pp. 599–653.
[14] J. M. Hyman, J. Li, and E. A. Stanley, *Threshold conditions for the spread of the HIV infection in age-structured populations of homosexual men*, J. Theoret. Biol., 166 (1994), pp. 9–31.
[15] J. M. Hyman and J. Li, *An intuitive formulation for the reproductive number for the spread of diseases in heterogeneous populations*, Math. Biosci., 167 (2000), pp. 65–86.
[16] H. Inaba, *Threshold and stability for an age-structured epidemic model*, J. Math. Biol., 28 (1990), pp. 411–434.
[17] J. Lin, V. Andreasen, and S. A. Levin, *Dynamics of influenza A drift: The linear three-strain model*, Math. Biosci., 162 (1999), pp. 33–51.
[18] P. Palese and J. F. Young, *Variation of influenza A, B, and C viruses*, Science, 215 (1982), pp. 1486–1474.
[19] A. Sasaki, *Evolution of antigen drift/switching: Continuously evading pathogens*, J. Theoret. Biol., 168 (1994), pp. 291–308.
[20] S. Sato, K. Suzuki, Y. Akahane, K. Akamatsu, K. Akiyama, K. Yunomura, F. Tsuda, T. Tanaka, H. Okamoto, Y. Miyakawa, and M. Mayumi, *Hepatitis B virus strains with mutations in the core promoter in patients with fulminant hepatitis*, Ann. Internal Medicine, 122 (1995), pp. 241–248.
[21] D. J. Smith, S. Forrest, S. H. Ackley, and A. S. Perelson, *Variable efficacy of repeated annual influenza vaccination*, Proc. Natl. Acad. Sci. USA, 96 (1999), pp. 14001–14006.
[22] H. R. Thieme and C. Castillo-Chavez, *How may infection-age dependent infectivity affect the dynamics of HIV/AIDS?*, SIAM J. Appl. Math., 53 (1993), pp. 1449–1479.
[23] R. G. Webster, *Influenza*, in Emerging Viruses, S. T. Morse, ed., Oxford University Press, Oxford, UK, 1993, pp. 37–44.
[24] R. G. Webster, J. R. Schafer, J. Suss, W. J. Bean, Jr., and Y. Kawaoka, *Evolution and ecology of influenza viruses*, in Options for the Control of Influenza II, C. Hannoun et al., eds., Elsevier Science Publishers, New York, 1993, pp. 177–185.
[25] R. G. Webster, *Influenza: An emerging disease*, Emerg. Infec. Dis., 4 (1998), pp. 436–441.

SIAM J. APPL. MATH.
Vol. 65, No. 1, pp. 24–42

COMPLETE TRANSMISSION THROUGH A TWO-DIMENSIONAL DIFFRACTION GRATING*

G. A. KRIEGSMANN†

Abstract. The propagation of a normally incident electromagnetic plane wave through a two-dimensional metallic grating is modeled and analyzed. The period of the structure A is on the order of the incident wave length λ, but the height of the channel H separating the blocks is very small. Exploiting the small parameter H/A, an approximate transmission coefficient is obtained for the grating. For a fixed frequency this coefficient is $O(H/A)$ except near resonant lengths where it is $O(1)$. That is, for certain widths the structure is transparent. Similarly, for a fixed length the transmission coefficient has the same resonant features as a function of frequency.

Key words. gratings, electromagnetics, diffraction, scattering matrices, asymptotic approximations

AMS subject classifications. 35J25, 41A60, 78A45, 78M35

DOI. 10.1137/S0036139903427398

1. Introduction.

There has been considerable recent interest in the study of electromagnetic propagation through a particular grating structure [1, 2]. This structure is taken at first approximation to be an infinite slab in the X-Y plane, but with a finite thickness L in the Z-direction. A periodic array of identical holes are bored through the slab, parallel to the Z-axis. The period of the structure A in the X-Y plane is on the order of the wavelength λ of the incident electromagnetic plane wave, the holes are very small in comparison, and the thickness $L \sim \lambda$. Experimental results [1, 2] show that very little of the wave propagates through the slab. This makes sense due to the size of the holes compared to λ. However, at certain resonant frequencies, there is significant transmission. It is precisely this feature that has generated interest in this structure as an element in photonic and microwave circuits. For example, the complete transmission at select frequencies makes this grating structure useful as a highly selective filter. On the other hand, utilizing its almost complete reflection for bands of frequencies suggests that the grating can be used as a mirror in Fabry–Perot resonators.

Although the electromagnetic boundary value problem describing this physical problem is easy to state, it has no closed form solution even for simply shaped holes such as circles [1], squares, and cross-like geometries [2]. This is basically because there is a mismatch between the geometries of the hole and the fundamental cell in the X-Y plane, and the complication of requiring continuity of tangential electromagnetic fields across the hole boundary. An approximate method, based on heuristic reasoning, has been employed to study this structure and to explain the phenomenon of complete transmittance [3].

In this paper we consider a two-dimensional version of this problem in which the structure is composed of perfectly conducting metal brick cylinders separated by thin channels which take the place of the holes. A schematic of the structure is shown

*Received by the editors May 7, 2003; accepted for publication (in revised form) February 27, 2004; published electronically September 24, 2004. This work was sponsored by the National Science Foundation under grant DMS0071368 and the Department of Energy under grant DE-FG0294ER25196.
http://www.siam.org/journals/siap/65-1/42739.html
†Department of Mathematical Sciences, Center for Applied Mathematics and Statistics, New Jersey Institute of Technology, University Heights, Newark, NJ 07102 (grkrie@micro.njit.edu).

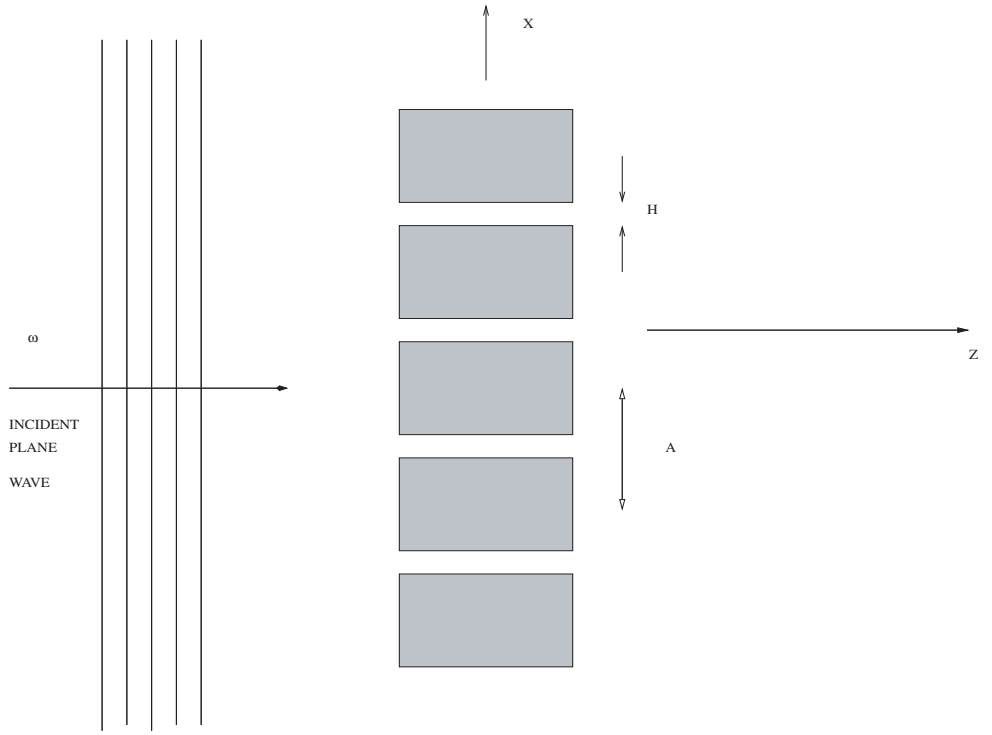


FIG. 1(a).

in Figure 1(a). The incident electromagnetic field is polarized so that the magnetic field is parallel to the Z-axis. The boundary value problem describing this physical configuration is scalar and amenable to analysis. We will show that this grating structure has the property of complete transmission at certain resonant frequencies.

We attack this problem by using S-matrix theory; we divide the problem into two pieces, analyze each separately, and then recombine the results to obtain a simple, explicit representation of the transmission coefficient. This approach is similar in spirit to the technique we employed to study large resonant structures [4]. Our analysis of the present problem shows that the transmission coefficient for the grating depends, remarkably, upon only a single real number $t(\lambda, H)$ which is a function of the wavelength of the incident plane wave and the height of the channel. This is the mathematical underpinning of the physics of the grating structure. The determination of t requires the solution of a single auxiliary scattering problem which involves the same grating structure with $L = \infty$. We solve this problem using a modal expansion procedure that is similar to the one used in [5]. It yields a single infinite system of algebraic equations. We make explicit use of the smallness of the channel height and derive an asymptotic approximation of the solution to these equations from which we deduce t. Our approximate transmission coefficient has the property seen in the three-dimensional grating structures. It is essentially zero for small channel heights, except at and near a discrete set of resonant frequencies.

A grating structure very similar to the one shown in Figure 1(a) was studied using a straightforward modal analysis [5]. By matching the modal expansions of the tangential electric and magnetic fields, which are valid inside the channel, with similar

expansions outside the channel, the authors arrived at two coupled infinite systems of linear equations whose solution yielded the required transmission and reflection coefficients. This approach can be modified to handle our grating, and the results can be simplified by considering a very narrow channel. The results of such an approach have recently been stated [6]. Another approach has been used to study a similar grating [7]. It involves a coordinate transformation, which reduces the grating to a flat surface, and a Fourier analysis of the resulting periodic Helmholtz equation. The novelty of our approach, compared to these, is its simplicity and its ability to produce a simple formula for the transmission coefficient of the grating which depends upon only $t(\lambda, H)$.

We finish this section by briefly outlining the remainder of this paper. In section 2 we formulate the scattering problem in dimensionless quantities. Section 3 contains the description of our S-matrix method. Specifically, we consider two auxiliary scattering problems and discuss their individual S-matrices. We show that both of these can be described by a single complex number τ_0 which physically is the transmission coefficient for the first auxiliary problem. Exploiting the unitary character of either S-matrix we show that τ_0 lies on a circle, C_0, in the complex plane whose center and radius are known. The number t described in the above paragraph is essentially the angle pinning τ_0 down. Using connection formulae, which ignore the effects of evanescent modes in the channel, the formula for the transmission coefficient is derived.

In section 4 the boundary value problem required to determine τ_0 is presented and a normal mode method is described to solve the problem. We derive an asymptotic approximation of the solution of these equations in the limit as $H/A \to 0$. From this result we obtain an approximation of t and hence τ_0. We find that this approximation of τ_0 lies on the circle C_0.

In section 5 we put all the ingredients together to obtain our approximation of the transmission coefficient. From this simple formula we are able to easily deduce the resonant structure of the problem. Specifically, we obtain a formula for the resonant lengths, for a fixed frequency, at which the grating is essentially transparent. We also show that there is very little transmission through the grating for other lengths, especially for $H/A \ll 1$. Similarly, when the length is fixed, we deduce a formula for the resonant frequencies of the structure. At these frequencies the grating is again transparent, and for others it is opaque.

In section 6 we briefly discuss the case when the incident wave obliquely strikes the grating. The problem is more difficult in general but asymptotically the same for $H/A \ll 1$. There is one slight modification which does not significantly alter our formulae and the resonant structure of the grating. Finally, in section 7 we offer a short conclusion and further discussion.

2. Formulation. A TM polarized electromagnetic plane wave impinges, normally, upon the periodic structure shown in Figure 1(a). The shaded regions indicate a perfectly conducting material. In all that follows the spatial variables have been scaled with respect to the period of the structure A and are denoted by lower case letters; the magnetic field P has been scaled with respect to the amplitude of the incident wave.

Since the structure is periodic, it is sufficient to study the wave propagation and scattering in the fundamental cell shown, in dimensionless form, in Figure 1(b). The magnetic field within this region satisfies the Helmholtz equation

$$\frac{\partial^2}{\partial x^2}P + \frac{\partial^2}{\partial z^2}P + k^2 P = 0, \tag{1a}$$

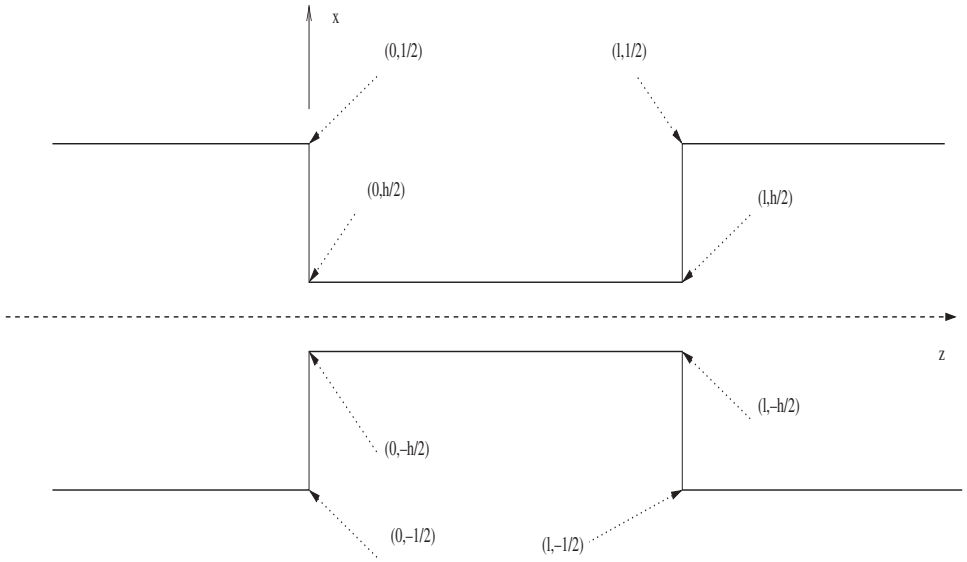


FIG. 1(b).

where $k = \omega A/c_0$, ω is the frequency in radians of the incident wave, and c_0 is the speed of light in air. The length of the structure, $l = L/A$, is order one in this scaling, and the height of the air-filled channel, $h = H/A \ll 1$, is very small. On the walls of the channel, the normal derivative of the magnetic field is zero; i.e.,

$$\frac{\partial}{\partial z} P = 0. \tag{1b}$$

On the boundaries of the fundamental cell the field is taken to periodic.

The field to the left of the structure, $z < 0$, is given by

$$P = e^{ikz}\psi_0 + \sum_{n=0}^{\infty} R_n \psi_n(x) e^{-i\beta_n z} \tag{2a}$$

and to the right of the structure, $z > l$,

$$P = \sum_{n=0}^{\infty} T_n \psi_n(x) e^{i\beta_n z}, \tag{2b}$$

where R_n and T_n are the unknown reflection and transmission coefficients, respectively. The orthonormal eigenfunctions and propagation constants are

$$\psi_0 = 1, \qquad \psi_n = \sqrt{2}\cos(2n\pi x), \quad n \neq 0, \tag{2c}$$

$$\beta_n = \sqrt{k^2 - 4n^2\pi^2}, \tag{2d}$$

where the sines have been omitted due to the symmetry of the incident wave about the z-axis. The field inside the channel is given by

$$P = \sum_{n=0}^{\infty} [A_n e^{-ik_n z} + B_n e^{ik_n z}]\phi_n, \tag{3a}$$

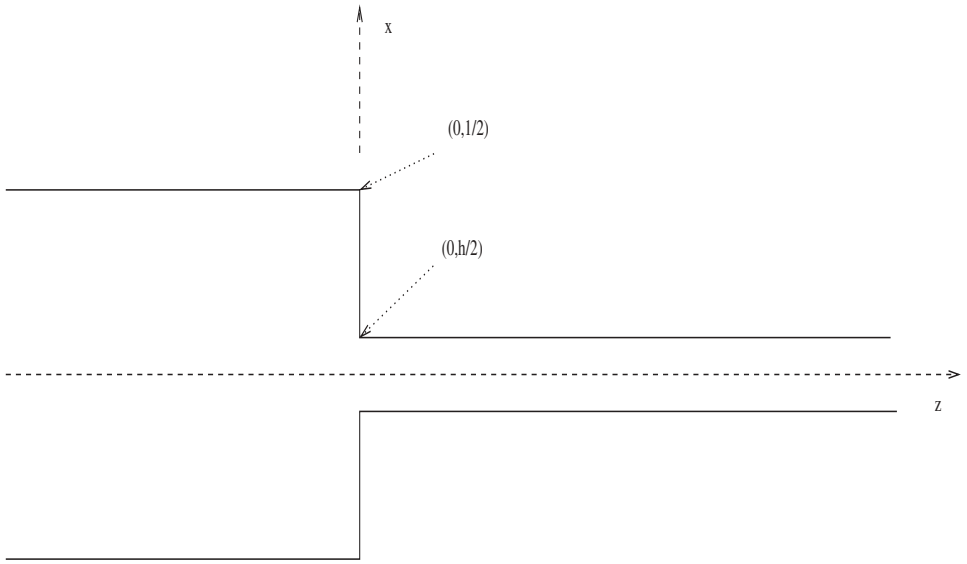


FIG. 2(a).

where the A_n and B_n are unknown constants, and the orthonormal eigenfunctions and propagation constants are

$$\phi_0 = \frac{1}{\sqrt{h}}, \qquad \phi_n = \sqrt{\frac{2}{h}} \cos \frac{2n\pi}{h} \left(x + \frac{h}{2} \right), \quad n \geq 1, \tag{3b}$$

$$k_n = \sqrt{k^2 - \frac{4n^2\pi^2}{h^2}}, \tag{3c}$$

where the sines again are omitted because of the symmetry of the incident field. It is assumed here that the wave number $k < 2\pi$ so that only β_0 is real and hence all of the other modes outside of the structure are evanescent. With this restriction on k and the fact that $h \ll 1$ it follows that only k_0 is real and all the other modes in the channel are evanescent, too. Finally, from their definitions it is clear that $\beta_0 = k_0 = k$.

3. The method. We shall begin this section by considering two auxiliary problems whose solutions can be used to construct an approximate solution to (1)–(3). These problems involve the same Helmholtz equation and boundary conditions, but now we let $l \to \infty$ so that the structure becomes semi-infinite. Accordingly, the fundamental cell becomes semi-infinite; its geometry is shown in Figure 2(a). These problems are two-dimensional versions of the three-dimensional acoustic cases studied in [8].

The auxiliary problems. In the first problem we consider an incident mode of unit strength striking this structure from the left; this corresponds to a normally incident plane wave. The field is given there by

$$P_1 = e^{ikz}\psi_0 + \sum_{n=0}^{\infty} r_n \psi_n(x) e^{-i\beta_n z}, \quad z < 0. \tag{4a}$$

The resulting field within the channel is

$$P_1 = \sum_{n=0}^{\infty} \tau_n \phi_n(x) e^{ik_n z}, \quad z > 0, \tag{4b}$$

where there are only modes propagating to the right. The determination of r_n and τ_n must be determined approximately by a numerical method; we shall describe a particular method in section 4.

At this stage we shall quickly derive a relationship between r_0 and τ_0 which will become very useful. It is

$$1 - r_0 = \sqrt{h}\,\tau_0. \tag{5}$$

This is derived by taking the partial derivative of (4a) with respect to z at $z = -\delta$, where $\delta \ll 1$, i.e., just to the left of the channel. Then multiplying this result by ψ_0, integrating with respect to x, and using the orthonormality of the eigenfunctions, we obtain

$$\int_{-1/2}^{1/2} \frac{\partial}{\partial z} P_1(x, -\delta)\psi_0 \, dx = ik(1 - r_0)e^{ik\delta}. \tag{6a}$$

A similar calculation applied to (4b) at $z = +\delta$ yields

$$\int_{-h/2}^{h/2} \frac{\partial}{\partial z} P_1(x, \delta)\phi_0 \, dx = ik\tau_0 e^{ik\delta}. \tag{6b}$$

We now let $\delta \to 0$ in (6a) and recall from (1b) that $P_{1z} = 0$ on the metal portion of the structure. This implies that the range of integration is now $(-h/2, h/2)$ in (6a). The exponential on the right-hand side goes to one. A similar limiting process applied to (6b) removes the exponential from its right-hand side. Since P_{1z} is continuous at $z = 0$ within the channel and $\phi_0 = \psi_0/\sqrt{h}$, (6a) and (6b) can be combined to arrive at (5).

In the second auxiliary problem we consider an incident mode of unit strength striking the structure from the right. This corresponds to an incident mode in the channel. The field is given by

$$P_2 = \sum_{n=0}^{\infty} \gamma_n \psi_n(x) e^{-i\beta_n z}, \quad z < 0, \tag{7a}$$

and

$$P_2 = e^{-ikz}\phi_0(x) + \sum_{n=0}^{\infty} \rho_n \phi_n(x) e^{ik_n z}, \quad z > 0. \tag{7b}$$

There is an analogous relation between γ_0 and ρ_0 that is very similar to (5) and is obtained along parallel lines of reasoning. It is

$$\sqrt{h}\,\rho_0 + \gamma_0 = \sqrt{h}. \tag{8}$$

The determination of ρ_0 and γ_0 must be found numerically.

There is a final and very useful relationship between γ_0 and τ_0; it is simply

$$\gamma_0 = \tau_0. \tag{9}$$

This follows from integrating the identity $\nabla\cdot\{P_1\nabla P_2 - P_2\nabla P_1\} = 0$ in the region $|z| < z_\infty$, applying the divergence theorem, using the boundary conditions, and neglecting the evanescent modes. Here $z_\infty \gg 1$ is a fixed number.

Finally, we note that the relationships contained in (5), (8), and (9) can be combined to express r_0, ρ_0, and γ_0 in terms of τ_0. Omitting the algebraic details, we list them here as

$$r_0 = 1 - \sqrt{h}\tau_0, \quad \gamma_0 = \tau_0, \tag{10a}$$

$$\rho_0 = 1 - \frac{\tau_0}{\sqrt{h}}. \tag{10b}$$

Thus, only the first auxiliary problem needs to be solved numerically to determine the reflection and transmission coefficients, of the propagating mode, for the second problem.

The S-matrix. The results of the proceeding subsection can be combined into two simple linear statements that constitute classical S-matrix theory used to characterize microwave circuits [9]. Let the semi-infinite structure, shown in Figure 2(a), be excited by a normally incident plane wave of strength a_0 from the left and by the lowest mode with strength c_0 from the right. A few wavelengths to the left of $z = 0$, where the evanescent modes may be neglected, the field is given by

$$P = \{a_0 e^{ikz} + b_0 e^{-ikz}\}\psi_0(x) \tag{11a}$$

and a few wavelengths to the right of $z = 0$ by

$$P = \{d_0 e^{ikz} + c_0 e^{-ikz}\}\phi_0(x). \tag{11b}$$

By the linearity of the Helmholtz equation and the boundary conditions (1), it follows that

$$b_0 = r_0\, a_0 + \gamma_0\, c_0, \tag{12a}$$

$$d_0 = \tau_0\, a_0 + \rho_0\, c_0. \tag{12b}$$

These can be rewritten with the aid of (10) as

$$b_0 = \left(1 - \sqrt{h}\tau_0\right)\, a_0 + \tau_0\, c_0, \tag{13a}$$

$$d_0 = \tau_0\, a_0 + \left(1 - \frac{\tau_0}{\sqrt{h}}\right)\, c_0. \tag{13b}$$

The coefficients in (13) form the S-matrix for our semi-infinite structure. We show in Appendix A that this matrix is unitary and, accordingly, that τ_0 must lie on the circle

$$\left|\tau_0 - \frac{\sqrt{h}}{1+h}\right| = \frac{\sqrt{h}}{1+h} \tag{14a}$$

in the complex plane, which has the equivalent representation

$$\tau_0 = \frac{2\sqrt{h}}{(1+h)+i\eta}, \quad -\infty < \eta < \infty. \tag{14b}$$

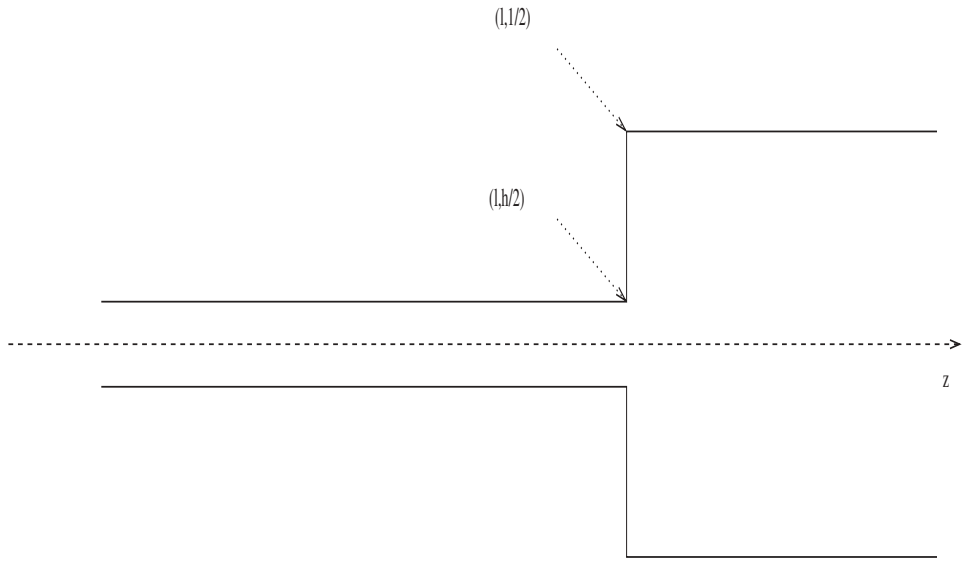


FIG. 2(b).

The real number η depends upon h and k: Remarkably, all of the scattering physics of our semi-infinite structure is contained in this number.

Finally, we state the S-matrix for the related structure shown in Figure 2(b), which is just the semi-infinite structure of Figure 2(a) shifted l units along the z-axis and reflected. If $\hat{c}_0$ is the strength of the incident mode to the left of $z = l$ and $\hat{a}_0$ is the amplitude of the plane wave striking from the right, then P is given by

$$P = \left\{ \hat{a}_0 e^{-ikz} + \hat{b}_0 e^{ikz} \right\} \psi_0(x) \tag{15a}$$

a few wavelengths to the right of $z = l$, and by

$$P = \left\{ \hat{d}_0 e^{-ikz} + \hat{c}_0 e^{ikz} \right\} \phi_0(x) \tag{15b}$$

a few to the left. The linear relationship analogous to (13) is

$$\hat{b}_0 = \left(1 - \sqrt{h}\tau_0\right) e^{-2ikl}\, \hat{a}_0 + \tau_0\, \hat{c}_0, \tag{16a}$$

$$\hat{d}_0 = \tau_0\, \hat{a}_0 + \left(1 - \frac{\tau_0}{\sqrt{h}}\right) e^{2ikl}\, \hat{c}_0, \tag{16b}$$

where the exponential factors take into account the shifted and reflected location of the channel.

Connection formulae. We shall now return to our original problem and begin the construction of its approximate solution. First, we reinterpret the modal structure of the field given by (3a) at the hypothetical plane $z = l/2$. At this plane only the propagating mode is present. This occurs because the evanescent modes are heavily damped, due to the fact that their propagation constants are purely imaginary and large in moduli; i.e., $\beta_m \sim \mathrm{Im}\pi/h$ for $m \geq 1$. For example, the largest evanescent

mode occurs when $m = 1$, and it is $O(e^{-\pi l/2h})$ as $h \to 0$. Thus in the vicinity of the hypothetical plane at $z = l/2$, we neglect these exponentially small terms and take

$$P = \{A_0 e^{-ikz} + B_0 e^{ikz}\}\phi_0. \tag{17}$$

From the perspective of an observer near the aperture at $z = 0$, a mode of unit strength impinges from the left and a mode of strength A_0 impinges from the right. The wave reflected to the left has an amplitude R_0 and to the right the amplitude is B_0. These amplitudes are connected via the S-matrix (13). Setting $a_0 = 1$, $b_0 = R_0$, $c_0 = A_0$, and $d_0 = B_0$, (13) becomes

$$R_0 = \left(1 - \sqrt{h}\tau_0\right) + \tau_0 A_0, \tag{18a}$$

$$B_0 = \tau_0 + \left(1 - \frac{\tau_0}{\sqrt{h}}\right) A_0. \tag{18b}$$

From the perspective of an observer near the aperture at $z = l$, there is no mode impinging from the right, just a mode of strength B_0 striking from the left. The wave transmitted to the right has an amplitude T_0 and reflected to the left A_0. These amplitudes are connected via the S-matrix (16). Setting $\hat{a}_0 = 0$, $\hat{b}_0 = T_0$, $\hat{c}_0 = B_0$, and $\hat{d}_0 = A_0$, (16) becomes

$$T_0 = \tau_0 B_0, \tag{19a}$$

$$A_0 = \left(1 - \frac{\tau_0}{\sqrt{h}}\right) e^{2ikl} B_0. \tag{19b}$$

Inserting (19b) into (18b) and solving for B_0 yields

$$B_0 = \frac{\tau_0}{1 - \left(1 - \frac{\tau_0}{\sqrt{h}}\right)^2 e^{2ikl}}. \tag{20}$$

Combining this result with (19a) gives

$$T_0 = \frac{\tau_0^2}{1 - \left(1 - \frac{\tau_0}{\sqrt{h}}\right)^2 e^{2ikl}}, \tag{21a}$$

the transmission coefficient for our periodic structure. Finally, inserting (20) into (19b) and the resulting expression into (18a) yields

$$R_0 = \left(1 - \sqrt{h}\tau_0\right) + T_0 \left(1 - \frac{\tau_0}{\sqrt{h}}\right) e^{2ikl}, \tag{21b}$$

the reflection coefficient of our periodic structure.

In closing this section it is interesting to note that the reflection and transmission coefficients for our periodic structure are determined completely by τ_0, the transmission coefficient for our first auxiliary problem. Remarkably then by (14b), the scattering properties of our grating are completely contained in the parameter η. All that is left now is to find an accurate approximation of this number.

4. An approximation of τ_0. The solution of the first auxiliary problem and the determination of τ_0 cannot be determined exactly; a numerical approximation must be sought. The approach we use here employs a Green's function argument and a normal mode expansion, using the ϕ_n as a basis, to obtain a system of equations for the τ_n. We then exploit the fact that $h \ll 1$ to obtain an approximation of τ_0.

We begin by representing the field P_1 in the region $z < 0$ by $P_1 = 2\cos(kz) + P_S$, where the first term incorporates the incident field and its rigid reflection and ignores the presence of the channel. The scattered field P_S takes this into account. It is clear from this representation that $\frac{\partial}{\partial z}P_1 = \frac{\partial}{\partial z}P_S$ on $z = 0$. It then follows from (1a) that $\frac{\partial}{\partial z}P_S = 0$ for $z = 0$ with $|x| > h/2$. Using this information we deduce, using standard Green's functions arguments, that

$$P_1(x,z) = 2\cos(kz) - \int_{-h/2}^{h/2} G(x,z|x',0)\,\frac{\partial}{\partial z'}P_1(x',0)\,dx', \tag{22a}$$

$$G(x,z|x',0) = \sum_{n=-\infty}^{\infty} \frac{1}{i\beta_n} e^{2n\pi i(x-x')}\, e^{-i\beta_n z}. \tag{22b}$$

Here G is the Green's function for the Helmholtz equation in $z < 0$ which is periodic in x, and satisfies $\frac{\partial}{\partial z'}G = 0$ when $z' = 0$.

We next set $z = 0$ in (22) and use the representation of P_1 given by (4b) in the resulting expression. This can be done because both the field and its z derivative are continuous across $z = 0$ with $|x| < h/2$. The result is

$$\sum_{m=0}^{\infty} \tau_m\phi_m = 2 - i\int_{-h/2}^{h/2} G_0(x,0|x',0) \sum_{m=0}^{\infty} k_m\tau_m\phi_m(x')\,dx'. \tag{23}$$

Multiplying this expression by ϕ_n, integrating between $\pm h/2$, and using the orthonormality of the $\{\phi_n\}$, we arrive at the linear system

$$\tau_n = 2\sqrt{h}\,\delta_{n0} - i\sum_{m=0}^{\infty} k_m\tau_m Z_{nm}, \tag{24a}$$

$$Z_{nm} = \int_{-h/2}^{h/2}\int_{-h/2}^{h/2} G(x,0|x',0)\phi_n(x)\phi_m(x')\,dx'\,dx, \tag{24b}$$

where δ_{n0} is the Kronecker delta function. Before proceeding to exhibit the formulae for the Z_{nm} we rewrite the system (24a) in two pieces, singling out the leading order coefficient τ_0:

$$\tau_0 = 2\sqrt{h} - ikZ_{00}\tau_0 + \sum_{m=1}^{\infty} |k_m| Z_{0m}\tau_m, \tag{25a}$$

$$\tau_n = -ikZ_{n0}\tau_0 + \sum_{m=1}^{\infty} |k_m| Z_{nm}\tau_m. \tag{25b}$$

Here we have recalled and used the fact that $k_m = i|k_m| \equiv i\sqrt{m^2\pi^2/h^2 - k^2}$ for $m \geq 1$.

We now make the substitution $\tau_n = -ik\tau_0\alpha_n$, for $n \geq 1$, into (25b) and find that the α_n must satisfy

$$\alpha_n = Z_{n0} + \sum_{m=1}^{\infty} |k_m| Z_{nm}\alpha_m, \quad n \geq 1. \tag{26a}$$

Once this system is solved for the α_n, then the τ_n are known for $n \geq 1$. Making the same substitution into (25a) and solving for τ_0 we find that

$$\tau_0 = \frac{2\sqrt{h}}{1 + ikZ_{00} + ik\sum\limits_{m=1}^{\infty} |k_m| Z_{m0}\,\alpha_m}. \tag{26b}$$

Thus, once the α_m are known, τ_0 is known, too.

The matrix elements are obtained by substituting (22b) into (24b) and interchanging the order of summation and integration. The resulting integrals are elementary, and we find that

$$Z_{00} = -i\frac{h}{k} - 2hS_{00}, \tag{27a}$$

$$Z_{n0} = Z_{0n} = 2hS_{0n}, \quad n \geq 1, \tag{27b}$$

$$Z_{nm} = hS_{nm}, \quad n, m \geq 1, \tag{27c}$$

where the S_{nm} are real and are defined by

$$S_{00} = \sum_{l=1}^{\infty} \frac{1}{|\beta_l|} \frac{\sin^2(l\pi h)}{(l\pi h)^2}, \tag{28a}$$

$$S_{0n} = \sqrt{2}\sum_{l=1}^{\infty} \frac{1}{|\beta_l|} \frac{\sin^2(l\pi h)}{[(n\pi)^2 - (l\pi h)^2]}, \quad n \geq 1, \tag{28b}$$

$$S_{nm} = -4\sum_{l=1}^{\infty} \frac{(l\pi h)^2 \sin^2(l\pi h)}{|\beta_l|[n^2\pi^2 - (lh\pi)^2][m^2\pi^2 - (lh\pi)^2]}, \quad m, n \geq 1. \tag{28c}$$

Finally, inserting (27a) and (27b) into (26b) we obtain the result

$$\tau_0 = \frac{2\sqrt{h}}{(1+h) + ih\,t}, \tag{29a}$$

$$t = 2\,k\left\{-S_{00} + \sum_{m=1}^{\infty} |k_m| S_{0m}\alpha_m\right\}. \tag{29b}$$

Thus, τ_0 satisfies (14b) with $\eta = h\,t$.

Now any approximation of τ_0 using the above formulation requires an approximation of the α_n. This is done by first truncating the infinite system (26a), so that

$1 \le n \le N$. Then for a fixed h and k the series defining the Z_{nm}, (28a)–(28c), can be truncated to yield a prescribed accuracy. Once this is done we can invert the finite system to determine the approximations of α_n , $1 \le n \le N$. These will be denoted by $\hat{\alpha}_n$. Next, the infinite sum in (29b) is likewise truncated, so that $1 \le m \le N$. We denote this approximation by $\hat{t}$. Finally, we insert this result into (29a) and obtain our approximation of the transmission coefficient

$$\hat{\tau}_o = \frac{2\sqrt{h}}{(1+h) + i\,h\,\hat{t}}. \tag{30}$$

Remarkably, this approximation satisfies (14b) regardless of N and the errors induced by truncating the series.

Up to this point we have not exploited the small channel width, $h \ll 1$, except in neglecting the exponentially small contributions of the evanescent modes to (17). From (27b) and (27c) we observe that the matrix elements Z_{nm}, $n \ge 1$ and $m \ge 0$, are formally $O(h)$ as $h \to 0$. To show this is true we need to demonstrate that the corresponding S_{nm} are $O(1)$. These sums are slowly convergent when $h \ll 1$. However, careful numerical computations on (28b) and (28c) show that they are order one, and a qualitative argument to this effect is presented in Appendix B.

Using the facts that Z_{n0} and Z_{nm} are small, it follows from (26a), truncated at $m = N$, that

$$\hat{\alpha}_n = 2\,h\,S_{n0} + O(h^2), \quad 1 \le n \le N, \tag{31a}$$

as $h \to 0$. Inserting this estimate into (29b), also truncated at $m = N$, and noting that $|k_m| \sim m\pi/h$, we deduce that

$$\hat{t} = 2k\left\{-S_{00} + 4\,\pi \sum_{m=1}^{N} m\,S_{0m}^2\right\}. \tag{31b}$$

Now the sum (28a) defining S_{00} is also slowly convergent, and we present a qualitative argument in Appendix B that $S_{00} \sim \ln(1/h)$; careful numerical computations bear this out, too. This suggests that the first term in (31b) is dominant and accordingly

$$\hat{t} = -2\,k\,S_{00}. \tag{31c}$$

Inserting (31c) into (30) yields our approximation $\hat{\tau}_0$ of the transmission coefficient of our first auxiliary problem.

The above approximations of $\hat{\alpha}_n$ and $\hat{t}$ are formal. To make them rigorous we must prove that the $O(h^2)$ estimate in (31a) remains valid as $N \to \infty$. Moreover, we must show $S_{0m} = O(1/m^p)$, where $p > 1$, so that the sum in (31b) converges in this limit. Of these two problems the second is intuitive due to the presence of $n^2\pi^2$ in the denominator of (28b). The first problem is not intuitive as it requires an estimate of the effects of the terms $\infty > m > N$. This is a difficult issue which we do not address here. We have, however, done careful numerical experiments on (26a) and the effects of truncation. We have found that for $h < 0.25$ and $0 < k < 2\pi$, choosing $N \ge 5$ gives accurate approximations of the α_n to six decimal places. Increasing N further changed the results in the higher decimal places. Furthermore, the estimate (31c) agreed to a similar degree of accuracy with (31b).

5. Transparency. All the ingredients are now in place to determine an approximation of the transmission coefficient T_0 for our periodic structure. This is done by inserting (30), with $\hat{t}$ given by (31c), into (21a); we denote the result by $\hat{T}_0$. The magnitude of this complex number is found, after some algebraic and trigonometric simplifications, to be

$$|\hat{T}_0| = \frac{4h}{\sqrt{16h^2A_1^2 + 4A_2^2}}, \tag{32a}$$

$$A_1 = \cos kl - 2kh^2 S_{00} \sin kl, \tag{32b}$$

$$A_2 = (1 + h^2 - 4k^2h^2S_{00}^2)\sin kl + 2khS_{00}\cos kl. \tag{32c}$$

Now if A_1 and A_2 are both $O(1)$, then $|T_0| = O(h)$. Since the channel width $h \ll 1$, the field will be quite small in the region $z > l$. On the other hand, if A_2 is small, then $|T_0|$ may be order one and the field will be substantial in $z > l$. In fact, for a fixed k and h we can force A_2 to be zero by choosing the length of our structure to satisfy

$$\tan kl = \frac{-2khS_{00}}{1 + h^2(1 - 4k^2S_{00}^2)}. \tag{33}$$

The approximate solution to this equation is

$$l = \frac{M\pi}{k} - 2hS_{00} + O(h^2\, S_{00}^2), \tag{34a}$$

as $h \to 0$, where M is an integer and the $O(h^2S_{00}^2)$ represents the error which is very small, even though $S_{00} = O(\ln(1/h))$, as $h \to 0$. For these choices of resonant l we find that $|\hat{T}_0| = 1 - O(h^2\, S_{00}^2)$; that is, our periodic structure is virtually transparent.

In Figure 3(a) we have plotted $|\hat{T}_0|$ as a function of l for $k = \pi/2$ and $h = 0.1$. The peaks of this function are slightly less than 1, and occur for values of l slightly less than $2, 4, 6$, and 8, completely in agreement with (34a). The value of this function away from these resonant lengths is relatively small and clearly on the order of h. Figure 3(b) contains the plot of $|\hat{T}_0|$ as a function of l for $k = \pi/2$ and $h = 0.01$. For this smaller channel width the peaks of this function are much more localized and occur even closer to $l = 2, 4, 6$, and 8. Furthermore, away from these peaks it is again on the order of h which is now even smaller.

Now for a fixed l we can rewrite (34a), neglecting the error term, to obtain

$$k = \frac{M\pi}{l + 2h\, S_{00}(k, h)}, \tag{34b}$$

where the dependence of S_{00} on k and h is explicitly shown. This is an implicit equation for the resonant k for our grating. However, the term involving hS_{00} is small so that the resonant k will be slightly less than $M\pi/l$. This is borne out in Figure 4(a), where $|\hat{T}_0|$ is plotted as a function of k for $h = 0.01$ and $l = 2$. Here we have restricted k to the interval $(0, 2\pi)$ to ensure single mode propagation in the regions $z < 0$ and $z > l$. We note that if l is increased from 2 to 4, then the function $|\hat{T}_0|$ will have 8 resonant spikes. This follows from (34b) and implies that the grating

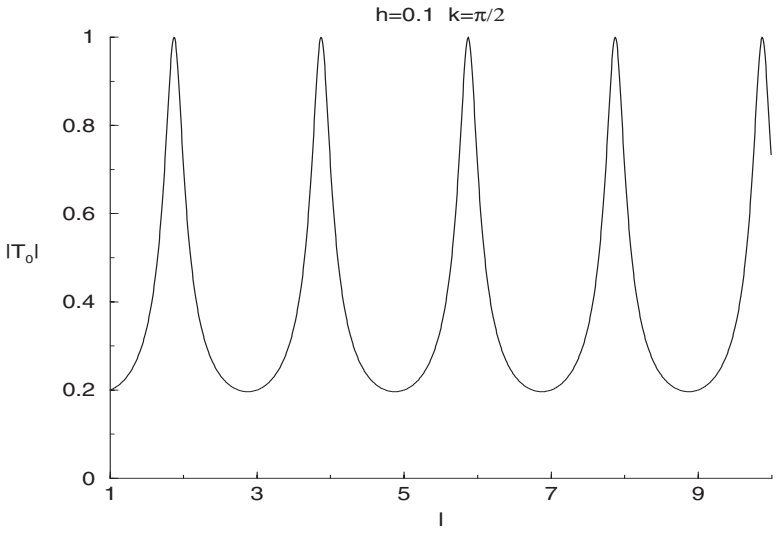


FIG. 3(a).

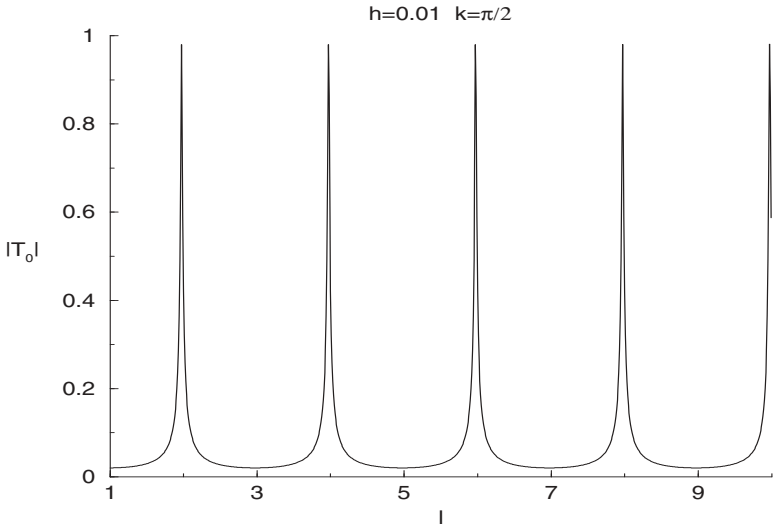


FIG. 3(b).

will be transparent at many frequencies as l increases, or equivalently as the structure becomes wider. We also observe that the pass bands for this grating structure are extremely localized about $k = k_M \sim M\pi/l$. This localization is mitigated somewhat when losses in the metallic grating are taken into account. Nonetheless this resonant structure suggests that the grating might be suitable as a highly selective filter.

In Figure 4(b) we plot $\hat{R}_0$ as a function of k for $h = 0.01$ and $l = 2$. This approximate reflection coefficient is obtained by combining (21a), (21b), (30), and (31c). It is evident from this plot that the grating stop bands are very broad, in the absence of grating losses. It is this feature of the grating that might make it a candidate for a lossy mirror in a Fabry–Perot resonator.

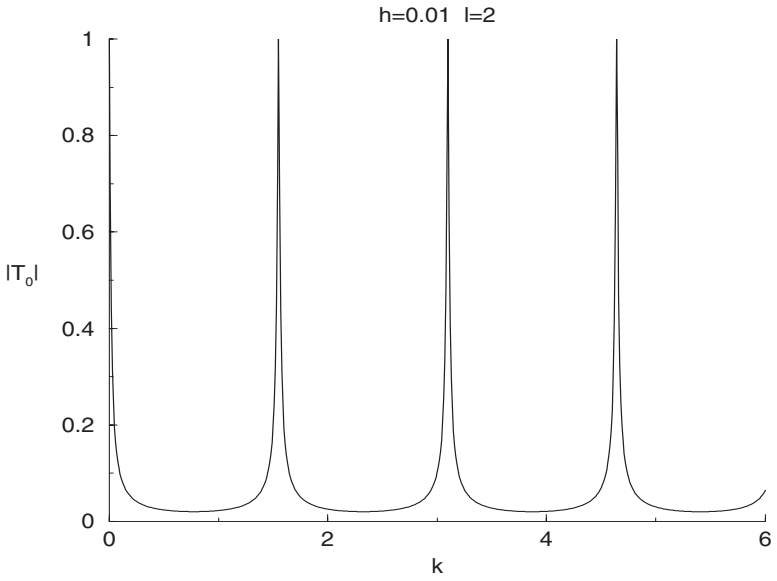


FIG. 4(a).

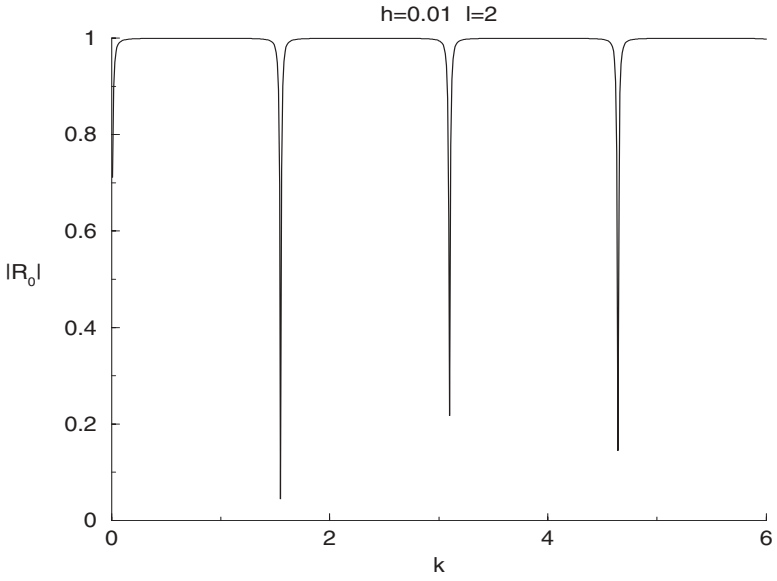


FIG. 4(b).

Finally, we offer a comparison of our results with those given in [6], where the transmission coefficient was determined from a full-modal solution of the entire grating problem. Using $A = 3.5\,\mu m$, $L = 4\,\mu m$, and $H = 0.5\,\mu m$, we find that $l = 1.14$ and $h = 1/7$. The full-modal solution predicted resonances at $4.9\,\mu m$ and $9.7\,\mu m$ which correspond to $k = 4.3$ and $k = 2.3$, respectively. Our approximation transmission coefficient for this case has resonances at $k = 2.4$ and $k = 4.7$ which differ by 4% and 8%. These errors are completely consistent with our asymptotic theory, which is accurate to $O(h)$.

6. Nonnormal incidence. The procedure developed and discussed in the previous sections can be modified to handle the case of an incident plane wave

$$P_{inc} = e^{ik(x\sin\theta_0 + z\cos\theta_0)}, \tag{35}$$

where θ_0 is the angle the wave makes with the positive z-axis. For an arbitrary channel height the simple results given in (10) are no longer true. Higher order reflection and transmission coefficients must be taken into account in these expressions. For example, the simple relationship $r_0 = 1 - \sqrt{h}\,\tau_0$ now becomes

$$r_0 = 1 - \sqrt{h}\tau_0 \operatorname{sinc}\left(\frac{kh\sin\theta_0}{2}\right) - \frac{i2\sqrt{h}}{k}\sum_{n=1}^{\infty} |k_n|\tau_n I_n \tag{36a}$$

where

$$\operatorname{sinc}(x) = \frac{\sin x}{x}, \qquad I_n = \int_0^1 e^{-ikht\sin\theta_0}\cos n\pi t\, dt. \tag{36b}$$

Thus, our simple theory based upon the single scalar number t no longer holds. However, if we exploit the fact that $h \ll 1$, then the sinc function and the I_n in (36a) all become $1 + O(h)$ and we obtain $r_0 = 1 - \sqrt{h}\tau_0 + O(h^{3/2})$. Ignoring the correction term we arrive at our original relationship. Similar reasoning shows that the other statements in (10) hold asymptotically and the rest of our argument proceeds accordingly until we arrive at the approximation of τ_0. A modified Green's function must be introduced at this point which takes into account the quasi-periodic nature of the solution in x. The analysis here follows along the same path and the results are the same with one exception. The formula for S_{00} must be modified by replacing $|\beta_l|$ by $\sqrt{(2l\pi - k\sin\theta_0)^2 - k^2}$ which takes into account the nonnormal incidence. The formulae (34a) and (34b) for the resonant length and wave number, respectively, still remain valid with this change.

7. Conclusion. We have derived a simple approximation of the transmission coefficient for the two-dimensional periodic structure shown in Figure 1. The assumptions used in our analysis were threefold. First, the incident wave was taken to be normally incident upon the structure. Second, the frequency of the incident wave was restricted to ensure that only a single mode propagated in the regions $z < 0$ and $z > l$. Third, the air filled channels were taken to be very small compared to the period of the structure. Our results show that for a fixed frequency and channel width h, the magnitude of the transmission coefficient is $O(h)$ except at certain resonant lengths where it is almost 1. Moreover, the smaller h is made, the more localized is this resonant behavior. The same phenomenon occurs when the length of the structure and the channel width are fixed. Then the magnitude of the transmission coefficient is $O(h)$ except in the neighborhood of resonant frequencies. These phenomena will be mitigated somewhat in actual gratings with finite conductivity. Nonetheless, our approximate transmission coefficient still gives accurate estimates for resonant lengths and the resonant frequencies for highly conductive gratings.

Our approximate transmission coefficient depends fundamentally upon τ_0, the transmission coefficient of a related auxiliary problem. This coefficient is deduced from the solution to an infinite system of linear algebraic equations. We have developed an asymptotic approximation of the solution of these equations, and hence of τ_0, in the limit $h \to 0$. Although these calculations were formal, careful numerical studies of the

system and of the resulting approximation of τ_0 strongly indicate that our asymptotic results are highly accurate.

We have also briefly addressed the case of nonnormal incidence and have provided the necessary changes. Finally, we note that removing the restriction on k so that more modes are present in the region exterior to our structure will cause the analyses to become much more involved. The dependence of the scattering physics on one number, regardless of the small channel height, will no longer be true. However, we conjecture that the resonant phenomena illustrated in this present work still will remain.

Appendix A. In this appendix we show that the matrix given in (13) is unitary and, accordingly, that τ_o satisfies (14a). We begin by letting P denote the solution to the Helmholtz equation which satisfies the boundary conditions enumerated in the text and behaves according to (11) away from $z = 0$. It follows from the former that $\mathrm{Im}\{\nabla \cdot \bar{P}\nabla P)\} = 0$. Integrating this relationship in the region $|z| < z_\infty$, applying boundary conditions, and using (11) we obtain

$$l|a_0|^2 + |c_0|^2 = |b_0|^2 + |d_0|^2. \tag{A.1}$$

This is a statement of the conservation of power. We now insert the expressions for b_0 and d_0 given by (13) into (A.1) and obtain

$$\begin{aligned} 0 = \{|1-\sqrt{h}\tau_0|^2 + |\tau_0|^2 - 1\}|a_0|^2 + \left\{\left|1-\frac{\tau_0}{\sqrt{h}}\right|^2 + |\tau_0|^2 - 1\right\}|c_0|^2 \\ +2\,\mathrm{Re}\left[\left\{\tau_0\left(1-\sqrt{h}\bar{\tau}_0\right) + \bar{\tau}_0\left(1-\frac{\tau_0}{\sqrt{h}}\right)\right\}\right]\bar{a}_0 c_0. \end{aligned} \tag{A.2}$$

Setting $a_0 = 1$ and $c_0 = 0$, we deduce from (A.2) that $|1-\sqrt{h}\tau_0|^2 + |\tau_0|^2 = 1$. This can be rearranged to give (14a). Similarly, setting $a_0 = 0$ and $c_0 = 1$ gives $|1-\frac{\tau_0}{\sqrt{h}}|^2 + |\tau_0|^2 = 1$. This too can be rearranged to give (14a). Using these two pieces of information it follows from (A.2) that $2\,\mathrm{Re}[\{\tau_0(1-\sqrt{h}\bar{\tau}_0) + \bar{\tau}_0(1-\frac{\tau_0}{\sqrt{h}})\}] = 0$. These results show that the matrix defined in (13) is unitary.

Appendix B. In this appendix we shall give a qualitative argument that S_{0n}, defined by (28b), is $O(1)$ as $h \to 0$. That is, S_{0n} is bounded in this limit. An analogous argument for S_{nm} has been developed, but for brevity will not be presented here.

We begin by rewriting (28b) as

$$S_{0n} = \frac{\sqrt{2}}{\pi}\sum_{l=1}^{\infty} \frac{\Delta x}{\sqrt{(2l\Delta x)^2 - (k\Delta x)^2}} \frac{\sin^2(l\Delta x)}{[(n\pi)^2 - (l\Delta x)^2]},$$

where $\Delta x = h\pi$. Now for small Δx the sum above can be interpreted as the Riemann sum which approximates the integral

$$I_{0n} = \frac{\sqrt{2}}{\pi}\int_{\Delta x}^{\infty} \frac{1}{\sqrt{4x^2 - (k\Delta x)^2}} \frac{\sin^2 x}{[(n\pi)^2 - x^2]}\, dx.$$

This integral can be split into three pieces:

$$I_{0n} = \frac{\sqrt{2}}{\pi}\{J_1 + J_2 + J_3\},$$

where

$$J_1 = \int_{k\Delta x/2}^{1} \frac{\sin^2 x}{\sqrt{4x^2 - (k\Delta x)^2}\,[(n\pi)^2 - x^2]}\,dx,$$

$$J_2 = -\int_{k\Delta x/2}^{\Delta x} \frac{\sin^2 x}{\sqrt{4x^2 - (k\Delta x)^2}\,[(n\pi)^2 - x^2]}\,dx,$$

$$J_3 = \int_{1}^{\infty} \frac{\sin^2 x}{\sqrt{4x^2 - (k\Delta x)^2}\,[(n\pi)^2 - x^2]}\,dx.$$

Now J_1 and J_3 can be approximated to $O(\Delta x)$ by replacing Δx by 0, yielding

$$J_1 \sim \int_{0}^{1} \frac{\sin^2 x}{2x\,[(n\pi)^2 - x^2]}\,dx,$$

$$J_3 \sim \int_{1}^{\infty} \frac{\sin^2 x}{2x\,[(n\pi)^2 - x^2]}\,dx.$$

Both of these integrals are finite. After making the substitution $x = (k\Delta x/2)p$ and using $\Delta x \ll 1$ the integral defining J_2 becomes

$$J_2 \sim -\left(\frac{k\Delta x}{2n\pi}\right)^2 \int_{1}^{2/k} \frac{p^2}{\sqrt{p^2 - 1}}\,dp.$$

Combining these three estimates we have

$$S_{0n} \sim I_{0n} = \frac{\sqrt{2}}{\pi}\{J_1 + J_2 + J_3\} = O(1)$$

as $h \to 0$.

Using similar reasoning on S_{00} we obtain

$$S_{00} \sim \frac{1}{\pi}\{K_1 + K_2 + K_3\},$$

where

$$K_1 = \int_{\Delta x}^{1} \frac{\sin^2 x}{\sqrt{4x^2 - (k\Delta x)^2}}\,\frac{dx}{x^2},$$

$$K_2 = \int_{k\Delta x/2}^{\Delta x} \frac{\sin^2 x}{\sqrt{4x^2 - (k\Delta x)^2}}\,\frac{dx}{x^2},$$

$$K_3 = \int_{1}^{\infty} \frac{\sin^2 x}{\sqrt{4x^2 - (k\Delta x)^2}}\,\frac{dx}{x^2}.$$

Now K_3 can be approximated to $O(h)$ by replacing Δx by 0, giving

$$K_3 = \int_1^\infty \frac{\sin^2 x}{2x^3}\, dx,$$

which is finite. Making the same substitution, as above in J_2, into K_2 we deduce that

$$K_2 = \int_1^{2/k} \frac{1}{\sqrt{p^2-1}}\, dp,$$

which is also finite. Now the integral defining K_1 is rewritten as

$$K_1 = \int_{\Delta x}^1 \frac{dx}{\sqrt{4x^2-(k\Delta x)^2}} - \int_{\Delta x}^1 \left(\frac{\sin^2 x}{x^2} - 1\right) \frac{dx}{\sqrt{4x^2-(k\Delta x)^2}}.$$

We observe that the second integral is bounded as $\Delta x \to 0$ and it can be approximated, to $O(\Delta x)$, by just setting $\Delta x = 0$. The first integral can be solved exactly and as $\Delta x \to 0$ it becomes $\frac{1}{2}\ln(\frac{4}{k\Delta x})$. Thus, we finally have

$$S_{00} \sim \frac{1}{2\pi} \ln(4/kh).$$

REFERENCES

[1] T. W. Ebbesen, H. J. Lezek, H. F. Ghaemi, T. Thio, and P. A. Wolff, *Extraordinary optical transmission through sub-wavelength hole arrays*, Nature, 391 (1998), pp. 667–669.

[2] A. P. Hibbins and J. R. Sambles, *Remarkable transmission of microwaves through a long wall of metallic bricks*, Appl. Phys. Lett., 79 (2001), pp. 2844–2846.

[3] I. Anderson, *Comment on "Remarkable transmission of microwaves through a wall of long metallic bricks,"* Appl. Phys. Lett., 82 (2003), pp. 308–309.

[4] G. A. Kriegsmann, *Scattering by large resonant cavity structures*, Wave Motion, 39 (1999), pp. 329–344.

[5] P. Sheng, R. S. Stepleman, and P. N. Sanda, *Exact eigenfunctions for square-wave gratings: Application to diffraction and surface-plasmon calculations*, Phys. Rev. B, 26 (1982), pp. 2907–2916.

[6] J. A. Porto, F. J. Garcia-Vidal, and J. B. Pendry, *Transmission resonances on metallic gratings with very narrow slits*, Phys. Rev. Lett., 83 (1999), pp. 2845–2848.

[7] U. Schroter and D. Heitmann, *Surface-plasmon-enhanced transmission through metallic gratings*, Phys. Rev. B, 58 (1998), pp. 15419–15421.

[8] A. N. Norris and H. A. Luo, *Acoustic radiation and reflection from a periodically perforated rigid solid*, J. Acoust. Soc. Amer., 82 (1987), pp. 2113–2122.

[9] D. S. Jones, *Acoustic and Electromagnetic Waves*, Oxford Science Publications, Clarendon Press, Oxford, UK, 1989.

SIAM J. APPL. MATH.
Vol. 65, No. 1, pp. 43–68

NUMERICAL HOMOGENIZATION AND CORRECTORS FOR NONLINEAR ELLIPTIC EQUATIONS*

Y. EFENDIEV† AND A. PANKOV‡

Abstract. In this paper we consider numerical homogenization and correctors for nonlinear elliptic equations. The numerical correctors are constructed for operators with homogeneous random coefficients. The construction employs two scales, one a physical scale and the other a numerical scale. A numerical homogenization technique is proposed and analyzed. This procedure is developed within finite element formulation. The convergence of the numerical procedure is presented for the case of general heterogeneities using G-convergence theory. The proposed numerical homogenization procedure for elliptic equations can be considered as a generalization of multiscale finite element methods to nonlinear equations. Using corrector results we construct an approximation of oscillatory solutions. Numerical examples are presented.

Key words. homogenization, multiscale, upscaling, scale-up, random, nonlinear, elliptic, finite element

AMS subject classification. 65N99

DOI. 10.1137/S0036139903424886

1. Introduction.

Consider the nonlinear elliptic equations

$$-\mathrm{div}(a_\epsilon(x, u_\epsilon, Du_\epsilon)) + a_{0,\epsilon}(x, u_\epsilon, Du_\epsilon) = f, \qquad u_\epsilon \in W_0^{1,p}(Q). \tag{1}$$

Here ϵ denotes the small scale of the problem. Direct numerical simulations of these kinds of problems are difficult because of scale disparity. Our objective is to find the approximation of the homogenized solution without solving the fine scale problem; i.e., (1) is solved on a grid of size h, where $h \gg \epsilon$. The numerical procedure introduced for this purpose can be regarded as numerical homogenization. The numerical homogenization procedure for (1) should account for the functional dependence of the macroscopic quantities on the solution and its gradients. Our motivation in considering (1) mostly stems from the applications of flow in porous media (multiphase flow in saturated porous media and flow in unsaturated porous media) and enhanced diffusion due to nonlinear heterogeneous convection, though many applications of nonlinear elliptic equations of these kinds occur in transport problems.

In this paper we consider two issues: (1) the calculation of the correctors and (2) the computation of the homogenized solution. The homogenization of nonlinear elliptic equations in a random media has been studied previously (see, e.g., [17]). It was shown that a solution u_ϵ converges (up to a subsequence) to u in an appropriate norm and where $u \in W_0^{1,p}(Q)$ is a solution of

$$-\mathrm{div}(a^*(x, u, Du)) + a_0^*(x, u, Du) = f. \tag{2}$$

The homogenized coefficients can be computed if we make an additional assumption on the heterogeneities such as periodicity, almost periodicity, or when the fluxes are

*Received by the editors March 23, 2003; accepted for publication (in revised form) January 29, 2004; published electronically September 24, 2004.
http://www.siam.org/journals/siap/65-1/42488.html

†Department of Mathematics, Texas A&M University, College Station, TX 77843-3368 (efendiev@math.tamu.edu). The work of this author is partially supported by NSF grants DMS 0327713 and EIA-0218229.

‡Department of Mathematics, College of William & Mary, Williamsburg, VA 23187-8795 (pankov@math.wm.edu).

strictly stationary fields with respect to spatial variables. In these cases one has an auxiliary problem of calculating a^* and a_0^*. The numerical homogenization procedure presented in this paper does not use the auxiliary problem for the calculation of the approximation of homogenized solutions.

To construct the numerical correctors we use two scales, a physical scale and a numerical scale that is much larger than the physical one, and construct the correctors in each numerical coarse block. The convergence for the corrector is obtained. These results show us a way to obtain numerically the fine scale features of the solution. We would like to note that the computation of the oscillation of solutions is important for the application to flow in porous media and other transport problems.

We present a procedure for calculating a coarse solution, the solution at the length scales h, $1 \gg h \gg \epsilon$. Our numerical homogenization procedure is based on general finite element computations of the coarse scale equations. It selectively solves the required local problems that reduce overall computations even in the periodic case. The solutions of the local problems are uniquely determined, which makes our discrete operator single-valued. The convergence of the numerical method is presented for general kinds of heterogeneities using G-convergence theory. Moreover, we show that the numerical homogenization approach presented in this paper can be considered as a generalization of multiscale finite element methods introduced in [10]. A related work in multiscale computations involves generalized finite element methods [2], residual free bubbles [3, 19], the variational multiscale method [12], two-scale finite element methods [15], two-scale conservative subgrid approaches [1], and the heterogeneous multiscale method (HMM) [6].

Some numerical examples are considered in this paper. We study numerically the effect of enhanced diffusion due to heterogeneous nonlinear convection,

$$\frac{\partial u_\epsilon}{\partial t} + \frac{1}{\epsilon} v_\epsilon(x) \cdot DF(u_\epsilon) - d\Delta u_\epsilon = f.$$

Since the elliptic part does not depend on t, the theory developed previously can be applied. In this application we are interested in the effect of the enhanced diffusion due to heterogeneous nonlinear convection. More precisely, assuming the existence of homogeneous stream function for the velocity field and zero mean drift, we calculate the approximation of the enhanced diffusion due to the convection using Buckley–Leverett flux that describes the convection. Other numerical examples for Richards equations are also studied.

The paper is organized as follows. In the next section we present some basic facts that are used later in the analysis. Section 3 is devoted to the construction of a numerical corrector and its convergence. Section 4 is devoted to the calculation of the homogenized solution and its analysis. In section 5 we present numerical results.

2. Preliminaries. We start with a description of random homogeneous fields on R^d, which are shown to be useful in homogenization problems (see, e.g., [13]). Let (Ω, Σ, μ) be a probability space. A random homogeneous field is a measurable function on Ω and $f(T(x)\omega)$ are realizations of the random field. The realizations are well-defined measurable functions on R^d for almost all $\omega \in \Omega$. Consider a d dimensional dynamical system on Ω, $T(x) : \Omega \to \Omega$, $x \in R^d$, that satisfies the following conditions: (1) $T(0) = I$, and $T(x + y) = T(x)T(y)$; (2) $T(x) : \Omega \to \Omega$ preserve the measure μ on Ω; (3) for any measurable function $f(\omega)$ on Ω, the function $f(T(x)\omega)$ defined on $R^d \times \Omega$ is also measurable (see [13, 18]). Let $L^p(\Omega)$ denote the space of all p-integrable

functions on Ω. Then $U(x)f(\omega) = f(T(x)\omega)$ defines a d-parameter group of isometries in the space $L^p(\Omega)$, and $U(x)$ is strongly continuous [13, 17]. Further, we assume that the dynamical system T is ergodic; i.e., any measurable T-invariant function on Ω is constant. Denote by $\langle\cdot\rangle$ the mean value over Ω,

$$\langle f\rangle = \int_\Omega f(\omega)d\mu(\omega) = E(f).$$

Now we explain briefly the relation between the standard definition of random homogeneous fields and the one we introduced here following, e.g., [17]. Let Ξ be a probability space endowed with a probability measure P. Let f be a random vector valued function, i.e., a measurable map $f : \Xi \times R^d \to R^N$. f is a random homogeneous field if all its finite dimensional distributions are translation invariant. The latter means that for any x^1, x^2, ..., $x^k \in R^d$, and any Borel subsets B_1, B_2, ..., $B_k \subset R^N$,

$$P\{\xi \in \Xi : f(\xi, x^1 + h) \in B_1, \dots, f(\xi, x^k + h) \in B_k\}$$

is independent of $h \in R^d$. Consider a new probability space Ω and a dynamical system $T(x)$ acting on Ω. We define Ω to be the set of all measurable functions $\omega : R^d \to R^N$ and set $T(x)\omega(y) = \omega(x + y)$, $x, y \in R^d$. Let $\mathcal{F}$ be the σ-algebra generated by "cylinder" sets, i.e., the sets of the form $B = \{\omega : \omega(x^1) \in B_1, \dots, \omega(x^k) \in B_k\}$, where x^1, x^2, ..., $x^k \in R^d$ and B_1, B_2, ..., B_k are Borel subsets in R^N. We define the measure μ on "cylinder" sets by $\mu(B) = P\{\xi \in \Xi : f(\xi, \cdot) \in B\}$ and then extend it to $\mathcal{F}$ by σ-additivity. Thus, the probability space Ω and the measure-preserving dynamical system $T(x)$, $x \in R^d$, on Ω are constructed. Moreover, consider the μ-measurable function $\hat{f} : \Omega \to R^N$ defined by the formula $\hat{f}(\omega) = \omega(0)$. Then $f(\xi, x) = \hat{f}(T(x)\omega)$, where $\omega(\cdot) = f(\xi, \cdot)$. More examples regarding the construction of T can be found in [13].

Denote by ∂^i_ω the generator of $U(x)$ along the ith coordinate direction, i.e.,

$$\partial^i_\omega = \lim_{\delta\to 0} \frac{f(T(\delta e_i)\omega) - f(\omega)}{\delta}.$$

The domains D^i of ∂^i_ω are dense in $L^2(\Omega)$, and the intersection of all D^i is also dense.

Next, following [17], we define potential and solenoidal fields. A vector field $f \in L^p(\Omega)$ is said to be potential (or solenoidal, respectively) if its generic realization $f(T(x)\omega)$ is a *potential* (or *solenoidal*, respectively) vector field in R^d. Denote by $L^p_{pot}(\Omega)$ (respectively, $L^p_{sol}(\Omega)$) the subspace of $L^p(\Omega)$ that consists of all potential (respectively, solenoidal) vector fields. Introduce the following notation:

$$V^p_{pot} = \{f \in L^p_{pot}(\Omega), \langle f\rangle = 0\}, \qquad V^p_{sol} = \{f \in L^p_{sol}(\Omega), \langle f\rangle = 0\}.$$

The following properties are known (see [17, page 138]):

$$L^p_{pot}(\Omega) = V^p_{pot} \oplus R^d, \quad L^p_{sol}(\Omega) = V^p_{sol} \oplus R^d, \quad L^q_{sol}(\Omega) = (V^p_{pot})^\perp, \quad L^q_{pot}(\Omega) = (V^p_{sol})^\perp.$$

Consider $u_\epsilon \in W^{1,p}_0(Q)$,

$$-\mathrm{div}(a(T(x/\epsilon)\omega, u_\epsilon, Du_\epsilon)) + a_0(T(x/\epsilon)\omega, u_\epsilon, Du_\epsilon) = f \ \text{ in } Q, \tag{3}$$

where f is a deterministic function that does not depend on ϵ and is sufficiently smooth, and $Q \subset R^d$ is a domain with Lipschitz boundaries.

Assume for all $\omega \in \Omega$

$$(a(\omega,\eta,\xi_1) - a(\omega,\eta,\xi_2),\xi_1-\xi_2) \geq C(1+|\xi_1|+|\xi_2|)^{p-\beta}|\xi_1-\xi_2|^{\beta}, \tag{4}$$

$$|a(\omega,\eta,\xi)| + |a_0(\omega,\eta,\xi)| \leq C(1+|\eta|+|\xi|)^{p-1}, \tag{5}$$

$$\begin{aligned} &|a(\omega,\eta_1,\xi_1) - a(\omega,\eta_2,\xi_2)| + |a_0(\omega,\eta_1,\xi_1) - a_0(\omega,\eta_2,\xi_2)| \\ &\quad \leq C(1+|\eta_1|^{p-1}+|\eta_2|^{p-1}+|\xi_1|^{p-1}+|\xi_2|^{p-1})\nu(|\eta_1-\eta_2|) \\ &\qquad + C(1+|\eta_1|^{p-1-s}+|\eta_2|^{p-1-s}+|\xi_1|^{p-1-s}+|\xi_2|^{p-1-s})|\xi_1-\xi_2|^s, \end{aligned} \tag{6}$$

where $0 < s \leq 1$, $\beta \geq \max(p,2)$, $p > 1$. Here $\nu(r)$ is a continuity modulus; i.e., $\nu(r)$ is a nondecreasing continuous function on $[0,+\infty)$ such that $\nu(0) = 0$, $\nu(r) > 0$ if $r > 0$ and $\nu(r) = 1$ if $r > 1$, and $\nu(r_1 + r_2) \leq C(\nu(r_1) + \nu(r_2))$. For the existence of the solution we need a coercivity condition,

$$(a(\omega,\eta,\xi),\xi) + a_0(\omega,\eta,\xi)\eta \geq C|\xi|^p - C_1. \tag{7}$$

It is known (e.g., [17]) that, as $\epsilon \to 0$, Du_ϵ converges to Du weakly in $L^p(Q)^d$ for almost every ω, and u is the solution of

$$-\mathrm{div}(a^*(u,Du)) + a_0^*(u,Du) = f, \qquad u \in W_0^{1,p}(Q). \tag{8}$$

Further, a^* and a_0^* can be constructed using the solution of the following auxiliary problem. Given $\eta \in R$ and $\xi \in R^d$, define $w_{\eta,\xi} \in V_{pot}^p$ such that

$$a(\omega,\eta,\xi + w_{\eta,\xi}(\omega)) \in L_{sol}^q(\Omega)^d. \tag{9}$$

Then $a^*(\eta,\xi)$ and $a_0(\eta,\xi)$ are defined as

$$\begin{aligned} a^*(\eta,\xi) &= \langle a(\omega,\eta,\xi + w_{\eta,\xi}(\omega))\rangle, \\ a_0^*(\eta,\xi) &= \langle a_0(\omega,\eta,\xi + w_{\eta,\xi}(\omega))\rangle. \end{aligned} \tag{10}$$

Moreover, $a^*(\eta,\xi)$ and $a_0^*(\eta,\xi)$ satisfy estimates similar to those of a and a_0 with different constants [17].

Remark 2.1. We would like to note that G-convergence and homogenization results presented in [17] were formulated under weaker than (4) conditions. In particular, it is assumed that

$$(a(\omega,\eta,\xi_1) - a(\omega,\eta,\xi_2),\xi_1-\xi_2) \geq C(1+|\eta|+|\xi_1|+|\xi_2|)^{p-\beta}|\xi_1-\xi_2|^{\beta}. \tag{11}$$

It turns out that G-convergence and homogenization results hold under more general assumptions such as (4). The proof is identical to the one presented in [17]. Moreover, following [17], it can be easily shown that the homogenized operator is also coercive and satisfies (7).

Throughout the paper C denotes a generic constant, $\|\cdot\|_p$ denotes $L^p(Q)$ (or the broken norm), and $L^p(Q)^d$ norms and q are defined by $1/p + 1/q = 1$. The notation a.e. (almost every) is often omitted.

3. Two-scale correctors. The corrector results obtained in this section will be used in the approximation of solution gradients. The importance of this approximation is motivated by some applications in which details of the fluxes play a key role in a physical phenomenon (e.g., flow in porous media). For the construction we assume that the homogenized solution is computed with a reliable accuracy in an appropriate norm which will be specified later. In the next section we will propose a numerical procedure for the computation of the homogenized solution for more general heterogeneities. For the construction of the correctors we introduce two scale correctors, where one scale represents the numerical scale h and the other the physical scale ϵ.

Define $M_h\phi(x)$ in the following way:

$$M_h\phi(x) = \sum_i 1_{Q_i} \frac{1}{|Q_i|} \int_{Q_i} \phi(y)dy,$$

where $\bigcup Q_i = Q$. Here Q_i are domains with diameter of order h, e.g., finite element triangles or some unions of the triangles. Note that $M_h\phi \to \phi$ in $L^p(Q)$ as $h \to 0$ (see [22]). Further, define

$$P(T(y)\omega, \eta, \xi) = \xi + w_{\eta,\xi}(T(y)\omega), \tag{12}$$

where $w_{\eta,\xi} \in V^p_{pot}(\Omega)$ is the solution of the auxiliary problem $a(\omega, \eta, \xi + w_{\eta,\xi}(\omega)) \in L^q_{sol}(\Omega)^d$. Here $w_{\eta,\xi}(T(y)\omega)$ satisfies

$$-\mathrm{div}(a(T(y)\omega, \eta, \xi + w_{\eta,\xi}(T(y)\omega))) = 0$$

in the sense of distribution [17, p. 155].

The main result of this section regarding the convergence of the correctors is the following.

THEOREM 3.1. *Let u_ϵ and u be solutions of* (3) *and* (8)*, respectively, and let P be defined by* (12) *in each Q_i. Then*

$$\lim_{h\to 0} \lim_{\epsilon\to 0} \int_Q |P(T(x/\epsilon)\omega, M_h u, M_h Du) - Du_\epsilon|^p dx = 0 \tag{13}$$

μ-a.e.

We will omit μ-a.e. notation in further analysis. To make the expressions in the proof more concise we introduce the notation

$$\mathcal{P}_\epsilon = P(T(x/\epsilon)\omega, M_h u, M_h Du).$$

Theorem 3.1 indicates that the gradient of solutions can be approximated by $P(T(x/\epsilon)\omega, M_h u, M_h Du)$. This quantity can be computed based on $M_h Du$ and $M_h u$, i.e., the gradient of the coarse solution in each coarse block, as we will show later. The following lemma [4] will be used in the proof.

LEMMA 3.2. *For any ϕ_1 and ϕ_2 belonging to $L_p(Q)$ we have*

$$\begin{aligned}\|\phi_1 - \phi_2\|_{p,Q} \le C \left(\int_Q |\phi_1 - \phi_2|^\beta (1 + |\phi_1| + |\phi_2|)^{p-\beta} dx \right)^{1/\beta} \\ \times (|Q|^{1/p} + \|\phi_1\|_{p,Q} + \|\phi_2\|_{p,Q})^{(\beta-p)/\beta}.\end{aligned} \tag{14}$$

For the proof of Theorem 3.1 we need the following lemma.

LEMMA 3.3. *For every $\eta \in R$ and $\xi \in R^d$*

$$\|P(\omega,\eta,\xi)\|_{p,\Omega}^p \le C(1+|\eta|^p+|\xi|^p).$$

Proof. Using Lemma 3.2 and (4), we obtain

$$\begin{aligned}\|P\|_{p,\Omega}^p &\le C(1+\|P\|_{p,\Omega}^p)^{(\beta-p)/\beta}\left(\int_\Omega |P|^\beta(1+|P|)^{p-\beta}d\mu(\omega)\right)^{p/\beta}\\ &\le C\delta^{(p-\beta)/p}\int_\Omega |P|^\beta(1+|P|)^{p-\beta}d\mu(\omega)+C\delta(1+\|P\|_{p,\Omega}^p).\end{aligned}$$

With a suitable choice of δ and using (4) and (5), we get

$$\begin{aligned}\|P\|_{p,\Omega}^p &\le C+C\int_\Omega |P|^\beta(1+|P|)^{p-\beta}d\mu(\omega)\le C+C\int_\Omega (a(\omega,\eta,P)-a(\omega,\eta,0),P)d\mu(\omega)\\ &\le C+C\left|\int_\Omega (a(\omega,\eta,P),P)d\mu(\omega)\right|+\left|\int_\Omega (a(\omega,\eta,0),P)d\mu(\omega)\right|\\ &\le C+\left|\int_\Omega (a(\omega,\eta,P),\xi)d\mu(\omega)\right|+(1+|\eta|^{p-1})\left|\int_\Omega P d\mu(\omega)\right|\\ &\le C+C\delta_1\|P\|_{p,\Omega}^p+C\delta_1^{-1/(p-1)}|\eta|^p+C\int_\Omega (1+|\eta|+|P|)^{p-1}|\xi|d\mu(\omega)\\ &\le C\delta_2\|P\|_{p,\Omega}^p+C\delta_2^{-1/(p-1)}(1+|\xi|^p)+C+C\delta_1(|\eta|^p+\|P\|_{p,\Omega}^p)+C\delta_1^{-1/(p-1)}|\eta|^p.\end{aligned}$$

With an appropriate choice of δ_1 and δ_2, we obtain the desired result. ∎

It follows from Lemma 3.3 that $P(T(y)\omega,\eta,\xi)\in L^p_{loc}(R^d)^d$ for a.e. ω and for each $\eta\in R$, $\xi\in R^d$. The next lemma will be also used in the proof of Theorem 3.1.

LEMMA 3.4. *If $u_k\to 0$ in $L^r(Q)$ $(1<r<\infty)$ as $k\to\infty$, then*

$$\int_Q \nu(u_k)|v_k|^p dx\to 0 \text{ as } k\to\infty$$

for all v_k either (1) *compact in $L^p(Q)$ or* (2) *bounded in $L^{p+\alpha}(Q)$, $\alpha>0$. Here $\nu(r)$ is the continuity modulus defined previously* (*see* (6)) *and $1<p<\infty$.*

Proof. Since u_k converges in L^r, it converges in measure. Consequently, for any $\delta>0$ there exists Q_δ and k_0 such that $\mathrm{meas}(Q\setminus Q_\delta)<\delta$ and $\nu(u_k)<\delta$ in Q_δ for $k>k_0$. Thus

$$\begin{aligned}\int_Q \nu(u_k)|v_k|^p dx &= \int_{Q_\delta}\nu(u_k)|v_k|^p dx+\int_{Q\setminus Q_\delta}\nu(u_k)|v_k|^p dx\\ &\le C\delta+C\int_{Q\setminus Q_\delta}|v_k|^p dx.\end{aligned}\tag{15}$$

Next we use the fact that if (1) or (2) is satisfied, then the set v_k has equi-absolute continuous norm [14] (i.e., for any $\epsilon>0$ there exists $\zeta>0$ such that $\mathrm{meas}(Q_\zeta)<\zeta$ implies $\|P_{Q_\zeta}v_k\|_{p,Q}<\epsilon$, where $P_Df=\{f(x) \text{ if } x\in D; 0 \text{ otherwise}\}$. Consequently, the second term on the right-hand side (r.h.s.) of (15) converges to zero as $\delta\to 0$. ∎

The proof of Theorem 3.1 will be based on the following estimate:

$$\begin{aligned}\int_Q |\mathcal{P}_\epsilon-Du_\epsilon|^p dx &\le \int_Q |\mathcal{P}_\epsilon-Du_\epsilon|^\beta(1+|\mathcal{P}_\epsilon|+|Du_\epsilon|)^{p-\beta}dx(|Q|^{1/p}\\ &\quad+\|\mathcal{P}_\epsilon\|_{p,Q}+\|Du_\epsilon\|_{p,Q})^{(\beta-p)/p}.\end{aligned}\tag{16}$$

$\|Du_\epsilon\|_{p,Q}$ is uniformly bounded for a.e. ω. $\|\mathcal{P}_\epsilon\|_{p,Q}$ is also uniformly bounded since $M_h u$ and $M_h Du$ are bounded in $L^p(Q)$ and $L^p(Q)^d$, respectively. Thus it remains to estimate $\int_Q |\mathcal{P}_\epsilon - Du_\epsilon|^\beta (1 + |\mathcal{P}_\epsilon| + |Du_\epsilon|)^{p-\beta} dx$. For this term, using (4), we have

$$
\begin{aligned}
&\int_Q |\mathcal{P}_\epsilon - Du_\epsilon|^\beta (1 + |\mathcal{P}_\epsilon| + |Du_\epsilon|)^{p-\beta} dx \\
&\qquad \le C \left| \int_Q (a(T(x/\epsilon)\omega, u_\epsilon, \mathcal{P}_\epsilon) - a(T(x/\epsilon)\omega, u_\epsilon, Du_\epsilon), \mathcal{P}_\epsilon - Du_\epsilon) dx \right| \\
&\qquad \le C \left| \int_Q (a(T(x/\epsilon)\omega, M_h u, \mathcal{P}_\epsilon) - a(T(x/\epsilon)\omega, u_\epsilon, Du_\epsilon), \mathcal{P}_\epsilon - Du_\epsilon) dx \right| \\
&\qquad\quad + C \left| \int_Q (a(T(x/\epsilon)\omega, u_\epsilon, \mathcal{P}_\epsilon) - a(T(x/\epsilon)\omega, M_h u, \mathcal{P}_\epsilon), \mathcal{P}_\epsilon - Du_\epsilon) dx \right|.
\end{aligned} \tag{17}
$$

To prove Theorem 3.1 we will need to estimate the first and second terms on the r.h.s. of (17). For the first term we have

$$
\begin{aligned}
&C \int_Q \big(a(T(x/\epsilon)\omega, M_h u, \mathcal{P}_\epsilon) - a(T(x/\epsilon)\omega, u_\epsilon, Du_\epsilon), \mathcal{P}_\epsilon - Du_\epsilon\big) dx \\
&\qquad = C \int_Q (a(T(x/\epsilon)\omega, M_h u, \mathcal{P}_\epsilon), \mathcal{P}_\epsilon) dx - C \int_Q (a(T(x/\epsilon)\omega, M_h u, \mathcal{P}_\epsilon), Du_\epsilon) dx \\
&\qquad\quad - C \int_Q (a(T(x/\epsilon)\omega, u_\epsilon, Du_\epsilon), \mathcal{P}_\epsilon) dx + C \int_Q (a(T(x/\epsilon)\omega, u_\epsilon, Du_\epsilon), Du_\epsilon) dx.
\end{aligned} \tag{18}
$$

We will investigate the r.h.s. of (18) in the limit as $\epsilon \to 0$. For the first term of the r.h.s. of (18) we have the following convergence.

LEMMA 3.5.

$$
\int_Q (a(T(x/\epsilon)\omega, M_h u, \mathcal{P}_\epsilon), \mathcal{P}_\epsilon) dx \to \int_Q (a^*(M_h u, M_h Du), M_h Du) dx
$$

as $\epsilon \to 0$.

Proof.

$$
\begin{aligned}
&\int_Q (a(T(x/\epsilon)\omega, M_h u, \mathcal{P}_\epsilon), \mathcal{P}_\epsilon) dx \\
&\qquad = \sum_i \int_{Q_i} \big(a(T(x/\epsilon)\omega, \eta_i, \xi_i + w_{\eta_i,\xi_i}(T(x/\epsilon)\omega)), \xi_i + w_{\eta_i,\xi_i}(T(x/\epsilon)\omega)\big) dx \\
&\qquad = \sum_i \int_{Q_i} \big(a(T(x/\epsilon)\omega, \eta_i, \xi_i + w_{\eta_i,\xi_i}(T(x/\epsilon)\omega)), \xi_i\big) dx \\
&\qquad\quad + \sum_i \int_{Q_i} \big(a(T(x/\epsilon)\omega, \eta_i, \xi_i + w_{\eta_i,\xi_i}(T(x/\epsilon)\omega)), w_{\eta_i,\xi_i}(T(x/\epsilon)\omega)\big) dx \\
&\qquad \to \sum_i \int_{Q_i} 1_{Q_i} (a^*(\eta_i, \xi_i), \xi_i) dx
\end{aligned}
$$

as $\epsilon \to 0$. In the last step we have used the Birkhoff ergodic theorem (see [13]) as well as (9), (10), and $w_{\eta,\xi} \in V^p_{pot}$. Next we note that the limit can be written as

$$
\sum_i \int_{Q_i} 1_{Q_i} (a^*(\xi_i), \xi_i) dx = \int_Q (a^*(M_h u, M_h Du), M_h Du) dx. \qquad \square
$$

For the second term of the r.h.s. of (18) we have the following convergence.

LEMMA 3.6.

$$\int_Q (a(T(x/\epsilon)\omega, M_h u, \mathcal{P}_\epsilon), Du_\epsilon)dx \to \int_Q (a^*(M_h u, M_h Du), Du)dx$$

as $\epsilon \to 0$.

Proof.

$$\int_Q (a(T(x/\epsilon)\omega, M_h u, \mathcal{P}_\epsilon), Du_\epsilon)dx = \sum_i \int_{Q_i} (a(T(x/\epsilon)\omega, \eta_i P(T(x/\epsilon)\omega, \eta_i, \xi_i)), Du_\epsilon)dx.$$

Du_ϵ is bounded in $L^p(Q)^d$ for a.e. ω. To show that $a(T(x/\epsilon)\omega, P(T(x/\epsilon)\omega, \eta_i, \xi_i))$ is bounded in $L^r(Q_i)^d$, where $r > q$, we will use Meyers' theorem [16]. Since $-\mathrm{div}(a(T(x/\epsilon)\omega, \eta_i, P(T(x/\epsilon)\omega, \eta_i, \xi_i))) = 0$ in $3 \times Q_i$ (where $3 \times Q_i$ is a domain that contains Q_i and is surrounded with a ring of size Q_i), using Meyers' theorem we can conclude that

$$\|P(T(x/\epsilon)\omega, \eta_i, \xi_i)\|_{p+\eta, Q_i} \le C\|P(T(x/\epsilon)\omega, \eta_i, \xi_i)\|_{p, 3\times Q_i},$$

where C is independent of ω and depends only on operator constants. Note that $P \in L^p_{loc}(R^d)^d$. Since $\|P(T(x/\epsilon)\omega, \eta_i, \xi_i)\|_{p,3\times Q_i}$ is bounded for a.e. ω (see Lemma 3.3), $\|P(T(x/\epsilon)\omega, \eta_i, \xi_i)\|_{p+\alpha, Q_i}$ is also bounded for a.e. ω. From here, using bounds for $a(T(y)\omega, \eta, \xi)$, we can easily obtain that $a(T(x/\epsilon)\omega, \eta_i, P(T(x/\epsilon)\omega, \eta_i, \xi_i))$ is bounded in $L^r(Q_i)^d$, where $r > q$ for a.e. ω. Indeed,

$$\begin{aligned}
&\int_{Q_i} |a(T(x/\epsilon)\omega, \eta_i, P(T(x/\epsilon)\omega, \eta_i, \xi_i)) - a(T(x/\epsilon)\omega, \eta_i, 0)|^r dx \\
&\quad \le C \int_{Q_i} (|1 + \eta_i + P(T(x/\epsilon)\omega, \eta_i, \xi_i)|)^{(p-2)r} |P(T(x/\epsilon)\omega, \eta_i, \xi_i)|^r dx \\
&\quad \le C(\|P\|_{r,Q_i} + \|P\|_{(p-1)r,Q_i}).
\end{aligned}$$

Since P is in $L^{p+\alpha}(Q_i)^d$ for a.e. ω, we can pick $r = q + \alpha/(p-1)$. Consequently, $(a(T(x/\epsilon)\omega, \eta_i, \xi_i + w_{\eta_i,\xi_i}(T(x/\epsilon)\omega)), Du_\epsilon)$ is bounded in $L^\sigma(Q_i)^d$, $\sigma > 1$, for every η_i and ξ_i. Thus it contains a subsequence that weak* converges to g_i for any i and a.e. ω. Since $-\mathrm{div}(a(T(x/\epsilon)\omega, \eta_i, P(T(x/\epsilon)\omega, \eta_i, \xi_i))) = 0$ in Q_i, using a compensated compactness argument we can obtain that as $\epsilon \to 0$, $g_i = (a^*(\eta_i, \xi_i), Du)$. The latter is true because Du_ϵ weakly converges to Du in $L^p(Q)^d$ for a.e. ω and $a(T(x/\epsilon)\omega, \eta_i, P(T(x/\epsilon)\omega, \eta_i, \xi_i))$ weakly converges to $a^*(\eta_i, \xi_i)$ in $L^r(Q)$. The fact that Du_ϵ weakly converges to Du for a.e. ω follows from general G-convergence results [17], and the weak convergence of $a(T(x/\epsilon)\omega, \eta_i, P(T(x/\epsilon)\omega, \eta_i, \xi_i))$ is a consequence of the Birkhoff ergodic theorem. Consequently,

$$\begin{aligned}
&\sum_i \int_{Q_i} (a(T(x/\epsilon)\omega, \eta_i, P(T(x/\epsilon)\omega, \eta_i, \xi_i)), Du_\epsilon)dx \\
&\quad \to \sum_i \int_{Q_i} (a^*(\eta_i, \xi_i), Du)dx = \int_Q (a^*(M_h u, M_h Du), Du)dx. \qquad \square
\end{aligned}$$

For the third term of the r.h.s. of (18) we have the following convergence.

LEMMA 3.7.

$$\int_Q (a(T(x/\epsilon)\omega, u_\epsilon, Du_\epsilon), \mathcal{P}_\epsilon)dx \to \int_Q (a^*(u, Du), M_h Du)dx$$

as $\epsilon \to 0$.

Proof.

$$\int_Q (a(T(x/\epsilon)\omega, u_\epsilon, Du_\epsilon), \mathcal{P}_\epsilon)dx = \sum_i \int_{Q_i} (a(T(x/\epsilon)\omega, u_\epsilon, Du_\epsilon), P(T(x/\epsilon)\omega, \eta_i, \xi_i))dx.$$

Since $|a(\omega, \eta, \xi)| \le C(1 + |\eta|^{p-1} + |\xi|^{p-1})$ and $P(T(x/\epsilon)\omega, \eta_i, \xi_i)$ converges to ξ_i in $L^p(Q)^d$ and is bounded in $L^{p+\eta}(Q)^d$, similar to the analysis for the Lemma 3.6 we can obtain that

$$\begin{aligned}&\sum_i \int_{Q_i} (a(T(x/\epsilon)\omega, u_\epsilon, Du_\epsilon), P(T(x/\epsilon)\omega, \eta_i, \xi_i))dx \\ &\quad \to \sum_i \int_{Q_i} (a^*(u, Du), \xi_i)dx = \int_Q (a^*(u, Du), M_h Du)dx. \qquad \square\end{aligned}$$

For the fourth term of the r.h.s. of (18), we have the following.

LEMMA 3.8.

$$\int_Q (a(T(x/\epsilon)\omega, u_\epsilon, Du_\epsilon), Du_\epsilon)dx \to \int_Q (a^*(u, Du), Du)dx$$

as $\epsilon \to 0$.

Proof.

$$\begin{aligned}&\int_Q (a(T(x/\epsilon)\omega, u_\epsilon, Du_\epsilon), Du_\epsilon)dx \\ &\quad = -\int_Q (\mathrm{div}(a(T(x/\epsilon)\omega, u_\epsilon, Du_\epsilon)), u_\epsilon)dx = -\int_Q f u_\epsilon dx \\ &\quad \to -\int_Q f u dx = \int_Q (a^*(u, Du), Du)dx. \qquad \square\end{aligned}$$

Next for the second term on the r.h.s. of (17), using (6), we have

(19)

$$\begin{aligned}&C\left|\int_Q (a(T(x/\epsilon)\omega, u_\epsilon, \mathcal{P}_\epsilon) - a(T(x/\epsilon)\omega, M_h u, \mathcal{P}_\epsilon), \mathcal{P}_\epsilon - Du_\epsilon)dx\right| \\ &\quad \le \frac{C}{\delta_1}\left|\int_Q a(T(x/\epsilon)\omega, u_\epsilon, \mathcal{P}_\epsilon) - a(T(x/\epsilon)\omega, M_h u, \mathcal{P}_\epsilon)\right|^q dx + C\delta_1 \int_Q |\mathcal{P}_\epsilon - Du_\epsilon|^p dx \\ &\quad \le \frac{C}{\delta_1}\sum_i \int_{Q_i} \nu(|u_\epsilon - \eta_i|)^q (1 + |\xi_i|^p)dx \\ &\qquad + \frac{C}{\delta_1}\sum_i \int_{Q_i} \nu(|u_\epsilon - \eta_i|)^q (1 + |w_{\eta_i,\xi_i}|^p)dx + C\delta_1 \int_Q |\mathcal{P}_\epsilon - Du_\epsilon|^p dx,\end{aligned}$$

where $\nu(r)$ is a continuity modulus defined earlier (see (6)). Here we have used the uniform boundedness of Du_ϵ as well as u_ϵ in $L^p(Q)^d$ and $L^p(Q)$, respectively. The first term on the r.h.s. converges to $\int_Q \nu(|u - M_h u|)^q (1 + |M_h Du|^p)dx$ by Lemma 3.4. For the second term, using Meyers' theorem (cf. Lemma 3.6), we obtain that w_{η_i,ξ_i} is bounded in $L^{p+\alpha}(Q_i)^d$, $\alpha > 0$. Thus, using Lemma 3.4, we have that the second term for each i converges to $\int_{Q_i} \nu(|u - \eta_i|)^q (1 + \langle |w_{\eta_i,\xi_i}|^p \rangle)dx$, which is not

greater than $\int_{Q_i} \nu(|u-\eta_i|)^q(1+|\eta_i|^p+|\xi_i|^p)dx$. Summing this over all i, we get $\int_Q \nu(|u-M_hu|)^q(1+|M_hu|^p+|M_hDu|^p)dx$. Thus (19) is not greater than

$$\int_Q \nu(|u-M_hu|)^q(1+|M_hu|^p+|M_hDu|^p)dx + C\delta_1 \int_Q |\mathcal{P}_\epsilon - Du_\epsilon|^p dx.$$

With an appropriate choice of δ_1, combining all the estimates, we have for (17) (cf. (16))

$$\begin{aligned}
&\lim_{\epsilon\to 0}\int_Q |\mathcal{P}_\epsilon - Du_\epsilon|^p dx \\
&\quad \le C\left(\int_Q (a^*(M_hu, M_hDu), M_hDu)dx - \int_Q (a^*(M_hu, M_hDu), Du)dx\right. \\
&\qquad \left. - \int_Q (a^*(u, Du), M_hDu)dx + \int_Q (a^*(u, Du), Du)dx\right) \\
&\qquad + \int_Q \nu(|u-M_hu|)^q(1+|M_hu|^p+|M_hDu|^p)dx.
\end{aligned} \tag{20}$$

Next it is not difficult to show that the r.h.s. of (20) approaches zero as $h \to 0$. For this reason we write

$$\begin{aligned}
&\int_Q (a^*(M_hu, M_hDu), M_hDu)dx - \int_Q (a^*(M_hu, M_hDu), Du)dx \\
&\qquad - \int_Q (a^*(u, Du), M_hDu)dx + \int_Q (a^*(u, Du), Du)dx \\
&\qquad = \int_Q (a^*(u, Du) - a^*(M_hu, M_hDu), Du - M_hDu)dx.
\end{aligned} \tag{21}$$

Next, using the estimate $|a^*(\eta_1,\xi_1) - a^*(\eta_2,\xi_2)| \le C(1+|\eta_1|^{p-1}+|\eta_2|^{p-1}+|\xi_1|^{p-1}+|\xi_2|^{p-1})\nu(|\eta_1-\eta_2|) + C(1+|\eta_1|^{p-1-\tilde{s}}+|\eta_2|^{p-1-\tilde{s}}+|\xi_1|^{p-1-\tilde{s}}+|\xi_2|^{p-1-\tilde{s}})|\xi_1-\xi_2|^{\tilde{s}}$, $0<\tilde{r}\le 1$ (see [17]), we can obtain that the r.h.s. of (21) converges to zero as $h\to 0$. Indeed,

$$\begin{aligned}
&\int_Q (a^*(u, Du) - a^*(M_hu, M_hDu), Du - M_hDu)dx \\
&\le C\int_Q (1+|u|^{p-1}+|Du|^{p-1}+|M_hu|^{p-1}+|M_hDu|^{p-1})\nu(|u-M_hu|)|Du-M_hDu|dx \\
&+ C\int_Q (1+|u|^{p-1-\tilde{s}}+|Du|^{p-1-\tilde{s}}+|M_hu|^{p-1-\tilde{s}}+|M_hDu|^{p-1-\tilde{s}})|Du-M_hDu|^{\tilde{s}}dx.
\end{aligned} \tag{22}$$

Using the Holder inequality, it can be easily shown that the second term here converges to zero as $h\to 0$. Since M_hu converges to u in $L^p(Q)$ and M_hDu converges to Du in $L^p(Q)^d$ from Lemma 3.4, the first term in (22) also converges to zero. Similarly one can show that the last term on the r.h.s. of (20) converges to zero as $h\to 0$. This completes the proof of Theorem 3.1.

As an example we consider the correctors for

$$\operatorname{div}(a(T(x/\epsilon)\omega)k_r(u_\epsilon)Du_\epsilon) = f. \tag{23}$$

We assume that the operator satisfies the conditions stated previously. In this case $P(T(y)\omega,\eta,\xi)=\xi+w_{\eta,\xi}(T(y)\omega)$, where $w_{\eta,\xi}\in L^p_{pot}(\Omega)$ satisfies

$$-\mathrm{div}(a(T(x/\epsilon)\omega)k_r(\eta)(\xi+w_{\eta,\xi}))=0.$$

Introducing a notation N such that $w^i_{\eta,\xi}(\omega)=N_{ij}(\omega)\xi_j$, we have the classical equation (see [13]) for $N(\omega)$, i.e., $a(\omega)(I+N)\in L^q_{sol}(\Omega)$. Consequently, the correctors for (23) have the form

$$P(T(y)\omega,\eta,\xi)=\xi(I+N(T(y)\omega)).$$

From this we conclude that u satisfies

$$\mathrm{div}(a^*k_r(u)Du)=f,$$

where a^* is the homogenized tensor corresponding to a linear elliptic operator. The approximation for the gradient of the solution is defined by

$$P(T(x/\epsilon)\omega,M_hu,M_hDu)=M_hDu(I+N(T(x/\epsilon)\omega)).$$

Theorem 3.1 shows a way to compute an approximation for the gradient of u_ϵ, although this computation is difficult since it involves the solution of the auxiliary problem. In the next section we will present the numerical computation of the oscillatory solution.

4. Numerical computation of the homogenized solution.

4.1. Numerical homogenization method. Consider $u_\epsilon\in W^{1,p}_0(Q)$,

$$-\mathrm{div}(a_\epsilon(x,u_\epsilon,Du_\epsilon))+a_{0,\epsilon}(x,u_\epsilon,Du_\epsilon)=f, \tag{24}$$

where $a_\epsilon(x,\eta,\xi)$ and $a_{0,\epsilon}(x,\eta,\xi)$, $\eta\in R$, $\xi\in R^d$, satisfy (4)–(6) and (7). As we mentioned in the introduction, the numerical homogenization procedure and its analysis can be studied for more general heterogeneities using G-convergence theory. The main idea of the numerical homogenization procedure is to find the homogenized solution without using the auxiliary problem. Consider a finite dimensional space over the standard triangular partitions K of $Q=\bigcup K$, and let

(25)
$$S^h=\{v_h\in C^0(\overline{Q})\colon \text{the restriction } v_h \text{ is linear for each element } K \text{ and } v_h=0 \text{ on } \partial Q\},$$

$\mathrm{diam}(K)\le Ch$. Here we assume that $h\gg\epsilon$ is chosen for the approximation of the homogenized solution. The numerical homogenization procedure consists of finding an approximation, $u_h\in S^h$, of a homogenized solution u such that

$$(A_{\epsilon,h}u_h,v_h)=\int_Q fv_h\,dx, \tag{26}$$

where

$$(A_{\epsilon,h}u_h,v_h)=\sum_K\int_K((a_\epsilon(x,\eta^{u_h},Du_{\epsilon,h}),Dv_h)+a_{0,\epsilon}(x,\eta^{u_h},Du_{\epsilon,h})v_h)dx. \tag{27}$$

Here $u_{\epsilon,h}$ satisfies

$$-\mathrm{div}(a_\epsilon(x,\eta^{u_h},Du_{\epsilon,h}))=0 \text{ in } K, \tag{28}$$

$u_{\epsilon,h}=u_h$ on ∂K, and $\eta^{v_h}(=M_hv_h)=\frac{1}{|K|}\int_K v_h dx$ in each K. Our numerical homogenization procedure consists of (26), (27), and (28). In some sense, (27) attempts to approximate $\int_Q[(a^*(x,u_h,Du_h),Dv_h)+a_0^*(x,u_h,Du_h)v_h]dx$, which is a finite element formulation of the homogenized equation. Note that solutions, u_h, of (26) depend on ϵ, which we do not explicitly write because $u_h\in S^h$. We would like to point out that different boundary conditions can be chosen; e.g., one can use an oversampling technique [10], where the solution of the larger problem is used in the calculation of the solution of local problems. We have implemented and shown the advantages of an oversampling technique in our recent work [7]. In the next subsection we will show that the numerical homogenization approach can be considered as a generalization of MsFEM.

Next we briefly describe the numerical implementation of MsFEM for nonlinear elliptic problems. For each $u_h=\sum_i\theta_i\phi_0^i(x)\in S^h$, where $\phi_0^i(x)$ is a basis in S^h, (26) is equivalent to solving

$$F(\theta)=b, \tag{29}$$

where $F(\theta)$ is defined by (27) with $v_h=\phi_0^i(x)$ and $b_i=\int_Q f\phi_0^i(x)dx$. Equation (29) can be solved using Newton's method or its modifications. This involves the inversion of the Jacobian corresponding to $F(\theta)$. When using MsFEM, the Jacobian is a matrix assembled on the coarse grid, which gives us the advantage in the computations.

The following convergence result will be shown.

THEOREM 4.1. *Let u_h and u be solutions of* (26) *and* (2)*, respectively. Then*

$$\lim_{h\to 0}\lim_{\epsilon\to 0}\|u_h-u\|_{W_0^{1,p}(Q)}=0 \tag{30}$$

(up to a subsequence) under some nonrestrictive assumptions on $a^*(x,\eta,\xi)$.

Remark 4.1. Since the proof uses G-convergence theory, the limiting a^* (as well as a_0^*) is not unique, and the convergence of the numerical solutions is up to a subsequence in ϵ; i.e., u_h converges to *a* solution of *a* homogenized equation. We note that for the random homogeneous case the limiting operator is unique and the whole sequence converges. In later analysis, all the limits are taken up to a subsequence.

Note that because of the lack of scale separation, the above result cannot be improved, because there are all the scales $\alpha(\epsilon)$, such that $\alpha(\epsilon)\to 0$ as $\epsilon\to 0$ are present. We have observed significant improvement in the numerical calculation when an oversampling technique is used for problems without scale separation. To show that $u_{\epsilon,h}$ approximates u_ϵ in $W_0^{1,p}(Q)$ we will use the corrector results presented in the previous section.

4.2. Numerical homogenization method and MsFEM. To present the relation between the numerical homogenization approach and MsFEM we introduce the multiscale mapping, $E^{MsFEM}: S^h\to V_\epsilon^h$, a one-to-one operator which is constructed in the following way. For each $v_h\in S^h$, $v_{\epsilon,h}$ is the solution of

$$-\mathrm{div}(a_\epsilon(x,\eta^{v_h},Dv_{\epsilon,h}))=0 \quad \text{in } K; \tag{31}$$

in each K, $v_{\epsilon,h} = v_h$ on ∂K, and $\eta^{v_h} = \frac{1}{|K|}\int_K v_h dx$. In [11] the authors introduce MsFEM, where a basis for V_ϵ^h is constructed by mapping a basis of S^h. The extension of this approach to nonlinear problems yields a nonlinear space for the approximation of heterogeneities. Note that $v_{\epsilon,h}$ are uniquely determined because (31) enjoys the monotonicity property. Now the numerical homogenization procedure can be written in the following way. Find $u_h \in S^h$ (consequently, $u_{\epsilon,h} = E^{MsFEM} u_h \in V_\epsilon^h$) such that

$$(A_{\epsilon,h} u_h, v_h) = \int_Q f v_h dx \qquad \forall v_h \in S^h, \tag{32}$$

where $A_{\epsilon,h}$ is given by (26). Later on we will show that $Du_{\epsilon,h}$ approximates Du_ϵ in $L^p(Q)^d$, assuming that the fluxes $a_\epsilon(x,\eta,\xi)$ and $a_{0,\epsilon}(x,\eta,\xi)$ are random homogeneous fields. Clearly, for periodic problems, (31) can be solved in a period of size ϵ and extended periodically to K. This solution will approximate the solution of (31) and can be used in the construction of $A_{\epsilon,h}$ and in setting up (32) (cf. HMM [6]). The convergence analysis for this case can be easily carried out using periodic correctors, and this will be presented elsewhere. Finally, we would like to note that one can adopt the oversampling technique [10] for nonlinear multiscale finite element methods.

4.3. Proof of Theorem 4.1. The proof of Theorem 4.1 will be carried out in the following way. First we show the coercivity of $A_{\epsilon,h}$ defined by (27). Next we study the limit as $\epsilon \to 0$ of (26) and show that the solution of the limiting equation approximates homogenized solutions. For the sake of simplification of the proof, we assume $\beta = p$ in (4).

LEMMA 4.2. *Let $A_{\epsilon,h}$ be defined by* (27). *Then for sufficiently small h, there exists a constant $C > 0$ such that for any $v_h \in S^h$*

$$(A_{\epsilon,h} v_h, v_h) \geq C \|Dv_h\|_{p,Q}^p - C_1.$$

Proof. Let $\tilde{v}_{\epsilon,h} = v_{\epsilon,h} - v_h$. It follows that $\tilde{v}_{\epsilon,h} \in W_0^{1,p}(K)$ satisfies the following problem:

$$-\text{div } a_\epsilon(x, \eta^{v_h}, D\tilde{v}_{\epsilon,h} + Dv_h) = 0 \quad \text{ in } K. \tag{33}$$

Using (33) and applying Green's theorem and (7), we have the following estimate:

$$\begin{aligned}
(A_{\epsilon,h} v_h, v_h) &= \sum_K \int_K [(a_\epsilon(x, \eta^{v_h}, Dv_{\epsilon,h}), Dv_h) + a_{0,\epsilon}(x, \eta^{v_h}, Dv_{\epsilon,h}) v_h] dx \\
&= \sum_K \int_K [(a_\epsilon(x, \eta^{v_h}, Dv_h + D\tilde{v}_{\epsilon,h}), Dv_h + D\tilde{v}_{\epsilon,h}) + a_{0,\epsilon}(x, \eta^{v_h}, Dv_{\epsilon,h}) v_h] dx \\
&= \sum_K \int_K [(a_\epsilon(x, \eta^{v_h}, Dv_h + D\tilde{v}_{\epsilon,h}), Dv_h + D\tilde{v}_{\epsilon,h}) + a_{0,\epsilon}(x, \eta^{v_h}, Dv_{\epsilon,h}) \eta^{v_h}] dx \\
&\quad + \sum_K \int_K a_{0,\epsilon}(x, \eta^{v_h}, Dv_{\epsilon,h})(v_h - \eta^{v_h}) dx \\
&\geq C \sum_K \int_K |Dv_h + D\tilde{v}_{\epsilon,h}|^p dx - Ch \left(1 + \sum_K \int_K |Dv_h|^p dx\right) - C_1.
\end{aligned}$$

Here we have also used the fact that $\int_K |\eta^{v_h}|^p dx \le C \int_K |v_h|^p dx$. Next we will show that

$$\sum_K \int_K |Dv_h + D\tilde{v}_{\epsilon,h}|^p dx = \sum_K \int_K |Dv_{\epsilon,h}|^p \, dx \ge C \sum_K \int_K |Dv_h|^p dx.$$

We note that v_h is piecewise linear on ∂K for triangular mesh, i.e., $v_{\epsilon,h}|_{\partial K} = v_h = \beta + (Dv_h, x - x_0)$, for some constants β and x_0 independent of Dv_h. We set $\bar{v}_{\epsilon,h} = v_{\epsilon,h} - \beta$. Then, by change of variable and homogeneity argument and applying the trace theorem, we have

$$\begin{aligned}\sum_K \int_K |Dv_{\epsilon,h}|^p dx &\ge C \sum_K \frac{h^d}{h^p} \int_{K_r} |D_y \bar{v}_{\epsilon,h}|^p dy \\ &\ge C \sum_K \frac{h^d}{h^p} \int_{\partial K_r} |(Dv_h, y\, h)|^p dy = C \sum_K h^d |Dv_h|^p C(e_{Dv_h}),\end{aligned}$$

where K_r is a reference triangle of size of order 1, e_{Dv_h} is the unit vector in the direction of Dv_h, and

$$C(e_{Dv_h}) = \int_{\partial K_r} |(e_{Dv_h}, y)|^p dy.$$

Here we have used the trace inequality, $\|u\|_{L^p(\partial Q)} \le C \|u\|_{W^{1,p}(Q)}$, and taken into account the equivalence of finite dimensional norms for every h. One can further show that $C(e_\xi)$ is bounded from below independent of ξ and h. By contradiction, suppose that the claim is not true. Then there exists a sequence $\{e_{\xi_n}\}$ which has a subsequence (denoted by the same notation) such that $e_{\xi_n} \to e_*$ and $C(e_{\xi_n}) \to 0$ as $n \to \infty$. Since $C(e_\xi)$ is continuous, it follows that $C(e_*) = 0$. This further implies that $(e_*, y) = 0$ on ∂K_r, and hence $e_* = 0$. This is a contradiction. ∎

Next we show that $A_{\epsilon,h}$ is equicontinuous for any h in any compact set.

LEMMA 4.3. *For any $v_h \in S^h$ and $w_h \in S^h$ in a compact set we have*

$$\|A_{\epsilon,h} v_h - A_{\epsilon,h} w_h\|^p \le C \left(\sum_K \int_K (|D(v_h - w_h)|^p + \nu(|\eta^{v_h} - \eta^{w_h}|)) dx \right)^{1/p},$$

where C does not depend on ϵ.

Since this result is for fixed h (i.e., finite dimensional), we do not specify the norm.

Proof.

$$\begin{aligned}\|A_{\epsilon,h} v_h - A_{\epsilon,h} w_h\| &= \sum_K \int_K |a_\epsilon(x, \eta^{v_h}, Dv_{\epsilon,h}) - a_\epsilon(x, \eta^{w_h}, Dw_{\epsilon,h})| dx \\ &\quad + \sum_K \int_K |a_{0,\epsilon}(x, \eta^{v_h}, Dv_{\epsilon,h}) - a_{0,\epsilon}(x, \eta^{w_h}, Dw_{\epsilon,h})| dx.\end{aligned} \tag{34}$$

Next we will estimate the first term on the r.h.s. of (34). The estimate for the second

term is analogous.

$$
\begin{aligned}
&\sum_K \int_K |a_\epsilon(x,\eta^{v_h},Dv_{\epsilon,h}) - a_\epsilon(x,\eta^{w_h},Dw_{\epsilon,h})|dx \\
&\quad \le C\sum_K \int_K (1+|\eta^{v_h}|+|\eta^{w_h}|+|Dv_{\epsilon,h}|+|Dw_{\epsilon,h}|)^{p-1}\nu(|\eta^{v_h}-\eta^{w_h}|) \\
&\quad + C\sum_K \int_K (1+|\eta^{v_h}|+|\eta^{w_h}|+|Dv_{\epsilon,h}|+|Dw_{\epsilon,h}|)^{p-1-s}|Dv_{\epsilon,h}-Dw_{\epsilon,h}|^s \\
&\quad \le C\left(\sum_K \int_K \nu(|\eta^{v_h}-\eta^{w_h}|)^p dx\right)^{1/p} + C\left(\sum_K \int_K |Dv_{\epsilon,h}-Dw_{\epsilon,h}|^p\right)^{1/p}.
\end{aligned} \tag{35}
$$

Here we have used the Cauchy inequality along with the facts that $\|Dv_{\epsilon,h}\|_{p,K} \le C\|Dv_h\|_{p,K}$, $\|Dw_{\epsilon,h}\|_{p,K} \le C\|Dw_h\|_{p,K}$, and $\|Dv_h\|_{p,Q} \le C$, $\|Dw_h\|_{p,Q} \le C$. It remains to estimate the second term on the r.h.s. of (35).

(36)

$$
\begin{aligned}
&\sum_K \int_K |Dv_{\epsilon,h}-Dw_{\epsilon,h}|^p \\
&\quad \le C\sum_K \int_K (a_\epsilon(x,\eta^{v_h},Dv_{\epsilon,h}) - a_\epsilon(x,\eta^{v_h},Dw_{\epsilon,h}), Dv_{\epsilon,h}-Dw_{\epsilon,h})dx \\
&\quad \le \sum_K \int_K (a_\epsilon(x,\eta^{v_h},Dw_{\epsilon,h}) - a_\epsilon(x,\eta^{w_h},Dw_{\epsilon,h}), Dv_{\epsilon,h}-Dw_{\epsilon,h})dx \\
&\qquad + \sum_K \int_K (a_\epsilon(x,\eta^{w_h},Dw_{\epsilon,h}) - a_\epsilon(x,\eta^{v_h},Dw_{\epsilon,h}), Dv_{\epsilon,h}-Dw_{\epsilon,h})dx \\
&\quad \le C\sum_K \int_K (a_\epsilon(x,\eta^{v_h},Dv_{\epsilon,h}) - a_\epsilon(x,\eta^{w_h},Dw_{\epsilon,h}), Dv_h + D\tilde{v}_{\epsilon,h} - Dw_h - D\tilde{w}_{\epsilon,h})dx \\
&\qquad + C\sum_K \int_K (a_\epsilon(x,\eta^{w_h},Dw_{\epsilon,h}) - a_\epsilon(x,\eta^{v_h},Dw_{\epsilon,h}), Dv_{\epsilon,h}-Dw_{\epsilon,h})dx \\
&\quad \le C\sum_K \int_K (a_\epsilon(x,\eta^{v_h},Dv_{\epsilon,h}) - a_\epsilon(x,\eta^{w_h},Dw_{\epsilon,h}), Dv_h - Dw_h)dx \\
&\qquad + C\sum_K \int_K \nu(|\eta^{w_h}-\eta^{v_h}|)^p dx \\
&\quad \le C\left(\sum_K \int_K |Dv_h-Dw_h|^p dx\right)^{1/p} + C\sum_K \int_K \nu(|\eta^{w_h}-\eta^{v_h}|)^p dx.
\end{aligned}
$$

Here we have used Holder and Cauchy inequalities along with the facts that $\|Dv_{\epsilon,h}\|_{p,K} \le C\|Dv_h\|_{p,K}$, $\|Dw_{\epsilon,h}\|_{p,K} \le C\|Dw_h\|_{p,K}$, and $\|Dv_h\|_{p,Q} \le C$, $\|Dw_h\|_{p,Q} \le C$, and that $v_{\epsilon,h} = v_h + \tilde{v}_{\epsilon,h}$, where $\tilde{v}_{\epsilon,h} \in W_0^{1,p}(K)$ satisfies $-\text{div } a_\epsilon(x,\eta^{v_h}, Dv_h + D\tilde{v}_{\epsilon,h}) = 0$. The estimates (35) and (36) give us the estimate for the first term of the r.h.s. of (34). A similar estimate for the second term can be obtained in a very analogous manner. □

The coercivity and continuity of $A_{\epsilon,h}$ guarantee the existence of a solution for the

discrete equation

$$(A_{\epsilon,h}u_{\epsilon,h}, w_h) = \int_Q f w_h dx. \tag{37}$$

LEMMA 4.4. *For any $v_h \in S^h$ and $w_h \in S^h$*

$$\lim_{\epsilon\to 0}(A_{\epsilon,h}v_h, w_h) = (A_h v_h, w_h)$$

(up to a sub-sequence), where the r.h.s. is defined as

$$(A_h v_h, w_h) = \sum_K \int_K [(a^*(x, \eta^{v_h}, Dv_0), Dw_h) + a_0^*(x, \eta^{v_h}, Dv_0)w_h]dx$$

and v_0 is the solution of $v_0 - v_h \in W_0^{1,p}(K)$,

$$-\mathrm{div}(a^*(x, \eta^{v_h}, Dv_0)) = 0.$$

Here $a^*(x, \eta, \xi)$ is a G-limit of the corresponding limit operator.

Proof. Using the theorem on G-convergence of arbitrary solutions [17, p. 87], we obtain that solutions $v_{\epsilon,h}$ of (31) weakly converge to v_0 in $W^{1,p}(K)$, and $a_\epsilon(x, \eta^{v_h}, Dv_{\epsilon,h})$ weakly converges to $a^*(x, \eta^{v_h}, Dv_0)$ in $L^q(K)^d$, and $a_{0,\epsilon}(x, \eta^{v_h}, Dv_{\epsilon,h})$ weakly converges to $a_0^*(x, \eta^{v_h}, Dv_0)$ in $L^q(K)$ (up to a subsequence), where $a^*(x, \eta, \xi)$ and $a_0^*(x, \eta, \xi)$ are the fluxes corresponding to a G-limit of the original operators. Thus,

$$\lim_{\epsilon\to 0}(A_{\epsilon,h}v_h, w_h) = \lim_{\epsilon\to 0}\sum_K \int_K [(a_\epsilon(x, \eta^{v_h}, Dv_{\epsilon,h}), Dw_h) + a_{0,\epsilon}(x, \eta^{v_h}, Dv_{\epsilon,h})w_h]dx$$

$$= \sum_K \int_K [(a^*(x, \eta^{v_h}, Dv_0), Dw_h) + a_0^*(x, \eta^{v_h}, Dv_0)w_h]dx = (A_h v_h, w_h). \qquad \square$$

It can be easily shown that A_h is coercive for small h. Since $A_{\epsilon,h}$ is equicontinuous in any compact set, the results of Lemma 4.4 hold for any $v_h \in S^h$ and $w_h \in S^h$ that are uniformly bounded (finite dimensional). Thus, taking the limit $\epsilon \to 0$ of (37) (up to a subsequence), we obtain

$$(A_h u_h, w_h) = \int_Q f w_h dx.$$

From Lemma 4.2 and the continuity of A_h (which can be easily verified) it follows that u_h exists and is uniformly bounded in $W_0^{1,p}(Q)$, and thus $u_h \to u$ (up to a subsequence) weakly in $W_0^{1,p}(Q)$. Our task to show that u is a solution of the homogenized equation. The following lemma is needed for this purpose.

LEMMA 4.5. *Assume that $v_h \in S^h$ and Dv_h are uniformly bounded in $L^{p+\alpha}(Q)^d$ (with $\alpha > 0$) and $w_h \in S^h$ and Dw_h are uniformly bounded in $L^p(Q)^d$. Then*

$$\lim_{h\to 0}(A_h v_h - A^* v_h, w_h) = 0,$$

where

$$(A^* v_h, w_h) = \sum_K \int_K [(a^*(x, v_h, Dv_h), Dw_h) + a_0^*(x, v_h, Dv_h)w_h]dx.$$

Proof.

$$(A_h v_h - A^* v_h, w_h) = \sum_K \int_K [(a^*(x, \eta^{v_h}, Dv_0) - a^*(x, v_h, Dv_h), Dw_h) + (a_0^*(x, \eta^{v_h}, Dv_0) - a_0^*(x, v_h, Dv_h))w_h]dx. \tag{38}$$

Next we will show that the first and second terms on the r.h.s. of (38) converge to zero. For the first term we have

(39)

$$\begin{aligned}
&\sum_K \int_K (a^*(x, \eta^{v_h}, Dv_0) - a^*(x, v_h, Dv_h), Dw_h)dx \\
&\quad \le C \sum_K \int_K \nu(|v_h - \eta^{v_h}|)(1 + |\eta^{v_h}|^{p-1} + |v_h|^{p-1} + |Dv_0|^{p-1} + |Dv_h|^{p-1})|Dw_h|dx \\
&\qquad + C \sum_K \int_K (1 + |\eta^{v_h}| + |v_h| + |Dv_0| + |Dv_h|)^{p-1-s}|Dv_h - Dv_0|^s |Dw_h|dx \\
&\quad \le C \sum_K \left(\int_K \nu(|v_h - \eta^{v_h}|)^q (1 + |Dv_h|^p)dx \right)^{1/q} \left(\int_K |Dw_h|^p dx \right)^{1/p} \\
&\qquad + C \sum_K \left(\int_K (1 + |Dv_h|^p)dx \right)^{(p-qs)/pq} \left(\int_K |D(v_h - v_0|^p dx \right)^{s/p} \left(\int_K |Dw_h|^p dx \right)^{1/p} \\
&\quad = C \left(\int_Q \nu(|v_h - \eta^{v_h}|)^q (1 + |Dv_h|^p)dx \right)^{1/q} + C \left(\int_Q |D(v_h - v_0)|^p dx \right)^{s/p}.
\end{aligned}$$

Here we have used the Cauchy inequality along with the facts that $\|Dw_h\|_{p,Q} \le C$, $\int_K |Dv_0|^p dx \le C \int_K |Dv_h|^p dx$, $\int_K |v_h|^p dx \le C \int_K |Dv_h|^p dx$, and $\int_K |\eta^{v_h}|^p dx \le C \int_K |Dv_h|^p dx$. Next we will show that $\|Dv_h - Dv_0\|_{p,Q} \to 0$ as $h \to 0$ under some assumptions regarding the regularity of $a^*(x, \eta, \xi)$ with respect to spatial variables. Moreover, this convergence is uniform for a uniformly bounded family of Dv_h. Define $\overline{a^*}_K(x, \eta, \xi)$ as a piecewise constant function on each K and η, ξ defined in each K by

$$\overline{a^*}_K(\eta, \xi) = \frac{1}{|K|} \int_K a^*(x, \eta, \xi).$$

We assume that in each K

$$|a^*(x, \eta, \xi) - \overline{a^*}_K(\eta, \xi)| \le \alpha_h (1 + |\eta|^{p-1} + |\xi|^{p-1}), \tag{40}$$

where α_h is a generic sequence such that $\alpha_h \to 0$ as $h \to 0$ and is independent of K. For example, this condition is satisfied if $a^*(x, \eta, \xi)$ is a Holder function with respect to spatial variables. Note that for random homogeneous operators, (40) trivially holds because a^* is independent of x.

Then

$$\begin{aligned}
\|Dv_h - Dv_0\|^p_{p,Q} &\le C\sum_K \int_K (a^*(x,\eta^{v_h},Dv_h) - a^*(x,\eta^{v_h},Dv_0), Dv_h - Dv_0)dx \\
&= C\sum_K \int_K (a^*(x,\eta^{v_h},Dv_h) - \overline{a^*}_K(\eta^{v_h},Dv_h), Dv_h - Dv_0)dx \\
&\quad + C\sum_K \int_K (\overline{a^*}_K(\eta^{v_h},Dv_h) - a^*(x,\eta^{v_h},Dv_0), Dv_h - Dv_0)dx \\
&= C\sum_K \int_K (a^*(x,\eta^{v_h},Dv_h) - \overline{a^*}_K(\eta^{v_h},Dv_h), Dv_h - Dv_0)dx \\
&\le C\alpha_h \sum_K \int_K (1 + |\eta^{v_h}|^{p-1} + |Dv_h|)^{p-1}|Dv_h - Dv_0|dx \\
&\le C\alpha_h (1 + \|Dv_h\|^{p/q}_{p,Q})\|Dv_h - Dv_0\|_{p,Q}.
\end{aligned}$$

From here it follows that $\|Dv_h - Dv_0\|_{p,Q} \to 0$ as $h \to 0$.

We note that the r.h.s. of (39) converges to zero because $\|Dv_h - Dv_0\|_{p,Q} \to 0$ as $h \to 0$ and because of Lemma 3.4. Thus, the first term on the r.h.s. of (38) converges to zero.

For the second term on the r.h.s. of (38) we have

(41)

$$\begin{aligned}
&\sum_K \int_K (a_0^*(x,\eta^{v_h},Dv_0) - a_0^*(x,v_h,Dv_h))w_h dx \\
&\quad \le C\sum_K \int_K \nu(|v_h - \eta^{v_h}|)(1 + |\eta^{v_h}|^{p-1} + |v_h|^{p-1} + |Dv_0|^{p-1} + |Dv_h|^{p-1})|w_h|dx \\
&\qquad + C\sum_K \int_K (1 + |\eta^{v_h}| + |v_h| + |Dv_0| + |Dv_h|)^{p-1-s}|Dv_h - Dv_0|^s|w_h|dx.
\end{aligned}$$

Clearly one can do the same manipulations as the those for the first term of the r.h.s. of (38) to show that the r.h.s. of (41) converges to zero as $h \to 0$. □

For the proof of Theorem 4.1 we assume that Du_h is uniformly bounded in $L^{p+\alpha}(Q)^d$ for some $\alpha > 0$. One can assume additional nonrestrictive regularity assumptions [16] for input data and obtain Meyers-type estimates, $\|Du\|_{p+\alpha,Q} \le C$, for the homogenized solutions. In this case it is reasonable also to assume that the discrete solutions are uniformly bounded in $L^{p+\alpha}(Q)^d$. We have obtained results on Meyers-type estimates for our approximate solutions in the case $p = 2$ (see [8]). We are currently studying the generalization of these results to arbitrary p. One can impose different kinds of assumptions for which the Lemma 4.5 holds without assuming that Du_h is uniformly bounded in $L^{p+\alpha}(Q)^d$, e.g.,

$$|a^*(x,\eta,\xi) - a^*(s,\eta',\xi)| \le C(1 + |\eta|^{p-1} + |\eta'|^{p-1} + |\xi|^{p-1-r})|\eta - \eta'|^r$$

$(0 < r < 1)$.

To conclude the proof of Theorem 4.1 we note that $u_h \to u$ (up to a sub-sequence) weakly in $W_0^{1,p}(Q)$, and our goal is to show that u is a solution of the homogenized equation. Using Lemma 4.5, we obtain that

$$(A_h u_h - A^* u_h, v_h) = \int_Q f v_h dx - (A^* u_h, v_h).$$

Thus it follows from Lemma 4.5 that $A^*u_h \to f$ weakly in $W^{-1,p}(Q)$. Moreover, using Lemma 4.5, we obtain that $(A^*u_h, u_h) \to \int_Q f v_h dx$. Since the operator A^* is of type M (see [20]), we obtain that $A^*u = f$; i.e., u is a solution of a homogenized equation. Moreover, A^* is also of type S_+ (see [21]), which allows us to state that $u_h \to u$ *strongly* in $W_0^{1,p}(Q)$.

Remark 4.2. We would like to note that in the periodic and random homogeneous cases Theorem 4.1 holds in the limit $\epsilon/h \to 0$; i.e., $h = h(\epsilon) \gg \epsilon$. This will be presented elsewhere.

Remark 4.3. Finally we would like to note that Theorem 4.1 is proved under the assumptions (40) and $\|Du_h\|_{p+\alpha,Q} \le C$, $\alpha > 0$. The latter has been shown for $p = 2$ in [8].

4.4. Approximation of the oscillations. In order to approximate solutions u_ϵ in the $W^{1,p}$-norm, we assume $a_\epsilon(x,\eta,\xi) = a(T(x/\epsilon),\eta,\xi)$ and $a_{0,\epsilon}(x,\eta,\xi) = a_0(T(x/\epsilon),\eta,\xi)$. Then the following theorem holds.

THEOREM 4.6.

$$\lim_{h\to 0}\lim_{\epsilon\to 0}\|D(u_{\epsilon,h} - u_\epsilon)\|_{p,Q} = 0,$$

where $u_{\epsilon,h} = E^{MsFEM}u_h$, *defined by* (31) *(or* (28) *in each* K*).*

Proof. Because of Theorem 3.1 we need to show only that

$$\lim_{h\to 0}\lim_{\epsilon\to 0}\|Du_{\epsilon,h} - P(T(x/\epsilon)\omega, M_h u, M_h Du)\|_{p,Q} = 0.$$

Similarly,

$$\lim_{\epsilon\to 0}\|Du_{\epsilon,h} - P(T(x/\epsilon)\omega, M_h u_h, M_h Du_h)\|_{p,K} = 0. \tag{42}$$

Equation (42) follows from the fact that $-\mathrm{div}(a^*(\eta^{u_h}, D_x u_h)) = 0$, i.e., the homogenized solution for $u_{\epsilon,h}$ is u_h. Consequently,

$$\lim_{\epsilon\to 0}\|Du_{\epsilon,h} - P(T(x/\epsilon)\omega, M_h u_h, M_h Du_h)\|_{p,Q} = 0.$$

It remains to show that

$$\lim_{h\to 0}\lim_{\epsilon\to 0}\|P(T(x/\epsilon)\omega, M_h u_h, M_h Du_h) - P(T(x/\epsilon)\omega, M_h u, M_h Du)\|_{p,Q} = 0.$$

To show this we need the estimate for $\int_\Omega |P(\omega,\eta_1,\xi_1) - P(\omega,\eta_2,\xi_2)|^p d\mu(\omega)$. Define $P_1 = P(\omega,\eta_1,\xi_1)$ and $P_2 = P(\omega,\eta_2,\xi_2)$. Then

$$\begin{aligned}
\int_\Omega |P_1 - P_2|^p d\mu(\omega) &\le C\int_\Omega (a(\omega,\eta_1,P_1) - a(\omega,\eta_1,P_2), P_1 - P_2) d\mu(\omega)\\
&= \int_\Omega (a(\omega,\eta_1,P_1) - a(\omega,\eta_2,P_2), P_1 - P_2) d\mu(\omega)\\
&\quad + \int_\Omega (a(\omega,\eta_2,P_2) - a(\omega,\eta_1,P_2), P_1 - P_2)) d\mu(\omega)\\
&\le \int_\Omega (a(\omega,\eta_1,P_1) - a(\omega,\eta_2,P_2), \xi_1 - \xi_2) d\mu(\omega)\\
&\quad + \frac{C}{\delta_1}\int_\Omega (1 + |\eta_1|^p + |\eta_2|^p + |P_2|^p)\nu(|\eta_1 - \eta_2|) d\mu(\omega)\\
&\quad + C\delta_1 \int_\Omega |P_1 - P_2|^p d\mu(\omega).
\end{aligned}$$

Choosing δ_1 appropriately small and using (10), we have

(43)
$$\int_\Omega |P_1 - P_2|^p d\mu(\omega)$$
$$\leq (a^*(\eta_1,\xi_1) - a^*(\eta_2,\xi_2), \xi_1 - \xi_2) + C\int_\Omega (1 + |\eta_1|^p + |\eta_2|^p + |P_2|^p)\nu(|\eta_1 - \eta_2)d\mu(\omega).$$

Using (43), we have

$$\lim_{h\to 0}\lim_{\epsilon\to 0} \|P(T(x/\epsilon)\omega, M_h u_h, M_h Du_h) - P(T(x/\epsilon)\omega, M_h u, M_h Du)\|_{p,Q}$$
$$\leq \lim_{h\to 0}\sum_K \int_K (a^*(M_h u_h, M_h Du_h) - a^*(M_h u, M_h Du), M_h Du_h - M_h Du)dx$$
$$+ C\lim_{h\to 0}\sum_K \int_K (1 + |M_h u_h| + |M_h u| + |M_h Du|)^p \nu(|M_h u_h - M_h u|)dx.$$

The r.h.s. of (4.4) converges to zero, which can be established in a manner similar to the convergence analysis of the r.h.s. of (20). □

5. Numerical results. Our first numerical example is a nonlinear convection diffusion equation in two dimensions:

$$\frac{1}{\epsilon} v(T(x/\epsilon)\omega)\cdot DF(u_\epsilon) - d\Delta u_\epsilon = f, \tag{44}$$

where div$v = 0$. Assuming that there exists a homogeneous stream function $\mathcal{H}(T(x/\epsilon)\omega)$,

$$\mathcal{H} = \begin{pmatrix} 0 & H(T(x/\epsilon)\omega) \\ -H(T(x/\epsilon)\omega) & 0 \end{pmatrix},$$

such that div$\mathcal{H} = v$, we obtain

$$\text{div}(-d\delta_{ij} Du_\epsilon + \mathcal{H}(T(x/\epsilon)\omega)DF(u_\epsilon)) = f$$

or

$$-\text{div}(a(T(x/\epsilon)\omega, u_\epsilon)Du_\epsilon) = f,$$

where

$$a = \begin{pmatrix} -d & H(T(x/\epsilon)\omega)F'(u) \\ -H(T(x/\epsilon)\omega)F'(u) & -d \end{pmatrix}.$$

We assume that a satisfies the assumptions imposed in previous sections. The auxiliary problem is defined as follows: $w_{\eta,\xi}(\omega) \in V^p_{pot}$ is the solution of

$$\text{div}(a(T(y)\omega, \eta)(\xi + w_{\eta,\xi}(T(y)\omega))) = 0.$$

Introducing $N^j_\eta(T(y)\omega) \in V^p_{pot}$ such that $w^i_{\eta,\xi}(T(y)\omega) = N^{ij}_\eta(T(y)\omega)\xi_j$, we have

$$\text{div}(a(T(y)\omega, \eta)(I + N_\eta(T(y)\omega))) = 0,$$

where I is the identity matrix. Using $w_{\eta,\xi}$, we can compute the homogenized operator that is given by

$$\operatorname{div}(a^*(u)Du) = f,$$

where $a^*_{ij}(\eta) = -d\delta_{ij} + \langle H_{ik}F'(\eta)N^{kj}_\eta\rangle$. Here we have taken into account that $\langle N\rangle = 0$ since $N \in V^p_{pot}$. The term $\langle H_{ik}F'(\eta)N^{kj}_\eta\rangle$ can be regarded as an enhanced diffusion due to heterogeneous convection, similar to the linear case [9]. A numerical corrector is defined as

$$\mathcal{P}_\epsilon = M_h Du(I + w_{M_h u}(T(x/\epsilon)\omega)).$$

Next we present numerical examples where the enhanced diffusivity is approximately computed locally. It is more transparent for this purpose to use a parabolic equation,

$$\frac{\partial u_\epsilon}{\partial t} + \frac{1}{\epsilon}v(T(x/\epsilon)\omega)\cdot DF(u_\epsilon) = d\Delta u_\epsilon. \tag{45}$$

Using general G-convergence theory, we have the following equation for the homogenized solution:

$$\frac{\partial u}{\partial t} = \operatorname{div}(a^*(u)u), \tag{46}$$

where $a^*(\eta)$ is the homogenized operator derived from the elliptic problem shown above. In particular, $a^*_{ij} = d\delta^{ij} + a^c_{ij}$, where $a^c_{ij}(\eta) = -\langle H_{ik}F'(\eta)N^{kj}_\eta\rangle$ is the enhanced diffusion due to nonlinear heterogeneous convection. It can be shown that the corrector has the same form as in the elliptic case

$$P(T(x/\epsilon)\omega, M_h u(t,x), M_h Du(t,x)) = M_h Du(t,x)(I + w_{M_h u(t,x)}(T(x/\epsilon)\omega)),$$

i.e., all the time dependence is in the homogenized solution. The proof is the same as in the elliptic case.

To illustrate the significance of the enhanced diffusion, we present some numerical examples. Numerical tests are performed using the finite element method. First we present the total diffusivity as a function of η (i.e., average of the solution) for two different heterogeneous velocity fields given by the stream functions $H = \sin(2\pi y/\epsilon) + \sin(2\sqrt{(}2)\pi y/\epsilon)$. We take $\epsilon = 0.1$ and $d = 0.1$ (molecular diffusion). The flux function is chosen to be the Buckley–Leverett function $F(u) = u^2/(u^2 + 0.2(1-u)^2))$ motivated by porous media flows. The enhanced diffusion is computed by solving the problem in a unit square, and thus it is only an approximate value of it. In Figure 1 we plot the total diffusivity. The left plot in this figure represents the total diffusivity in the horizontal direction (along the layers), and the right plot represents the total diffusivity in the vertical direction. Clearly, the diffusion is enhanced dramatically in the horizontal direction, that is, along the convection (note the ten-fold difference between the y-axis scales). As we see for $\eta \approx 0.4$, there is a 15-fold increase in the diffusion relative to molecular diffusion, d. Moreover, since $F'(0) = F'(1) = 0$, there is no enhancement if $\eta = 0$ or $\eta = 1$ (this corresponds to pure phases). For the cellular flow, $H(x,y) = \sin(2\pi y/\epsilon)\sin(2\pi x/\epsilon)$, we obtain isotropic diffusion, which is shown in Figure 2.

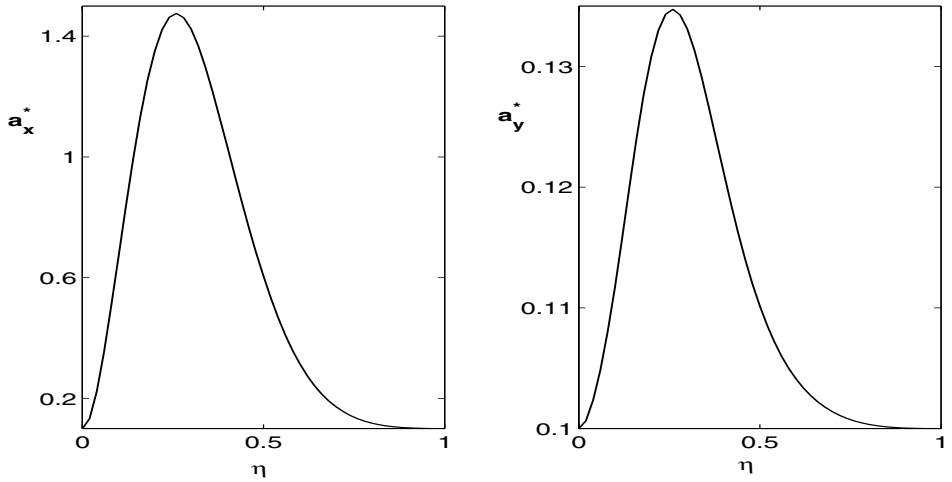


FIG. 1. *Horizontal (left) and vertical (right) effective diffusivity for the layered media with stream function* $H(x,y) = \sin(2\pi y/\epsilon)$ *and flux function* $F(u) = u^2/(u^2 + 0.2(1-u)^2)$.

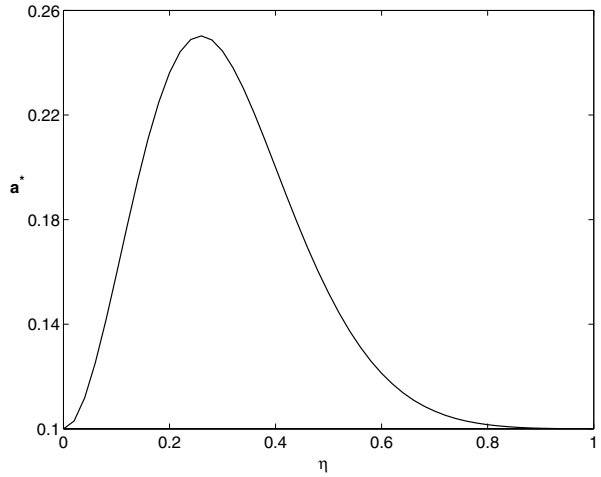


FIG. 2. *Effective diffusivity for the isotropic media with stream function* $H(x,y) = \sin(2\pi y/\epsilon) \times \sin(2\pi x/\epsilon)$ *and flux function* $F(u) = u^2/(u^2 + 0.2(1-u)^2)$.

The next set of numerical examples is designed to compare the solutions of the original (fine scale equation) with the solutions of the equations obtained using numerical homogenization with and without enhanced diffusion. Our goal here is to illustrate the importance of nonlinear enhanced diffusion. We consider

$$\frac{\partial u_\epsilon}{\partial t} + \frac{1}{\epsilon} v_\epsilon \cdot DF(u_\epsilon) = d\Delta u_\epsilon \tag{47}$$

in a unit square domain with boundary and initial conditions as follows: $u_\epsilon = 1$ at the inlet ($x_1 = 0$), $u_\epsilon = 0$ at the outlet ($x_1 = 1$), and no flow boundary conditions on the lateral sides $x_2 = 0$ and $x_2 = 1$. We have tested various heterogeneous fields for the velocity, and we present here a result for the layered flow, $H = \sin(2\pi y/\epsilon)$.

In Figure 3 we plot the average (over the whole domain) of the solutions of (47) ($\frac{1}{Q}\int_Q u_\epsilon(x,t)dx$) as a function of time. We compare the fine scale solution with the

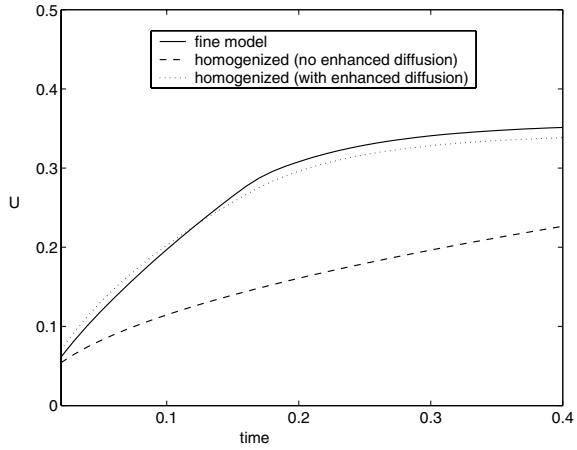


FIG. 3. *Comparison of the average u over the whole domain for three problems:* (1) *fine scale (designated with a solid line),* (2) *homogenized solution with no enhanced diffusion (designated with a dashed line), and* (3) *homogenized solution with enhanced diffusion (designated with a dotted line). In this case* $H(x,y) = \sin(2\pi y/\epsilon)$ *and* $F(u) = u^2/(u^2 + 0.2(1-u)^2)$.

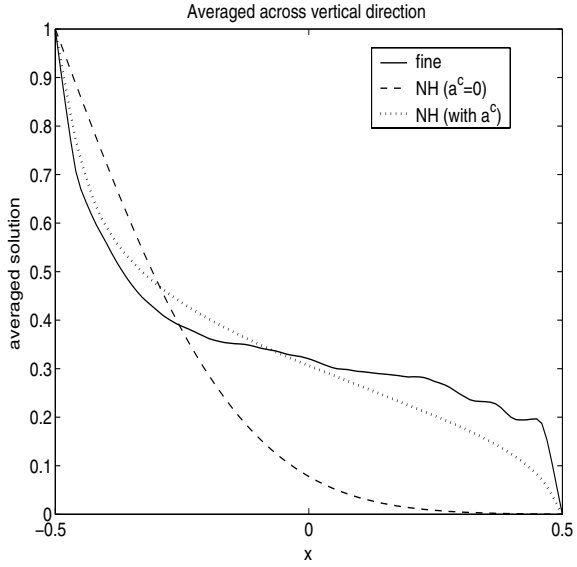


FIG. 4. *Vertical average (across the heterogeneities) of the solution for the layered media with stream function* $H(x,y) = \sin(2\pi y/\epsilon)$ *and flux function* $F(u) = u^2/(u^2 + 0.2(1-u)^2)$ *at time* 0.4.

coarse scale solutions where the enhanced diffusivity is taken into account; i.e., it can be considered as a numerical homogenization procedure with one coarse block. We also consider the coarse scale solution where the enhanced diffusion is neglected, i.e, $u_t = d\Delta u$. As we see from this figure, the solution computed with enhanced diffusion performs well and gives a reasonable approximation of the fine scale solution. On the other hand, the average solution that does not account for enhanced diffusion performs very poorly. In Figure 4 we plot the average along the horizontal direction (x_2)

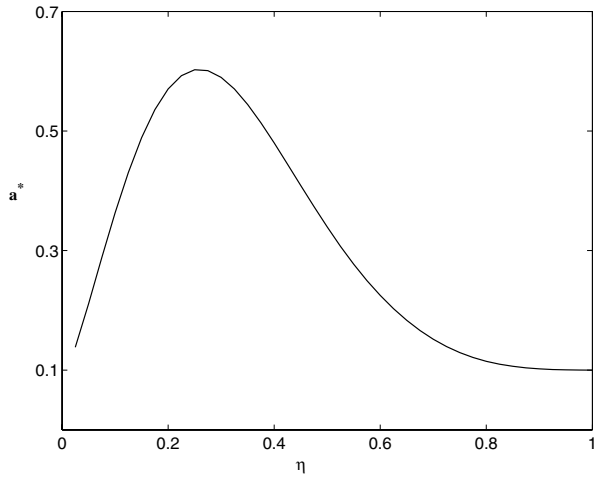


FIG. 5. *Effective diffusivity for the isotropic media with Gaussian stream function which has correlation lengths* $l_x = l_y = 0.1$*, mean zero, and variance* 0.5*. The flux function is chosen as* $F(u) = u^2/(u^2 + 0.2(1-u)^2)$ *and* $d = 0.1$ *(see* (44)*).*

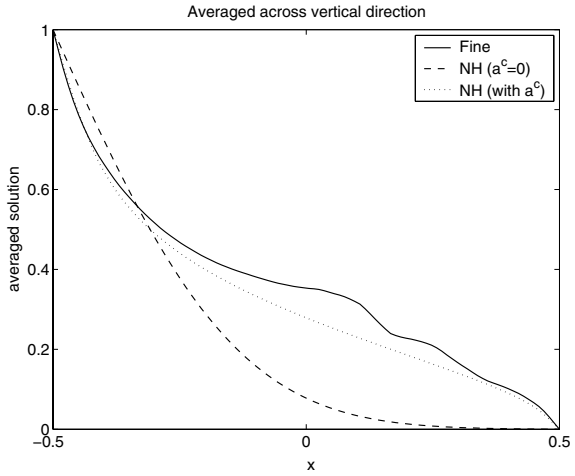


FIG. 6. *Average of the solution along the horizontal direction at* $t = 0.4$ *for Gaussian stream function which has correlation lengths* $l_x = l_y = 0.1$*, mean zero, and variance* 0.5*. The flux function is chosen as* $F(u) = u^2/(u^2 + 0.2(1-u)^2)$ *and* $d = 0.1$ *(see* (44)*).*

(across heterogeneities) of the solutions at time 0.4. The figure clearly indicates the importance of having enhanced diffusion in the homogenized setting of the problem. Next we present an example where the stream function $H(x, y)$ is a realization of the random field with Gaussian distribution. To generate a realization of the random field with prescribed correlation lengths, we use GSLIB [5]. In particular, $H(x, y)$ is a realization of a Gaussian field with correlation lengths $l_x = l_y = 0.1$, and with mean zero and variance 0.5. Here $d = 0.1$ and $F(u) = u^2/(u^2+0.2(1-u)^2)$ are used in (44). In Figure 5 we plot the total diffusivity. As we see, the enhancement of the diffusion can be up to 6 times the molecular diffusion, d. Since the stream field is isotropic, it is sufficient to consider total diffusion in one direction. In Figure 6 we plot a cross section

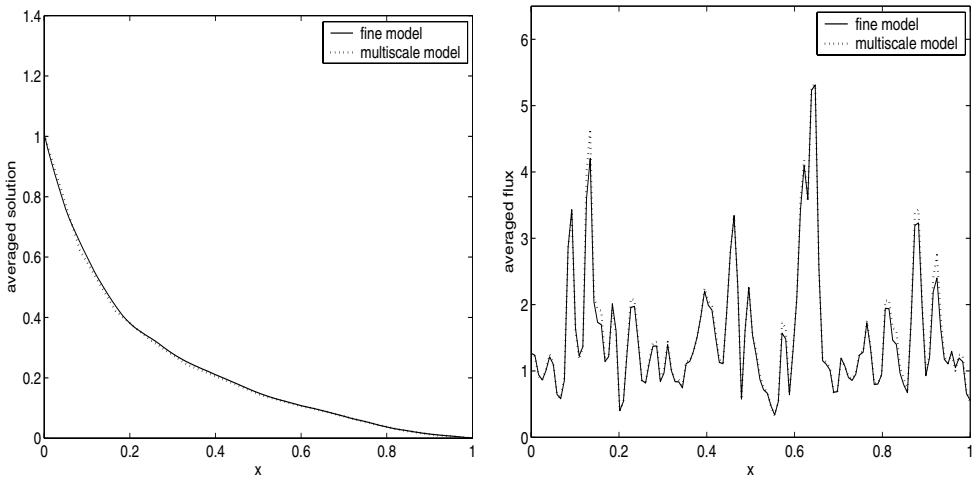


FIG. 7. *Left: the solutions are averaged in the vertical direction. Right: the fluxes are averaged in the vertical direction.*

of the solution at time 0.4. These results clearly indicate the importance of enhanced diffusion. For different realizations of the random field we have observed similar results. We note that this example can be easily generalized to nonlinear convection diffusion of more general form, $\frac{1}{\epsilon}v(T(x/\epsilon)\omega)\cdot DF(u_\epsilon) - \mathrm{div}(a(T(x/\epsilon)\omega, u_\epsilon, Du_\epsilon)) = f$.

Finally, we consider an application of the numerical homogenization procedure to Richards equations, $\mathrm{div}(a_\epsilon(x,u_\epsilon)D_x u_\epsilon) = 0$, where $a_\epsilon(x,\eta) = k_\epsilon(x)/(1+\eta)^{\alpha_\epsilon(x)}$. $k_\epsilon(x) = \exp(\beta_\epsilon(x))$ is chosen such that $\beta_\epsilon(x)$ is a realization of a random field with the spherical variogram [5], correlation lengths $l_x = 0.2$, $l_y = 0.02$, and variance $\sigma = 1.5$. Here $\alpha_\epsilon(x)$ is chosen such that $\alpha_\epsilon(x) = k_\epsilon(x) + const$ with the spatial average of 2. In Figure 7 we compare the solutions (u_ϵ) and the fluxes $(-a_\epsilon(x,u_\epsilon)D_x u_\epsilon)$ corresponding to this equation with boundary and initial conditions given as previously. The solutions are rescaled for comparison purposes. The solid line designates the fine scale model results computed on a 120×120 grid, and the dotted line designates the coarse scale results computed using the numerical homogenization procedure on a 12×12 coarse grid. These results demonstrate the robustness of our approach for anistropic fields where h and ϵ are nearly the same. For different realizations of the random field we have observed similar results. Currently, we are studying the application of the oversampling technique to the numerical homogenization procedure.

Acknowledgments. Efendiev would like to thank T. Hou for inspiring discussions. We would also like to acknowledge the anonymous reviewers for their helpful comments.

REFERENCES

[1] T. ARBOGAST, *An overview of subgrid upscaling for elliptic problems in mixed form*, in Current Trends in Scientific Computing (Xi'an, 2002), Contemp. Math. 329, AMS, Providence, RI, 2003, pp. 21–32.

[2] I. BABUŠKA, G. CALOZ, AND J. E. OSBORN, *Special finite element methods for a class of second order elliptic problems with rough coefficients*, SIAM J. Numer. Anal., 31 (1994), pp. 945–981.

[3] F. BREZZI, *Interacting with the subgrid world*, in Numerical Analysis 1999 (Dundee), Chapman & Hall/CRC, Boca Raton, FL, 2000, pp. 69–82.

[4] G. Dal Maso and A. Defranceschi, *Correctors for the homogenization of monotone operators*, Differential Integral Equations, 3 (1990), pp. 1151–1166.
[5] C. V. Deutsch and A. G. Journel, *GSLIB: Geostatistical Software Library and User's Guide*, 2nd ed., Oxford University Press, New York, 1998.
[6] W. E and B. Engquist, *The heterogeneous multi-scale methods*, Comm. Math. Sci., 1 (2003).
[7] Y. Efendiev, T. Hou, and V. Ginting, *Multiscale finite element methods for nonlinear problems and their applications*, Comm. Math. Sci., submitted.
[8] Y. Efendiev and A. Pankov, *Meyers type estimates for approximate solutions of nonlinear elliptic equations and their applications*, submitted to Numer. Math.; also available online at http://www.math.tamu.edu/∼yalchin.efendiev/ep-meyers-elliptic.ps.
[9] A. Fannjiang and G. Papanicolaou, *Convection enhanced diffusion for periodic flows*, SIAM J. Appl. Math., 54 (1994), pp. 333–408.
[10] T. Y. Hou and X. H. Wu, *A multiscale finite element method for elliptic problems in composite materials and porous media*, J. Comput. Phys., 134 (1997), pp. 169–189.
[11] T. Y. Hou, X. H. Wu, and Z. Cai, *Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients*, Math. Comp., 68 (1999), pp. 913–943.
[12] T. Hughes, G. Feijoo, L. Mazzei, and J. Quincy, *The variational multiscale method—A paradigm for computational mechanics*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 3–24.
[13] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, New York, 1994.
[14] M. A. Krasnosel′skiĭ, P. P. Zabreĭko, E. I. Pustyl′nik, and P. E. Sobolevskiĭ, *Integral Operators in Spaces of Summable Functions*, Noordhoff International Publishing, Leiden, The Netherlands, 1976 (translated from the Russian by T. Ando).
[15] A.-M. Matache and C. Schwab, *Homogenization via p-FEM for problems with microstructure*, in Proceedings of the Fourth International Conference on Spectral and High Order Methods (ICOSAHOM 1998, Herzliya), 2000, Vol. 33, pp. 43–59.
[16] N. G. Meyers and A. Elcrat, *Some results on regularity for solutions of non-linear elliptic systems and quasi-regular functions*, Duke Math. J., 42 (1975), pp. 121–136.
[17] A. Pankov, *G-convergence and Homogenization of Nonlinear Partial Differential Operators*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.
[18] G. Papanicolaou and S. R. S. Varadhan, *Boundary value problems with rapid oscillating random coefficients*, Seria Colloquia Mathematica Societatis Janos Bolyai, 27 (1981), pp. 835–873.
[19] G. Sangalli, *Capturing small scales in elliptic problems using a residual-free bubbles finite element method*, Multiscale Model. Simul., 1 (2003), pp. 485–503.
[20] R. E. Showalter, *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*, Math. Surveys Monogr. 49, American Mathematical Society, Providence, RI, 1997.
[21] I. V. Skrypnik, *Methods for Analysis of Nonlinear Elliptic Boundary Value Problems*, Transl. Math. Monogr., American Mathematical Society, Providence, RI, 1994 (translated from the 1990 Russian original by Dan D. Pascali).
[22] A. Žaanen, *An Introduction to the Theory of Integration*, North–Holland, Amsterdam, 1961.

SIAM J. APPL. MATH.
Vol. 65, No. 1, pp. 69–92

ANALYSIS OF A CANARD MECHANISM BY WHICH EXCITATORY SYNAPTIC COUPLING CAN SYNCHRONIZE NEURONS AT LOW FIRING FREQUENCIES*

JONATHAN DROVER†, JONATHAN RUBIN†, JIANZHONG SU‡, AND BARD ERMENTROUT†

Abstract. A population of oscillatory Hodgkin–Huxley (HH) model neurons is shown numerically to exhibit a behavior in which the introduction of excitatory synaptic coupling synchronizes and dramatically slows firing. This effect contrasts with the standard theory that recurrent synaptic excitation promotes states of rapid, sustained activity, independent of intrinsic neuronal dynamics. The observed behavior is not due to simple depolarization block nor to standard elliptic bursting, although it is related to these phenomena. We analyze this effect using a reduced model for a single, self-coupled HH oscillator. The mechanism explained here involves an extreme form of delayed bifurcation in which the development of a vortex structure through interaction of fast and slow subsystems pins trajectories near a surface that consists of unstable equilibria of a certain reduced system, in a canard-like manner. Using this vortex structure, a new passage time calculation is used to approximate the interspike time interval. We also consider how changes in the synaptic opening rate can modulate oscillation frequency and can lead to a related scenario through which bursting may occur for the HH equations as the synaptic opening rate is reduced.

Key words. neuronal oscillations, Hodgkin–Huxley equations, synaptic excitation, slow passage, canard

AMS subject classifications. 34C15, 34C23, 34C25, 37G15, 37N25, 92C20

DOI. 10.1137/S0036139903431233

1. Introduction.

Recurrent excitatory networks of neurons are purported to underlie persistent activity in the nervous system. Such networks have been used as models for wave propagation and short-term memory [2, 17]. Long-lasting excitatory synaptic connectivity is generally sufficient to enable such densely coupled neurons to fire repetitively at high rates after some transient input, even when the individual neurons do not intrinsically oscillate. The ability of an excitatory network to maintain a persistent state depends on several interacting factors. In many types of cortical neuron models, excitatory coupling leads to asynchronous firing when the synaptic time course lasts long enough [10]. Shortening the time constant leads to two effects; first, the neurons can synchronize, and second, thus synchronized, the network cannot reignite due to the refractory period of the neurons. Studies of persistent activity have not generally focused on differences from this standard scenario that arise due to the intrinsic dynamics of individual neurons.

In this paper we report on a new mechanism through which persistent activity is drastically slowed by excitatory coupling in a network of Hodgkin–Huxley (HH) neurons. In fact, even if the neurons are intrinsically active (say, through current injection), the excitatory coupling dramatically slows them down. We will show that the mechanism for this slowing down is a consequence of an interesting mathematical

*Received by the editors July 9, 2003; accepted for publication (in revised form) March 18, 2004; published electronically September 24, 2004.
http://www.siam.org/journals/siap/65-1/43123.html
†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (jddst25@pitt.edu, rubin@math.pitt.edu, bard@math.pitt.edu). This work was partially supported by the National Science Foundation.
‡Department of Mathematics, University of Texas, Arlington, TX 76019 (su@uta.edu).

structure (a canard) in which a trajectory passes close to a curve of points that are critical points for the intrinsic neuronal dynamics without coupling and that switch from attracting to repelling with respect to these dynamics as synaptic excitation decays [3, 22]. While delayed bifurcation resulting from slow passage infinitesimally close to such a critical curve has been studied previously [3, 15, 16, 4], we shall see that the extreme slowing that we observe involves a novel "vortex" structure and does not fit into the standard class of slow passage problems that have been considered. Indeed, the dynamics controlling the slow passage here, namely, the synaptic decay, do not need to be particularly slow for the extreme delay in activity to occur. Moreover, the slowing phenomenon occurs over a broad parameter range, which distinguishes it from typical canard scenarios.

Our results relate to those of Guckenheimer et al. [8, 9], who found prolonged interspike intervals in a model of the lateral pyloric (LP) cell of the lobster stomatogastric ganglion (see Figure 5 in [8]) and analyzed a normal form of the subcritical Hopf-homoclinic bifurcation that gives rise to this phenomenon in the LP model. To compare our work to theirs, we note that the system we study has a unique, unstable critical point, at which the synaptic variable is zero. This critical point can be made to undergo a subcritical Hopf bifurcation as certain parameters are varied, although we do not do this. It is also quite possible that we are working in a parameter regime that is near a homoclinic bifurcation curve, although we do not consider this aspect of the dynamics directly. What Guckenheimer et al. analyze, however, is not a slow passage problem. Indeed, a crucial difference arising in the present work is that the decay of the synaptic variable sweeps a critical point of a reduced subsystem through a Hopf bifurcation, whereas their analysis treats periodic orbits with the full system held at a fixed distance from bifurcation. The slow passage that we consider leads to a delayed escape from a repelling branch of critical points for the subsystem; the normal form asymptotic analysis in [9] does not involve delayed bifurcation, multiple timescales, or reduced subsystems, although a slow variable does bring trajectories closer to the Hopf bifurcation on successive oscillation cycles in the LP model. Further, we give a directly computable estimate for the change in the synaptic variable during the passage through the vortex structure that traps it, which translates directly into an estimate of passage time, and we analyze the contribution of the synaptic decay rate to the delay. The work in [9] does give an estimate for oscillation period, but this is stated in terms of normal form variables and includes some abstract constants. We note that a prolonged silent phase in the HH equations was also observed in the thorough numerical study of Doi and Kumagai [5]. There, the slowing down was attributed simply to a decrease in the instability of the unstable equilibrium of a certain fast subsystem; no further analysis was given, and the vortex phenomenon was not uncovered.

In section 2 of this paper, we begin by demonstrating the extreme delay effect, first in a large network of HH neurons, then in a reduced model, and finally in a single self-coupled neuron. Since we show that the HH networks oscillate in near synchrony, the self-coupled neuron represents a reasonable approximation of the full network behavior. In the self-coupled neuron, we show how the slowed firing rate depends on the coupling strength, the time constant of the synapses, and the reversal potential of the synapses. In section 3, we review the phase plane for the reduced HH model for a single self-coupled neuron and illustrate the slowing mechanism there. In section 4, we introduce a polynomial approximation of the model that encapsulates the behavior of the reduced HH neuron in the silent phase. We analyze this model in some detail, first showing that the usual approach to delayed bifurcations [3, 15, 16] does

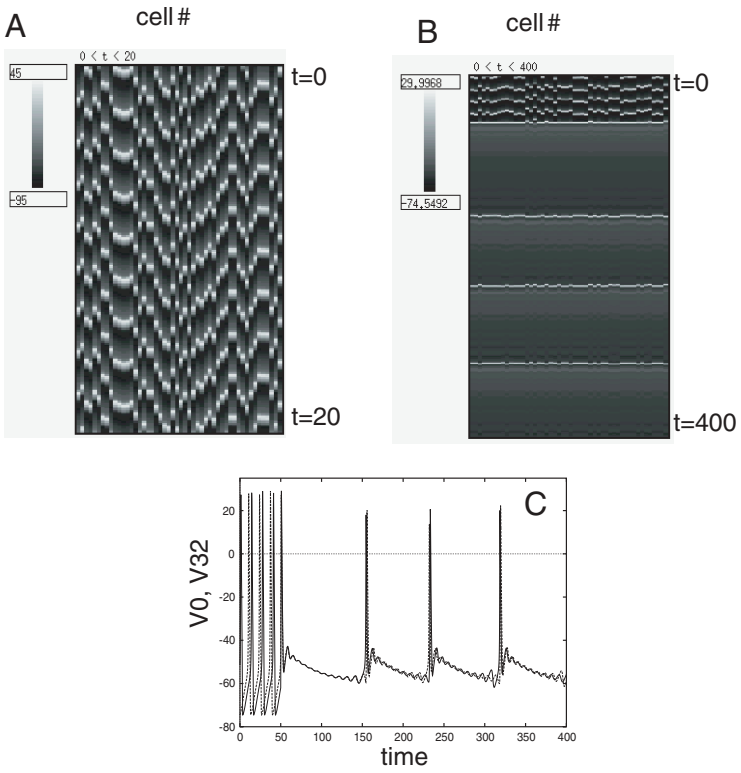


FIG. 1.1. *Behavior of networks of excitatorily coupled neurons depends on the intrinsic dynamics.* (a) *Persistent activity in a network of* 50 *cells with Traub's pyramidal cell dynamics. Neurons are indexed horizontally and time increases downward along the vertical axis. Grey scale depicts the membrane potential.* (b) *A similar network using the dynamics due to Hodgkin and Huxley. The first* 50 *milliseconds show the behavior of the uncoupled network; coupling is then turned on showing rapid synchronization and a* 10*-fold increase in the oscillation period.* (c) *Voltage traces from cells* 0 *and* 32 *(out of* 50*) from the simulation in* (b).

not capture the slowing down that we observe and then deriving a novel approach to analyze the delay, including its dependence on the synaptic decay rate. This approach focuses on the effect of a vortex structure in which the interaction of fast and slow subsystems pins trajectories in a certain neighborhood of the critical curve mentioned above. More specifically, we use this structure to derive an appropriate way-in–way-out function [3, 15, 16] that can be used to compute a good estimate of the change in the synaptic variable as a trajectory passes through the vortex. In section 5, we show how this vortex mechanism carries over to the HH system, and we explore the role of the active phase in the slow oscillations. In particular, we see how the slowing mechanism can contribute to a form of bursting, or alternation of sustained silent periods with periods of spiking, in the HH equations. Finally, in section 6, we give a further discussion of how this work relates to some earlier results and of the open questions that remain.

2. Numerical simulations of networks. If a network of excitatory cells is coupled together, often the network activity is asynchronous and has a much higher frequency than the individual cell [11, 12]. This is illustrated in Figure 1.1(a) for 50 cells coupled together in an all-to-all manner using a biophysical model for the fast

currents in a hippocampal neuron and synapses with a decay constant of 5 milliseconds [23]. Note that simulations shown in this figure, as well as all other simulations in this paper, were done using XPPAUT [7]. In the model simulated, individual cells do not fire on their own; the applied current is below threshold. However, coupled together, they produce a rhythm that is nearly 400 Hz. This is an example of strong persistent activity in an excitatory network. Contrast this behavior with another biophysical model based on the HH equations [13], with the same initial conditions and all-to-all coupling. The upper part of Figure 1.1(b) shows asynchronous output of the network when there is no coupling; the frequency is around 100 Hz. Here the neurons receive drive so that they fire spontaneously. After the first 50 milliseconds, the coupling is turned on and the network rapidly synchronizes and fires at a frequency of only about 10 Hz. Stronger coupling or longer decay rates lead to even lower frequencies. Both networks contain only three currents: a transient sodium current, a potassium current, and a leak. The individual voltage traces of two cells in Network B are shown in Figure 1.1(c). They are nearly synchronous, with out-of-phase subthreshold oscillations.

The difference in synchronization properties between these two example networks is fairly well understood, at least in the weak coupling limit. It is known that excitatory coupling can synchronize or desynchronize coupled neurons depending on many factors, such as the synaptic time constant. A very important factor is the nature of the individual neuron. In models for which the onset of repetitive firing is through a saddle node on a limit cycle (e.g., Figure 1.1(a)), excitatory coupling desynchronizes [6], while in models for which the onset is through a Hopf bifurcation (e.g., Figure 1.1(b)), excitatory coupling synchronizes [11]. As it turns out, the extreme slowing observed in the HH network also contributes to the synchronization through a form of fast threshold modulation [20]. We will return to this point in the discussion.

Our goal in much of the rest of this paper is to understand how the frequency of the synchronized oscillations is reduced to the extremely low rates observed in the HH simulations. To understand this, we first reduce the four-variable model to a two-variable system in the manner of Rinzel [18]. This will make the analysis simpler in the subsequent sections. The same network of 50 cells for the reduced system exhibits the same behavior as the full model (not shown); however, the cells synchronize perfectly, unlike in the four-variable cell model. Since synchrony (or near synchrony) appears to be a stable state of the network, we can understand the slowing down of the full network by studying a single self-coupled reduced HH cell:

$$
\begin{aligned}
C\frac{dV}{dt} &= -g_L(V-V_L) - g_K n^4(V-V_K) - g_{Na} m^3 h(V-V_{Na}) \\
&\quad + I_0 - g_{syn} s(V-V_{syn}), \\
\frac{dh}{dt} &= \frac{h_\infty(V)-h}{\tau_h(V)}, \\
m &= m_\infty(V), \\
n &= \max(.801 - 1.03h, 0), \\
\frac{ds}{dt} &= \alpha(V)(1-s) - s/\tau_{syn}.
\end{aligned} \tag{2.1}
$$

The specific values of the gating functions and parameters in (2.1) are given in Appendix A. Note that the synapse has dynamics gated by the potential, V, and the reversal potential of the synapse is V_{syn}. Figure 2.1(a) shows the period of the self-coupled cell as a function of the strength of coupling, g_{syn}, for several different synaptic

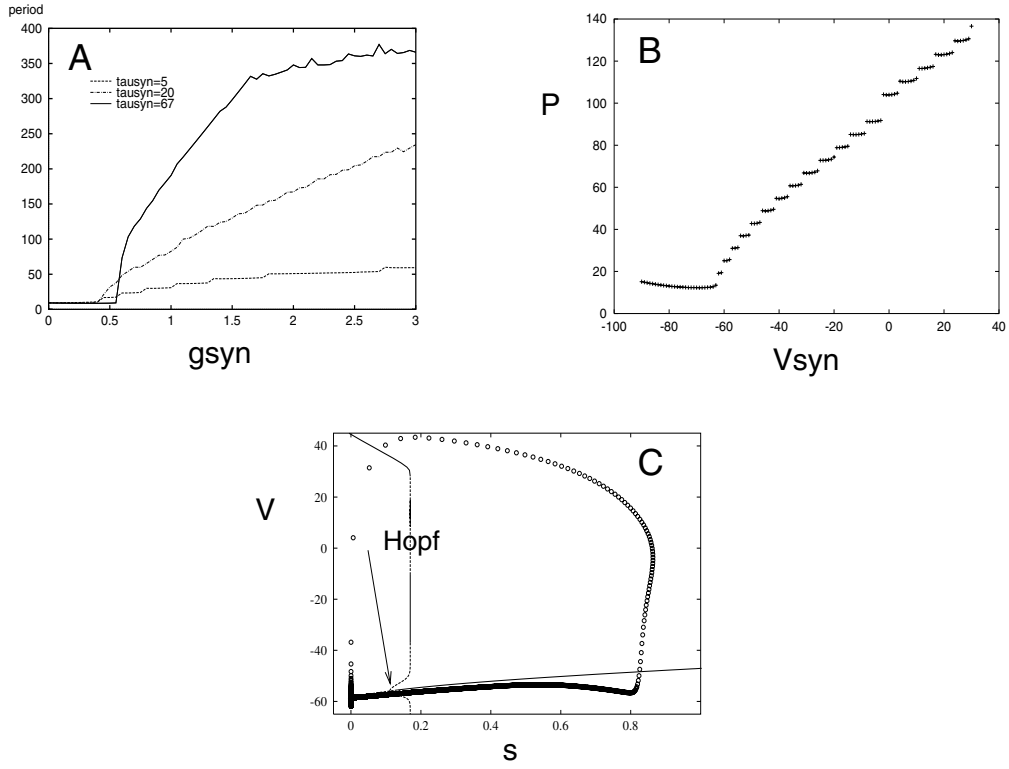


FIG. 2.1. *Properties of the self-coupled reduced HH model.* (a) *The variation of the period as a function of the maximal synaptic conductance for different synaptic decay times.* (b) *Dependence of the period on the reversal potential of the synapse;* $g_{syn} = 4\frac{ms}{cm^2}$ *and* $\tau_{syn} = 10s$. *The resting potential of the neuron is about* $-65\ mV$. *The discontinuities in the curve occur because the trajectory cannot release until after an integral number of subthreshold oscillations (see Figure* 1.1(c)*).* (c) $V-s$ *phase plane during a slow oscillation (trajectory shown with circles and thick solid line) superimposed on the bifurcation diagram (thin solid and dashed lines) for which* s *is treated as a parameter. The arrow depicts the value of* s *at which there is a Hopf bifurcation. To compute the bifurcation diagram, we replaced the piecewise linear definition of* n *in* (2.1) *with a smooth approximation.*

decay rates, τ_{syn}. This dramatic slowing down is not due to simple depolarization; the period is a monotonically decreasing function of the applied current, I_0. Furthermore, for g_{syn} fixed and s held constant as a parameter, the period is roughly constant as s increases. The mechanism for slowing down depends on the transient nature of $s(t)$ and its interplay with the intrinsic dynamics of the reduced HH model. Furthermore, synaptic *excitation* is required for this; Figure 2.1(b) shows the period as a function of the reversal potential of the synapse V_{syn}.

We can give a rather crude explanation for the behavior by treating the synapse as a slow variable. Thus, in (2.1), we treat s as a parameter in the voltage dynamics. For sufficiently large values of s and for g_{syn} large, the membrane dynamics have a stable fixed point corresponding to depolarization block of the sodium current. (The resting potential is so large that the sodium channels are inactivated by the synapse.) As s is decreased, there is a Hopf bifurcation leading to large amplitude periodic solutions. Figure 2.1(c) shows the $V - s$ phase plane with the bifurcation diagram superimposed. The trajectory winds around in a clockwise motion. Essentially, the

slow oscillation is a one-spike elliptic burster [19, 24, 14]. That is, for large values of s, the resting state is stable and the neuron cannot fire. Thus, the synaptic gating variable decays. As this variable gets smaller, the trajectory passes through the Hopf bifurcation (shown by the arrow) and the resting state becomes unstable. However, as can be seen in the figure and is known to occur in elliptic bursting, the trajectory continues along the curve of unstable fixed points, to s-values well below the Hopf point, before jumping away.

While this explanation seems somewhat satisfactory, it cannot account for the drastic slowing down and extreme decay (to nearly 0) of s that we observe. Moreover, the time constant of the decay in the figure ($\tau_{syn} = 10$ msec) is not particularly slow; in this range it is about twice the decay rate of the inactivation variable, h. The mechanism for the extended period is actually quite subtle, and it turns out to be better to treat the recovery variable, h, as the slow variable and to study the dynamics in the $V-h$ plane. Moreover, we shall see that standard treatment of elliptic bursting and associated delay does not predict the extent to which the period increases with τ_{syn} here, as seen in Figure 2.1(a).

3. The $V-h$ plane. We rewrite the equations for the reduced HH model:

$$C\frac{dV}{dt} = f(V,h) - g_{syn}s(V - V_{syn}), \tag{3.1}$$

$$\frac{dh}{dt} = \alpha_h(V)(1-h) - \beta_h(V)h, \tag{3.2}$$

where

$$f(V,h) = I_0 - g_{Na}h(V - V_{Na})m_\infty^3(V) - g_K(V - V_K)n^4(h) - g_L(V - V_L).$$

The equation for the synapse is

$$\frac{ds}{dt} = \alpha(V)(1-s) - s/\tau_{syn}. \tag{3.3}$$

While h and s have similar time courses, h evolves much more slowly than V, so we refer to (3.1) as the fast equation and (3.2) as the slow equation, and we refer to this pair of equations as (PS), for projected system. For each fixed value of s, the solution to the equation $dV/dt = 0$ forms a triple-branched curve in (V,h)-phase space, which constitutes the fast nullcline (Figure 3.1). We will also refer to the slow nullcline, given by $dh/dt = 0$ (Figure 3.1(b)). Note that as s evolves, the fast nullcline of system (3.1)–(3.2) evolves correspondingly, while the slow nullcline is independent of s. Alternatively, for the full system (3.1)–(3.3), there exist two-dimensional fast and slow nullsurfaces in (V,h,s)-phase space.

Solutions to the system (3.1)–(3.3) are strongly attracted to the left and right branches of the fast nullsurface, except during fast jumps between branches (see Figure 3.1(a)). We refer to a time period when a solution is near the left (right) branch as a silent phase (active phase). For our analysis, we will make use of projections of solutions to (V,h)-phase space, but it is important to note that s continues to evolve along with V and h.

3.1. Attraction to the intersection of nullclines and extended delay. The left panel of Figure 3.1 shows a numerically generated trajectory of (3.1)–(3.3),

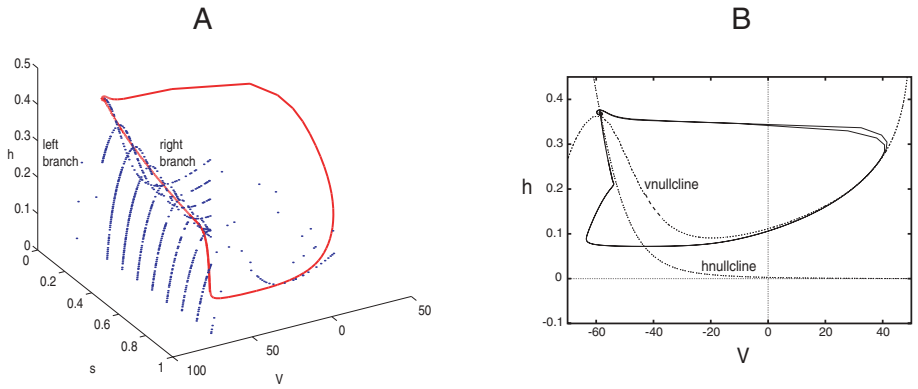


FIG. 3.1. *An orbit of* (3.1)–(3.3) *together with relevant nullclines. In the left panel, it is apparent that the orbit spends a long time in the silent phase near the left knees of the V-nullclines. In the right panel, it is clear that the orbit hugs the h-nullcline until s decays very near to zero, and then there is a small oscillation followed by a jump up to the active phase. $\tau_{syn} = 20s$ in this figure.*

superimposed on V-nullclines of (PS) that were numerically generated for several different values of s. A projection of this trajectory into (V, h)-phase space appears in the right panel, along with the V- and h-nullclines for an arbitrary fixed s near 0. In Figure 3.1, we see that after jumping down to the left surface of the fast nullcline, the orbit travels very close to this surface, although this is not apparent in the right panel of Figure 3.1 because we have only plotted the fast nullcline for a single, very small value of s. The orbit also appears to hug the slow nullcline as the synaptic variable s slowly decays; in other words, the orbit is very close to the intersection of the fast and slow nullclines for each fixed s. After a long delay, the orbit spirals away from the intersection of the nullclines as if this intersection point, treated as a critical point of (PS), had suddenly become unstable through a Hopf bifurcation at some small s. This is not the case; although there is a Hopf bifurcation and a loss of stability as s decays, the orbit remains near the nullcline until s reaches values well below the bifurcation point.

The intersection of the nullclines may be viewed as a critical point of (PS) with s fixed as a parameter. The stability of the critical point changes when $s \approx 0.222$ for the default parameter set, while the escape seen in Figure 3.1 occurs when $s \approx 0.003$. This means that the orbit is attracted toward the intersection (or not repelled) while that intersection represents an unstable fixed point of (PS). The objectives for this and the following section are to explain why this delayed exit occurs and to derive an analytical expression that gives a good estimate of the duration of this delay.

3.2. Ingredients for the delay. The problem presented here is that orbits appear to be attracted to a curve of unstable critical points. However, each critical point is only unstable for fixed s. For the full system (3.1)–(3.3), s decays during the silent phase, and so there are no true critical points with $s > 0$. Thus, we cannot immediately assume that the intersection will repel the orbit once it is unstable with respect to (PS). Linear stability analysis for critical points of (PS) may not be appropriate for the system (3.1)–(3.3). Somehow, one needs to take into account the dynamics of s to explain the delay in escape from the silent phase. Previous authors have contended with this issue in slow passage problems [3, 15, 16, 1, 4] and in elliptic bursting in particular [19, 24, 14, 21]. Unless $1/\tau_{syn} \ll dh/dt$, however, (3.1)–(3.3)

do not fit the standard slow passage assumptions.

Also, the h-coordinate of the fast nullcline increases as s decreases, and the slow nullcline has negative slope with respect to the variable V (in the (V, h)-plane). Thus, the intersection of the nullclines is moving up and to the left in the phase plane as s decreases. Trajectories also move in this direction, as they approach the negatively sloped slow nullcline. Thus, trajectories may approach the intersection of the nullclines, even if the linearization about the intersection of the nullclines with fixed s yields eigenvalues with positive real parts. Below, we will discuss an additional trapping mechanism that holds trajectories near this intersection.

Finally, for a value of s near the Hopf bifurcation, the nullclines are in the fold canard configuration [3]. Although this lasts for only a short period, it may provide a mechanism for a canard to arise in the full system. In this paper we will not use a singular slow-fast decomposition, and we will not use the tools of nonstandard analysis [3]. Nevertheless, the canard configuration appears to be an imperative structural feature in any system that demonstrates this extended delay, for reasons that we shall see below.

4. A simple system. To do any analysis directly, a model simpler than (3.1)–(3.3) is useful to characterize the relevant dynamics in the silent phase, although the conclusions of the analysis are expected to hold for more general systems. For the sake of analysis, the system ideally will have nullclines that are represented by polynomials. Based on the observations from the previous subsection, our model must incorporate the following characteristics:

- The slow nullcline has a negative slope with respect to the fast variable, provided the trajectory approaches the slow nullcline from the left after it enters the silent phase (see Figure 3.1). If the approach is from the right, then the slope of the curve must be positive.
- The intersection of the fast and slow nullclines is a stable critical point (when parameterized by s) of the intrinsic equations for large values of s, and then changes stability via a Hopf bifurcation induced by a transversal crossing of a conjugate pair of eigenvalues through the imaginary axis, away from the origin, as s decays. For a value of s near the Hopf bifurcation, the nullclines must be in the regular fold canard configuration, discussed in [3].
- The vector field of the system is analytic [15, 16] and autonomous during the silent phase.

4.1. The model. The model used for all analysis during the silent phase is

$$\frac{dx}{dt} = -f(x) + y - I(s)x, \tag{4.1}$$

$$\frac{dy}{dt} = -\epsilon\left(y + \frac{1}{4}x^5\right), \tag{4.2}$$

$$\frac{ds}{dt} = -\frac{s}{\tau_{syn}}, \tag{4.3}$$

where $0 < \epsilon \ll 1$; note that we consider only $x < 0$. For simulations in this paper, the function f in (4.1) is

$$f(x) = \frac{1}{4}x^3 - 2x$$

and the synaptic current function I is

$$I(s) = \frac{3}{2}s.$$

Note that this model does not oscillate, but trajectories do jump up from the silent phase. This is sufficient for consideration of behavior during the silent phase. It is not necessary to consider the active phase (when spikes occur) to explain the slow release; however, we will return to the study of the role of the active phase for the HH equations, and bursting in particular, later in the paper.

4.2. Some notation. For the remainder of the paper, the following notation will be used. $N_f(x,s)$ is the y-coordinate of the fast nullcline ($\frac{dx}{dt} = 0$) for a given x and s. Similarly, $N_s(x)$ is the y-coordinate of the slow nullcline ($\frac{dy}{dt} = 0$) for a given value of x. Note that $\partial N_f/\partial s < 0$ for $x < 0$, that $N_s(x)$ does not depend on s, and that these two curves intersect for each fixed s. Let $(\tilde{x}(s), \tilde{y}(s))$ denote the curve of intersection points.

For the system given in (4.1), (4.2), the functions $N_f(x,s)$ and $N_s(x)$ are given by

$$N_f(x,s) = f(x) + I(s)x,$$

$$N_s(x) = -\frac{1}{4}x^5.$$

The intersection of these curves is easily found for each value of s.

4.3. The usual approach. Though the trajectory is visibly separated from the intersection of the fast and slow nullclines in the right panel of Figure 3.1, it is still possible that the release value of s can be approximated using the variational equation around $(\tilde{x}(s), \tilde{y}(s))$. Indeed, this approach has been taken previously to analyze delayed escape in slow passage through a Hopf bifurcation through use of a way-in–way-out function [3, 15, 16]. This function relates the attraction of the orbit before the Hopf bifurcation to the repelling of the orbit after the change of stability has taken place. We shall see that in our case, this approach is not necessarily appropriate.

We now demonstrate the poor performance of the standard way-in–way-out, computed using the equation of first variation along the curve $(\tilde{x}(s), \tilde{y}(s))$. Let J be the Jacobian matrix of the system defined by (4.1)–(4.2) along $(\tilde{x}(s), \tilde{y}(s))$. We have that

$$J(s) = \begin{pmatrix} -\frac{3}{4}\tilde{x}^2(s) + 2 - I(s), & 1 \\ -\epsilon\frac{5}{4}\tilde{x}^4(s), & -\epsilon \end{pmatrix}. \tag{4.4}$$

The equation of first variation is

$$\frac{d}{ds}\begin{pmatrix} x \\ y \end{pmatrix} = -\frac{\tau_{syn}}{s} J(s) \begin{pmatrix} x \\ y \end{pmatrix}. \tag{4.5}$$

The solution to (4.5), taken from a starting point (x_0, y_0, s_{enter}), is

$$\begin{pmatrix} x \\ y \end{pmatrix} = \exp\left(\int_{s_{enter}}^{s} -\frac{\tau_{syn}}{\omega} J(\omega) d\omega \right) \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}. \tag{4.6}$$

Given an s_{enter}, we may solve the equation

$$\left\| \exp\left(\int_{s_{enter}}^{s} -\frac{\tau_{syn}}{\omega} J(\omega) d\omega \right) \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \right\|_2 \tag{4.7}$$

for $s = s_{exit}$. The value s_{exit} is an approximation of the value of s such that

$$\left\| \begin{pmatrix} x \\ y \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \right\|_2,$$

where x, y are functions of s since they solve (4.5).

In typical slow passage problems [3, 15, 16, 4], this s_{exit} provides a good approximation for the release value of s. The results for the system under consideration here are not good, especially for the lower values of τ_{syn} tested. This poor performance does not contradict the standard theory; this approach breaks down precisely when the passage rate determined by the decay of s in (4.3) is not sufficiently slow in comparison with the rate of change in (4.2). The value of the approximated value of s_{exit} over a range of τ_{syn} is shown in Figure 4.1. The standard way-in–way-out analysis overestimates s_{exit}. Since s decays in the silent phase, this means that this approach underestimates the amount of time spent in the silent phase.

Notice further that the s_{exit} curve generated here is rather flat. This is expected because the linearization of the system when s is used as a parameter does not depend on τ_{syn}. The slight curvature of the s_{exit} curve that is visible in Figure 4.1 is due to the fact that different values of s_{enter} satisfy the entrance criterion (see caption) for different τ_{syn}. Simulations (solid line in Figure 4.1) suggest that the true value of s_{exit} varies as the logarithm of τ_{syn}. Correspondingly, the passage time from s_{enter} to s_{exit} grows linearly with τ_{syn}, and spike frequency decreases as $1/\tau_{syn}$ as τ_{syn} increases.

It is now apparent this is not a standard way-in–way-out problem about the curve of critical points of a slow-fast system. In the following sections, we will propose a mechanism for the increased delay, perform the corresponding analysis, and demonstrate that this approach gives a much better estimate of the observed delay than that given by the usual analysis done in this section, up to values of τ_{syn} for which $1/\tau_{syn} \ll \epsilon$. For values of τ_{syn} greater than this, the usual approach is sufficient, and as $\tau_{syn} \to \infty$ the two approaches are identical.

4.4. The trapping mechanism. As $s \to 0$, the fast nullcline moves upward in the y-coordinate, since $x < 0$ and thus $\partial N_f / \partial s < 0$. In simulations, it appears as if orbits of (4.1)–(4.3) (or of (3.1)–(3.3)) track very close to the intersection curve of the fast and slow nullclines. To understand what organizes the flow near this curve, it is useful to define the following set:

$$A(s) = \left\{ (x_0, y_0) \middle| \frac{dy}{dt}(x_0, y_0) < \frac{dN_f}{ds}(x_0, s) \frac{ds}{dt} \right\}. \tag{4.8}$$

This set consists simply of the points in the (x, y)-plane such that a trajectory that passes through the point $(x_0, y_0) \in A$ travels more slowly in the vertical direction (y-direction) than does the point on the fast nullcline with the same x-coordinate. Because $N_f(x, s)$ increases as s decreases for fixed $x < 0$, we have that $\frac{dN_f}{ds}(x_0, s)\frac{ds}{dt} > 0$, which guarantees that $A(s)$ is nonempty for each s. As $x \to -\infty$, $\frac{dy}{dt} \to \infty$ as well (see (4.2)), so for each fixed y, there exists x sufficiently negative such that $\frac{dy}{dt} > \frac{dN_f}{ds}\frac{ds}{dt}$; similarly, for each fixed $x < 0$, there exists y sufficiently negative such

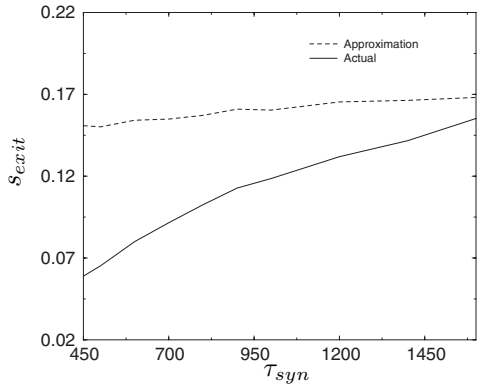


FIG. 4.1. *Values of s_{exit} computed numerically versus those computed from the usual way-in–way-out function, as τ_{syn} varies. The approximation obtained by solving* (4.7) *(dotted line) appears to be fairly invariant with respect to τ_{syn}, but simulations of* (4.1)–(4.3) *strongly suggest that this is not the case (solid line). Here, $\epsilon = .01$ and the entrance criterion used in* (4.7) *is $||x||_2 = 0.1$.*

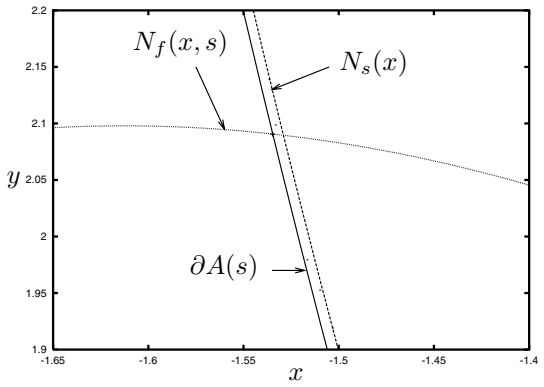


FIG. 4.2. *The curves N_f and N_s along with the boundary of the set $A(s)$ for $s = .0326$. The set $A(s)$ also includes a region to the right of $N_s(x)$, but only the shaded region is relevant.*

that this inequality holds. Thus, $A(s)$ is bounded to the left and below, and the boundary $\partial A(s)$ is a curve, which we denote $y_{\partial A(s)}(x)$, in the (x, y) plane. For the simple system (4.1)–(4.2), we can express the boundary curve $\partial A(s)$ as the graph of a function:

$$y_{\partial A(s)}(x) = -\frac{1}{4}x^5 + \frac{3xs}{2\epsilon\tau_{syn}}. \tag{4.9}$$

Notice that $y_{\partial A(0)}(x) = N_s(x)$, and that as $\tau_{syn} \to \infty$, $y_{\partial A(s)}(x) \to N_s(x)$.

Figure 4.2 shows the curve $\partial A(s)$ for $s = .0326$, along with $N_f(x, s)$ and $N_s(x)$. For the value of s in Figure 4.2, if the trajectory lies to the right of the curve $\partial A(s)$, then $N_f(x, s)$ will be moving upward faster than the trajectory. Likewise, if the trajectory lies to the left of the curve, then the nullcline will be moving upward slower than the trajectory.

The intersection of the curves $\partial A(s)$ and $N_f(x, s)$ turns out to be extremely important for the delay phenomenon under study. The curve defined by these intersection points for a range of s values forms an attractor for values of s for which, from the perspective of the analysis done in section 4.3, the intersection of N_f and N_s

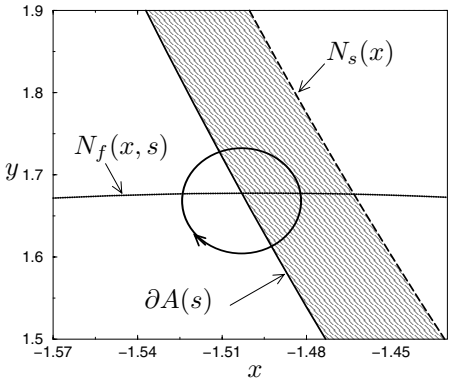


FIG. 4.3. *A sample trajectory as viewed by an observer riding the intersection of $\partial A(s)$ and $N_f(x,s)$. Trajectories to the left of $\partial A(s)$ pass to y-values above the observer, trajectories to the right fall behind. The left and right movement is dependent on whether the trajectory is above or below the curve $N_f(x,s)$.*

corresponds to a repelling set. Suppose that a trajectory lies below $N_f(x,s)$ and to the right of $\partial A(s)$. Thus, the trajectory and $N_f(x,s)$ are separating, but $\frac{dx}{dt} < 0$, and so eventually the trajectory crosses $\partial A(s)$ and then begins to catch up to $N_f(x,s)$. This may result in a net contraction toward $\partial A(s) \cap N_f(x,s)$. The y-coordinate of the trajectory will eventually increase through $N_f(x,s)$, such that $\frac{dx}{dt} > 0$ results. This causes the trajectory to again cross the curve $\partial A(s)$, and another contraction toward $\partial A(s) \cap N_f(x,s)$ may occur as $N_f(x,s)$ catches up to the trajectory. Thus, the intersection curve of $\partial A(s)$ and $N_f(x,s)$, while not itself invariant under the flow, creates a moving vortex, or core about which the flow spirals. The flow diagram around this core, projected to the (V,h)-phase plane, is shown in Figure 4.3.

This moving vortex structure generates a trapping mechanism within the flow. Simulations show that trajectories follow the vortex curve very closely during the silent phase. Using a change of variables, we next explore the stability of the vortex curve and its impact on delayed escape from the silent phase.

4.5. Equations of the moving vortex. To focus on the moving vortex, we will shift the system so that the intersection, say, $(\hat{x}(s), \hat{y}(s))$, of $\partial A(s)$ and $N_f(x,s)$ occurs at the origin for all s. For the simplified model, note that one can obtain explicit expressions for this intersection point. A linear change of variables, $z_1 = x - \hat{x}(s)$ and $z_2 = y - \hat{y}(s)$, yields the following system:

$$\frac{dz_1}{dt} = \frac{dx}{dt} - \frac{d\hat{x}}{ds}\frac{ds}{dt}, \tag{4.10}$$

$$\frac{dz_2}{dt} = \frac{dy}{dt} - \frac{d\hat{y}}{ds}\frac{ds}{dt}, \tag{4.11}$$

which can also be written

$$\frac{dz_1}{dt} = f_1(z_1, z_2, s), \tag{4.12}$$

$$\frac{dz_2}{dt} = f_2(z_1, z_2, s), \tag{4.13}$$

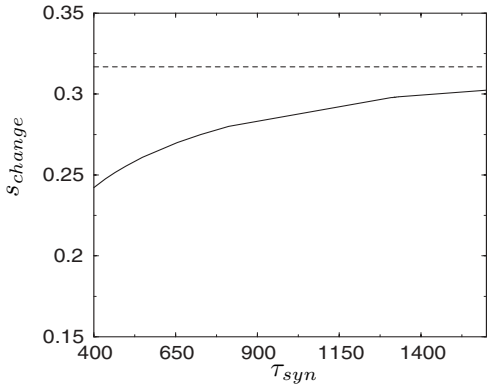


FIG. 4.4. *Change of stability. The solid line represents the value of s where the sign of the real part of the complex conjugate pair of eigenvalues changes along the curve $(\hat{x}(s), \hat{y}(s))$. The dotted line shows the value of s when the curve of critical points for* (4.1)–(4.2) *changes stability. This value is not dependent on τ_{syn}.*

where s is governed by (4.3).

If s is fixed as a parameter, then we may compute the linearization of system (4.12)–(4.13) about the vortex point $(z_1, z_2) = (0, 0)$. Although $(0, 0)$ is not a critical point for system (4.12)–(4.13), the sign of the real part of the complex conjugate pair of eigenvalues of the linearized system will still yield information about to what extent the neighborhood around the point acts as an attractor, as discussed above. Also, because the parameter τ_{syn} was incorporated into the linear component of the system during the change of variables, the value of s where the eigenvalues' real part changes sign is not invariant with respect to τ_{syn}, as it is using the regular approach discussed in section 4.3. The value of s where the eigenvalues' real part changes sign is shown in Figure 4.4. This is encouraging because it demonstrates a lower value for the change of stability in addition to a dependence on τ_{syn}, both of which are apparent in simulations but lacking in the analysis in section 4.3.

4.6. Release value for *s*. Because the real part of the eigenvalues crosses through zero for a smaller value of s in the linearization of system (4.12)–(4.13) about $(0, 0)$ than observed in the linearization of (4.1)–(4.2), we expect that the linearization of system (4.12)–(4.13) will provide an improved estimate of the exit value for s, relative to the analysis in section 4.3, at least until τ_{syn} becomes extremely large. In addition to the geometric argument given in section 4.4, an analytical justification for this expectation is given in Appendix B.

Now that we have transformed to the frame of the moving vortex, the analysis itself proceeds as in section 4.3. We rewrite (4.12)–(4.13) in vector form as

$$\frac{d\vec{z}}{ds} = -\frac{\tau_{syn}}{s}\vec{f}(\vec{z}, s). \tag{4.14}$$

The equation of first variation on the vortex curve $(z_1, z_2) = (0, 0)$ is

$$\frac{d\vec{z}}{ds} = -\frac{\tau_{syn}}{s}\vec{f}_{\vec{z}}(0, 0, s)\vec{z}. \tag{4.15}$$

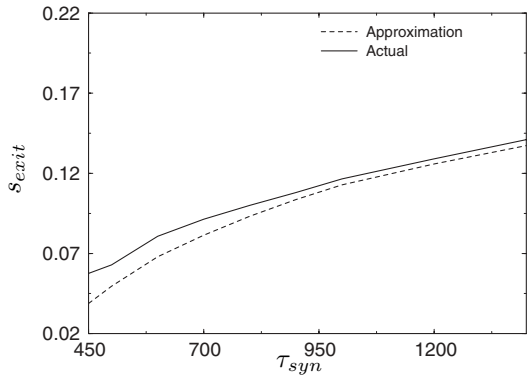


FIG. 4.5. *Improved estimate of s_{exit}. As a function of τ_{syn}, the exit value s_{exit} is derived from solution of* (4.17) *(dashed line) and numerical solution of the full translated model* (4.12)–(4.13) *(solid line). The entrance criterion for this figure was $||z||_2 = 0.03$, and again $\epsilon = 0.01$.*

The solution to (4.15) is given by

$$\vec{z}(s) = \exp\left(-\tau_{syn}\int_{s_0}^{s}\frac{1}{w}\vec{f}_{\vec{z}}(0,0,w)dw\right)\vec{z}(s_0). \tag{4.16}$$

To approximate the value of s where release begins to occur, we choose a value s_{enter} satisfying an entrance criterion, $||z||_2 = \eta$. We solve the equation

$$||\vec{z}(s)||_2 = \left\|\exp\left(-\tau_{syn}\int_{s_{enter}}^{s}\frac{1}{w}\vec{f}_{\vec{z}}(0,0,w)dw\right)\vec{z}(s_{enter})\right\|_2 = \eta. \tag{4.17}$$

The results of this estimation for a range of τ_{syn} are shown, along with results from full numerical simulations, in Figure 4.5. The approximation is much better than the one obtained in section 4.3 for low to moderately high values of τ_{syn}.

Remark 4.1. In principle, there exists some curve, say, $(x_{opt}(s), y_{opt}(s))$, such that linearization about this curve yields an optimal estimate of s_{exit}. Numerical simulation suggests that system (4.12)–(4.13) has a fixed point for each s, and this is the natural candidate about which to linearize this translated system. (In terms of Appendix B, linearization about this curve would yield a truly linear system in (8.8).) However, it is not clear how to access this curve numerically, and the geometric arguments and numerical computations done here, along with the analytical calculation in Appendix B, show that the moving vortex curve is a good approximation to $(x_{opt}(s), y_{opt}(s))$ to use for estimation of s_{exit}.

Remark 4.2. Unfortunately, for very large values of τ_{syn}, the approximation loses accuracy and gives a similar, but slightly less accurate, performance to the standard approach. Recall that the moving vortex point is defined as the intersection of $\partial A(s)$ with the fast nullcline $N_f(x,s)$ for each s. The boundary $\partial A(s)$ is given by $\frac{dy}{dt} = \frac{\partial N_f}{\partial s}\frac{ds}{dt} = -\frac{\partial N_f}{\partial s}\frac{s}{\tau_{syn}}$. As τ_{syn} increases, $\partial A(s)$ therefore approaches the slow nullcline, and correspondingly the moving vortex point approaches the intersection of the fast and slow nullclines, which is exactly the moving critical point used in the standard analysis. This explains why the moving vortex analysis is similar to the standard analysis for sufficiently large τ_{syn}. However, the transformation (4.10)–(4.11) brings τ_{syn} into (4.12)–(4.13), so the two approaches remain nonidentical.

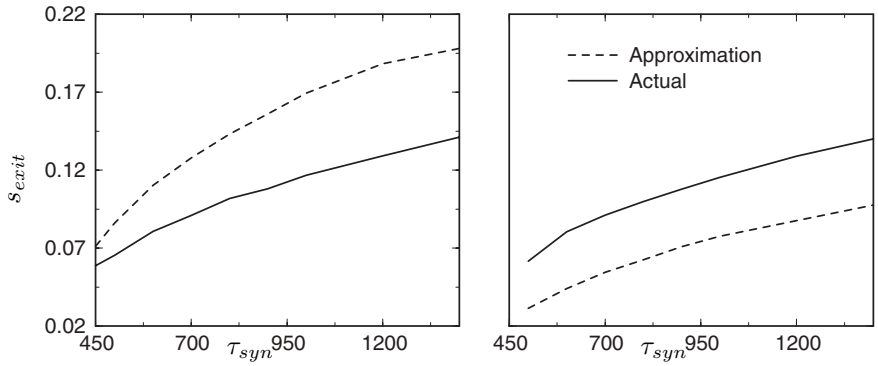


FIG. 4.6. *The approximation curve and the actual curve using the value* $\eta = 0.025$ *(left panel) or* $\eta = 0.035$ *(right panel) as the entrance criterion. The results are not as good as those in Figure* 4.5.

Remark 4.3. It is important to note that the results of our approach do depend on the value of η chosen for the entrance criterion. Because we take the equation of first variation of (4.12)–(4.13) about the vortex curve $(z_1, z_2) = (0, 0)$, rather than about the translated version of the optimal curve (x_{opt}, y_{opt}) discussed in Remark 4.1, we cannot choose η arbitrarily small. The behavior in a very small neighborhood of the origin, and the time to exit this neighborhood, do not perfectly capture the behavior near the optimal curve. Also, η cannot be chosen too large. Large η will result in failure of the approximation provided by the equation of first variation, and nonlinear terms may dominate. There must be an ideal entrance value, in the sense that the results obtained provide the most accurate approximations. Figure 4.6 shows the results derived from less appropriate values of η than that used in Figure 4.5. Note, however, that these results are still better than the standard approach (Figure 4.1) over the lower range of τ_{syn} values considered.

5. The HH equations.

5.1. Mechanism for slow oscillations. In section 4, a simplified model was used to elucidate a mechanism, involving trapping of trajectories near a vortex curve, by which slow synaptic decay results in an oscillation with a very long period. Because our simplified model satisfies the conditions listed at the start of section 4, this model is an appropriate subject for analysis, and we expect that the argument and findings from sections 4.4–4.6 carry over directly to the reduced HH model (3.1)–(3.3).

Indeed, numerical study strongly suggests that the mechanism for slow oscillations in the HH equations is identical to that of the simple model. Again, there is a vortex curve which is stable longer (for smaller s) than is the fixed-point curve created by the intersection of the fast and slow nullsurfaces. Figure 5.1 shows the analogue to Figure 4.2 for the reduced HH equations.

5.2. The active phase. Up to this point, our analysis has concerned only what occurs during the silent phase of oscillations. By changing the recovery capability of the synapse, either we can make the slow behavior discussed above more pronounced or we can eliminate the silent phase completely. The latter results in high-frequency oscillations, and for appropriate values of τ_{syn} this can induce bursting. Before discussing bursting, however, we take a closer look at how the recovery of the synapse depends on parameters in the model, assuming that a prolonged silent phase has occurred.

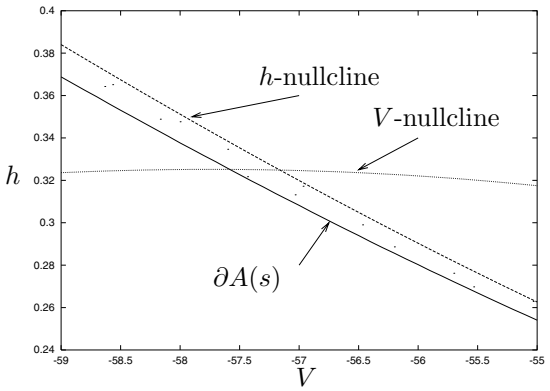


FIG. 5.1. *The set $A(s)$ for the HH equations (3.1)–(3.3) for fixed s. The shaded region is the numerically computed set of (V, h) (to the left of the slow h-nullcline) where the trajectory is moving more slowly in the direction of increasing h than is the fast V-nullcline.*

Under the flow of the reduced HH system (3.1)–(3.3), the synapse recovers (s increases) during the active phase, which begins when the cell jumps up from the vicinity of a left knee of the fast V-nullsurface and terminates when the cell jumps down from a right knee of this nullsurface. If we let $F(V, h, s)$ denote $f(V, h) - g_{syn}s(V - V_{syn})$, then the knees are the two solutions of $F(V, h, s) = \partial F(V, h, s)/\partial V = 0$, parametrized by s. More precisely, we can solve $F(V, h, s) = 0$ for $V = V(h, s)$, and then solve $\partial F(V(h, s), h, s)/\partial V = 0$ for $h = h(s)$, such that $V = V(h(s), s)$.

We can implicitly differentiate the equation

$$f(V(h(s), s), h(s)) - g_{syn}s(V(h(s), s) - V_{syn}) = 0$$

with respect to s to obtain

$$\frac{\partial f}{\partial V}\left[\frac{\partial V}{\partial h}\frac{dh}{ds} + \frac{\partial V}{\partial s}\right] + \frac{\partial f}{\partial h}\frac{dh}{ds} - g_{syn}s\left(\frac{\partial V}{\partial h}\frac{dh}{ds} + \frac{\partial V}{\partial s}\right) - g_{syn}(V(h(s), s) - V_{syn}) = 0. \tag{5.1}$$

Substitution of $\partial F(V(h(s), s), h(s), s)/\partial V = 0$ into (5.1) yields $\frac{\partial f}{\partial h}\frac{dh}{ds} = g_{syn} \times (V(h(s), s) - V_{syn})$. Rewriting this as a formula for dh/ds and substituting the currents in f from Appendix A, as well as $V_{syn} = 0$, yields

$$\frac{dh}{ds} = \frac{g_{syn}V}{-g_{Na}m^3(V)(V - V_{na}) - 4g_K n^3(h)(V - V_k)\frac{dn}{dh}}, \tag{5.2}$$

where $V = V(h(s), s)$ and $h = h(s)$. If we insert parameter values from Appendix A, as well as the range of V values found in the silent phase (say, $h = h_L(s)$) or the active phase (say, $h = h_R(s)$), into (5.2), we find that both dh_L/ds and dh_R/ds are quite small, at most about .02. Thus, we will assume that there is a fixed value h_L of h at the jump up from the silent phase to the active phase and a fixed value h_R of h at the jump down from the active phase to the silent phase.

Now, in the active phase, we have

$$\frac{dh}{ds} = \frac{\alpha_h(V)(1 - h) - \beta_h(V)h}{\alpha(V)(1 - s) - s/\tau_{syn}}. \tag{5.3}$$

Make the further approximations that $\alpha(V) \approx \alpha$ and $dh/dt \approx -\beta h$, for α, β constant, in the active phase, and let $\tau = \alpha + 1/\tau_{syn}$. Then direct integration of (5.3) from $(h, s) = (h_L, 0)$ to $(h, s) = (h_R, s_{max})$ yields

$$s_{max} = \frac{\alpha}{\tau}(1 - H^{\tau/\beta}), \tag{5.4}$$

where $H = h_R/h_L$. Equation (5.4) gives an estimate of how the level to which the synaptic variable s recovers in the active phase depends on the parameters of the HH equations, particularly α (the approximate value of $\alpha(V)$ in (3.3)), the synaptic decay rate τ_{syn}, and the active phase decay rate of h from (3.2), approximated by β.

In Figure 5.2, we compare this approximation of s_{max} to the value obtained from numerical simulation of (3.1)–(3.3) and to an alternative, naive approximation to s_{max}, namely, $\alpha/(\alpha + \tau_{syn}^{-1})$. This corresponds to the value of s that would be reached if synapses responded instantaneously to voltage. We show how s_{max} depends on α for several values of τ_{syn}, and also how s_{max} depends on τ_{syn} for $\alpha = 2$, corresponding to the default value of α_0 for the simulations in the other sections of this paper (see Appendix A). Note that there is some ambiguity in how to select the approximate decay rate β for h, since this rate typically remains near a constant value throughout much of the active phase but then decreases near the right knee, as the decay of h slows. We neglect the slowing near the right knee, which accounts for some of the error in Figure 5.2.

It is interesting to note that for fixed α, the value of s_{max} is roughly independent of τ_{syn}, such that the active phase contributes little to the slowing that occurs as τ_{syn} is increased, as discussed in the previous sections. As α increases, s_{max} increases correspondingly. This leads to a larger s_{enter} in (4.17), which in turn yields a smaller s_{exit}. Hence, the duration of the silent phase increases with α. We explore a further implication of this dependence in section 5.3.

5.3. Bursting. Consider Figure 5.3(a). This figure shows the bifurcation structure for (3.1)–(3.2) as s varies for $g_{syn} = 2$, while Figure 5.3(b) shows the voltage trace of a two-spike burst solution to (3.1)–(3.3). This solution was obtained by greatly reducing the function $\alpha(V)$, thereby reducing the turn-on of the synapse during the active (spiking) phase. Any number of spikes can be seen in a burst by scaling the recovery function appropriately.

As we have seen, during the time that a cell spends in the silent phase, its synaptic variable decays beyond the point where the fixed point (intersection of fast and slow nullclines) of the system (3.1)–(3.2) becomes unstable (s lies below the Hopf point at $s \approx 0.22$ in Figure 5.3(a)). During the active phase, the synaptic variable s increases as specified in (3.3). If s does not recover enough to reach a value for which the fixed point of (3.1)–(3.2) is stable ($s > 0.22$ in our example), then after it jumps down to the silent phase, it will not be attracted toward the slow nullcline or the vortex structure. Instead, the orbit tends toward the fast nullcline and the phase plane looks like a standard (oscillatory) relaxation oscillator. This results in a subsequent rapid jump to the active phase when the left knee of the fast nullcline is reached, corresponding to a rapid second spike, as seen, for example, at the start of the simulation in the right panel of Figure 5.3. Alternatively, if s does increase beyond the bifurcation point, then the silent phase becomes prolonged again; however, if it is still close to the bifurcation point, the silent phase duration is still reduced relative to that seen for large s, based on (4.17). Figure 5.3 shows the recovery of the synaptic variable, s, during the two-spike burst shown in the right panel of Figure 5.3.

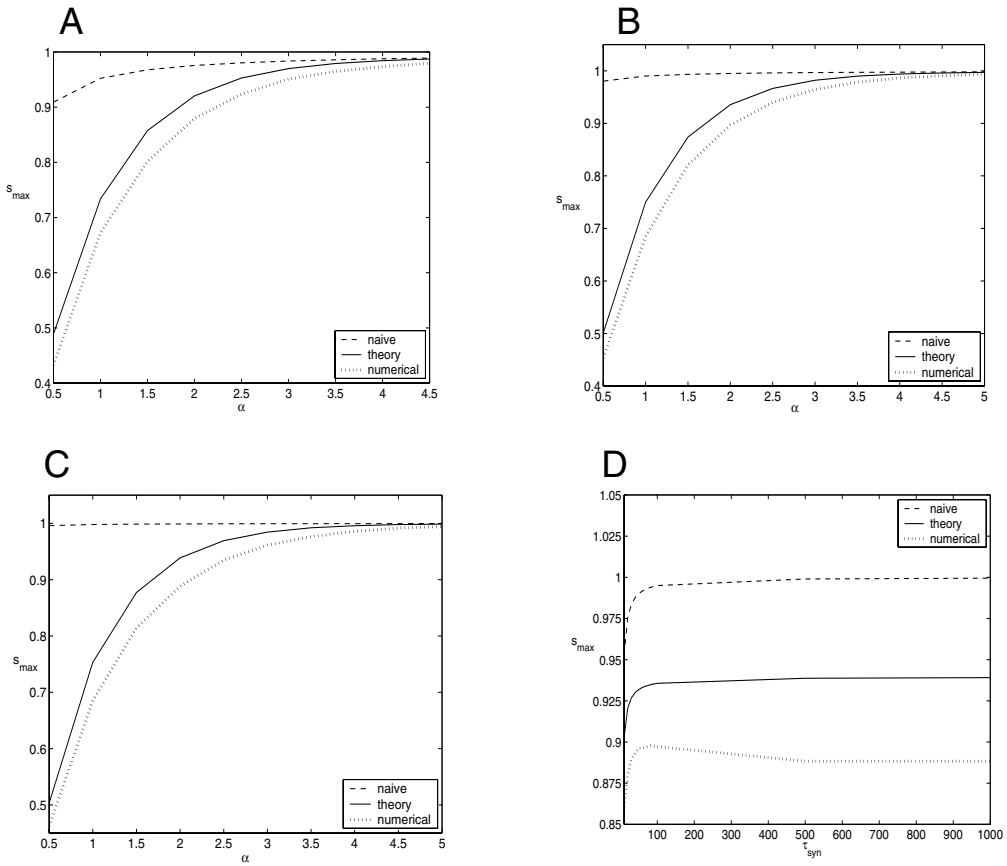


FIG. 5.2. *The dependence of synaptic recovery level,* s_{max}*, on the rate of synaptic rise* (α) *and decay* (τ_{syn})*. In each panel, the dashed line corresponds to the naive approximation* $s_{max} \approx \alpha/(\alpha + \tau_{syn}^{-1})$*, the solid line corresponds to* (5.4)*, and the thick dotted line corresponds to the actual value of* s_{max} *attained in numerical simulations of* (3.1)–(3.3). (a) $\tau_{syn} = 20$, (b) $\tau_{syn} = 100$, (c) $\tau_{syn} = 500$, (d) $\alpha = 2$.

6. Discussion. It is generally assumed that synaptic connections between excitatory neurons have the effect of strengthening and accelerating neuronal firing. Indeed, part of the accepted theory of computation in cortical circuitry is that if input is strong enough to make some excitatory cells fire, then recurrent excitation among excitatory cells amplifies this activity, whereas if inhibitory input comes in before the excitatory cells can become active, then this inhibition shuts them down. In this paper, we explore a scenario in which recurrent excitation instead causes a drastic slowing of firing. We find this effect, over a broad range of parameter values, in a network of standard, biophysically derived HH model neurons, coupled with slowly decaying synaptic excitation. This highlights the important point that the effects of synaptic inputs in neuronal networks depend on the intrinsic dynamics of the cells in the network, together with the timescale of the synaptic inputs. It remains to explore the functional consequences of this result, particularly in a network of interconnected excitatory cells and inhibitory interneurons.

Since we find that synaptic excitation is strongly synchronizing in this model network (up to small differences in subthreshold oscillations), we study the mechanism behind this synaptic slowing in a self-coupled neuron. The synchronization seen here

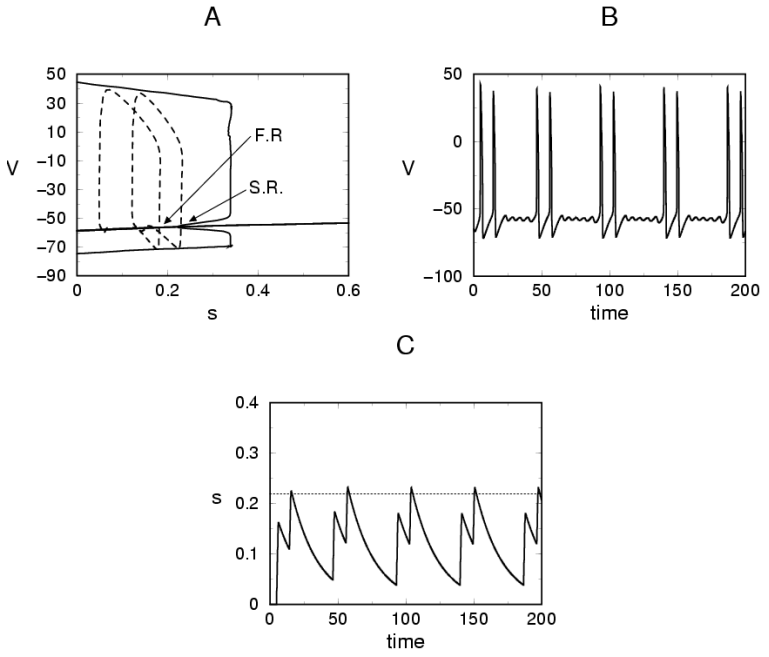


FIG. 5.3. *Bursting in the HH model.* (a) *Bifurcation diagram for the HH equations with s as the bifurcation parameter (as shown in Figure* 2.1(c)*). The curve at* $V \approx -60$ *corresponds to the critical point of system* (3.1)–(3.2) *formed by the intersection of the fast and slow nullclines. This becomes unstable via a subcritical Hopf bifurcation as s decreases. Here, F.R. and S.R. refer to the first and second return to the silent phase, respectively, of the dashed trajectory shown.* (b) *Two-spike burst solution. During the first spike of a two-spike burst, the s value does not recover enough to exit the regime where the critical point is unstable. The second recovery brings s into the stable regime, which yields a prolonged silent phase.* (c) *Synaptic variable,* $s(t)$ *during this burst. The dashed horizontal line is the value of s where the critical point (parametrized by s) changes stability. Because this stability is necessary to obtain a cycle with an arbitrarily long period, the oscillator experiences a prolonged silent phase only once s has exceeded this threshold. Parameter values for this plot are* $\tau_{syn} = 20$ *and* $\alpha_0 = 0.15$.

in part results from the phase response properties of HH neurons [11]. Further, the extreme slowing in the silent phase enhances the synchronization tendency. We have seen that this slowing involves a prolonged residence near the left knee curve of a fast nullsurface. In a population of many cells in a near-synchronized state, a strong spatial compression occurs during this residence. As soon as one cell jumps up to the active phase, fast threshold modulation (FTM) [20] will pull the other cells up as well. This compression and FTM easily overwhelm any desynchronization that may occur in the other stages of an oscillation.

We use a simplified model to elucidate the moving vortex canard mechanism by which slowly decaying synaptic excitation prolongs the silent phase between spikes, and this mechanism carries over to the HH model. The scenario that we study truly meets the criteria for a canard, since the fast (V) and slow (h) nullclines of the HH model, with s taken as a parameter, are in a regular fold canard configuration for an s-value near that at which the intersection of the nullclines loses stability via a Hopf bifurcation [3]; see also [22]. Moreover, the solutions to the full system spend a significant period of time traveling along the middle branch of the V-nullsurface

(although they remain extremely close to the curve of knees; see Figure 3.1). Unlike typical canards, however, the delayed solutions that we study are easy to find, occurring over a broad range of synaptic decay rates. We do not discuss the precise size of the region in phase space from which trajectories are drawn to the vortex region, for fixed parameter values. This may relate to attraction to a stable manifold of the $s = 0$ critical point of the HH model in the vicinity of a homoclinic bifurcation, as discussed in [9], but we have not explored this issue.

According to previous analytical results, one should be able to estimate the change in the slow variable s that will occur during the silent phase by using a way-in–way-out function [3, 15, 16]. This function incorporates information from the projected system derived by treating s as a parameter. Specifically, it involves the eigenvalues of the linearization of the projected system about an appropriate curve of critical points (parametrized by s). The eigenvalues correspond to rates of decay and growth toward this critical point curve. This approach was used previously in neuronal networks to study elliptic bursting, in which there is a delayed escape from a curve of critical points that are unstable with respect to a fast subsystem [19, 24, 14, 21]. However, the novel vortex phenomenon that we have identified causes this approach to underestimate the change in s in the silent phase, and correspondingly the time spent there, for a large range of synaptic decay rates.

The vortex structure develops through a breakdown in the distinction between fast and slow dynamics in the vicinity of the critical point curve for the projected system. The corresponding flow pins trajectories near a vortex curve, which itself lies close to the curve of critical points, for a prolonged period, as the synaptic strength gradually decays. We use the vortex curve to approximate a release threshold for the synaptic variable s, relative to a specified criterion for entrance into the trapping regime. This approach makes use of a set A, determined by the dynamics of the system, that is central to the vortex effect. In particular, A relates to the relative rates of change of the nonsynaptic slow variable and the position of the fast nullcline. Note that the position of the fast nullcline depends on the size of the synaptic variable s. Further, while there are three possible timescales corresponding to the rates of change of the three dependent variables (V, h, s) in the problem, the rate of change of the nonsynaptic slow variable (characterized by ϵ) and the synaptic decay rate $1/\tau_{syn}$ are comparable over much of the range of τ_{syn} that we consider. A full mathematical analysis of the vortex mechanism, and in particular the types of vector fields and range of timescales for which computations based on the vortex curve will always give small errors, remains open for consideration.

While we introduce the vortex mechanism and perform relevant calculations in the context of a simplified model related qualitatively to the silent phase features of the HH system, we illustrate numerically that the same ingredients are also present in the reduced HH equations (e.g., Figure 5.1). Numerical simulations of the full HH model show a similar prolongation of the silent phase, with a strong dependence on the synaptic decay rate τ_{syn}; indeed, such simulations led us to note and seek an explanation for the delay mechanism in the first place. In the reduced HH equations, we connect the active phase of oscillations to the silent phase by considering how the synaptic recovery rate α affects the level to which s recovers. This affects the level of s at which trajectories enter the trapping region (quantified by our choice of η), in turn affecting our estimation of s at release from the silent phase (see (4.17)); however, as discussed in section 5.2, the level of s at release feeds back little effect on the level to which s recovers in the active phase. By exploiting our understanding of the interaction of intrinsic and synaptic dynamics, we also describe how the fast-slow

structure allows for bursting in the HH equations. While this can be considered as elliptic bursting, the burst frequency can be quite slow, as the prolonged silent phase again occurs in the intervals between bursts of spikes.

7. Appendix A. The gating functions for h in (3.2) are

$$\alpha_h(V) = .07\exp(-(V+65)/20),$$

$$\beta_h(V) = 1/(1+\exp(-(V+35)/10)).$$

The m and n gating variables are slaved to V and h, respectively, by

$$m = \frac{\alpha_m(V)}{\alpha_m(V)+\beta_m(V)},$$

$$n = \max(.801 - 1.03h, 0),$$

where

$$\alpha_m(V) = \frac{0.1(V+40)}{1-\exp(-(V+40)/10)},$$

$$\beta_m(V) = 4\exp(-(V+65)/18).$$

The synaptic recovery function, $\alpha(V)$, is given by

$$\alpha(V) = \frac{\alpha_0}{1+\exp(-V/V_{shp})}.$$

Parameter values for all simulations are $V_{Na} = 50$, $V_K = -77$, $V_L = -54.4$, $g_{Na} = 120$, $g_K = 36$, $g_L = 0.3$, $C = 1$, $I_o = 13$, $V_{shp} = 5$, $g_{syn} = 2$, and $V_{syn} = 0$. Also, $\alpha_0 = 2$ in all sections except section 5.2, where it is varied, and section 5.3, where bursting is discussed. The units for the voltages are mV, the conductances (g_*) have units mS/cm^2, and the current (I_o) has units $\mu A/cm^2$.

8. Appendix B. Consider the model system (4.1)–(4.3), which we express as

$$\begin{aligned} \frac{dx}{dt} &= y - N_f(x,s), \\ \frac{dy}{dt} &= -\epsilon(y - N_s(x)), \\ \frac{ds}{dt} &= -\frac{s}{\tau_{syn}}. \end{aligned} \tag{8.1}$$

Note that we can express (8.1) as a pair of equations:

$$\begin{aligned} -\frac{s}{\tau_{syn}}\frac{dx}{ds} &= y - N_f(x,s), \\ \frac{s}{\epsilon\tau_{syn}}\frac{dy}{ds} &= y - N_s(x). \end{aligned} \tag{8.2}$$

To find the vortex point $(\hat{x}(s), \hat{y}(s))$ about which to linearize, we solve

$$\hat{y} = N_f(\hat{x}, s) \tag{8.3}$$

and

$$dy(\hat{x}, \hat{y})/ds = \partial N_f(\hat{x}, s)/\partial s. \tag{8.4}$$

Together with (8.4), the second equation of (8.2) gives

$$\epsilon\tau_{syn}(\hat{y} - N_s(\hat{x}))/s = \partial N_f(\hat{x}, s)/\partial s. \tag{8.5}$$

Implicit differentiation of (8.3) along the solution $(\hat{x}(s), \hat{y}(s))$ gives

$$\partial N_f(\hat{x}, s)/\partial s = d\hat{y}/ds - (\partial N_f(\hat{x}, s)/\partial x)(d\hat{x}/ds). \tag{8.6}$$

Together, (8.5) and (8.6) yield

$$\frac{d\hat{y}}{ds} = \frac{\epsilon\tau_{syn}}{s}(\hat{y} - N_s(\hat{x})) + \frac{\partial N_f(\hat{x}, s)}{\partial x}\frac{d\hat{x}}{ds}. \tag{8.7}$$

Substitute $(\hat{x}(s) + u(s), \hat{y}(s) + v(s))$ into (8.2) and linearize about $(\hat{x}, \hat{y})$ to obtain

$$\begin{aligned} -\frac{s}{\tau}\frac{du}{ds} &= \frac{s}{\tau}\frac{d\hat{x}}{ds} + \hat{y} + v - N_f(\hat{x}, s) - u(\partial N_f(\hat{x}, s)/\partial x), \\ \frac{s}{\epsilon\tau}\frac{dv}{ds} &= -\frac{s}{\epsilon\tau}\frac{d\hat{y}}{ds} + \hat{y} + v - N_s(\hat{x}) - u(dN_s(\hat{x})/dx). \end{aligned} \tag{8.8}$$

In the first equation of (8.8), $\hat{y} = N_f(\hat{x}, s)$. From (8.7), we have

$$\frac{s}{\epsilon\tau}\frac{d\hat{y}}{ds} = \hat{y} - N_s(\hat{x}) + \frac{s}{\epsilon\tau}\frac{\partial N_f(\hat{x}, s)}{\partial x}\frac{d\hat{x}}{ds}.$$

Thus, (8.8) becomes

$$\begin{aligned} -\frac{s}{\tau}\frac{du}{ds} &= \frac{s}{\tau}\frac{d\hat{x}}{ds} + v - u(\partial N_f(\hat{x}, s)/\partial x), \\ \frac{s}{\epsilon\tau}\frac{dv}{ds} &= v - u(dN_s(\hat{x})/dx) - \frac{s}{\epsilon\tau}(\partial N_f(\hat{x}, s)/\partial x)(d\hat{x}/ds). \end{aligned} \tag{8.9}$$

Note that while this is a linearized equation, the right-hand side is not linear in (u, v) because the vortex point is not a critical point of (8.2).

At this point, we make a key assumption. Since the trajectory lies in the vicinity of the knee during the time over which the vortex calculation is done, we henceforth assume that $\partial N_f(\hat{x}, s)/\partial x = 0$. In some sense, this amounts to assuming that the system is in a vortex canard configuration, since it specifies that the boundary $\partial A(s)$ should intersect $N_f(x, s)$ at the knee of $N_f(x, s)$. Clearly this assumption is not precisely satisfied; however, a straightforward generalization of the calculation below shows that any error resulting from the violation of this assumption will be of the same order of magnitude as $(\partial N_f(\hat{x}, s)/\partial x)(d\hat{x}/ds)$.

Next, we express $(u(s), v(s)) = (u_1(s), v_1(s)) + (\tilde{u}(s), \tilde{v}(s))$, where (u_1, v_1) is a zero of the right-hand side of (8.9) with $\partial N_f/\partial x = 0$; that is, (u_1, v_1) solves

$$\begin{aligned} 0 &= \frac{s}{\tau}\frac{d\hat{x}}{ds} + v, \\ 0 &= v - u(dN_s(\hat{x})/dx). \end{aligned} \tag{8.10}$$

Note that $(u_1(s), v_1(s)) = O(1/\tau_{syn})$, while $(u_1'(s), v_1'(s)) = O(1/\tau_{syn})$ as well since the determinant of coefficients $(dN_s(\hat{x})/dx) \neq 0$. Substitution of this decomposition of $(u(s), v(s))$ into (8.9) yields

$$\begin{aligned}
-\frac{s}{\tau}\frac{d\tilde{u}}{ds} &= \frac{s}{\tau}\frac{du_1}{ds} + \frac{s}{\tau}\frac{d\hat{x}}{ds} + v_1 + \tilde{v} \\
&= \frac{s}{\tau}\frac{du_1}{ds} + \tilde{v} \\
&= O(1/\tau_{syn}^2) + \tilde{v}, \\
\frac{s}{\epsilon\tau}\frac{d\tilde{v}}{ds} &= -\frac{s}{\epsilon\tau}\frac{dv_1}{ds} + v_1 - u_1(dN_s(\hat{x})/dx) + \tilde{v} - \tilde{u}(dN_s(\hat{x})/dx) \\
&= -\frac{s}{\epsilon\tau}\frac{dv_1}{ds} + \tilde{v} - \tilde{u}(dN_s(\hat{x})/dx) \\
&= O(1/\tau_{syn}) + \tilde{v} - \tilde{u}(dN_s(\hat{x})/dx),
\end{aligned}$$

where we have assumed in the final line that $\epsilon\tau_{syn} = O(1)$. Thus, when $\epsilon\tau_{syn} = O(1)$, the error in using the equation of variations in the vortex approach is of $O(1/\tau_{syn})$.

Contrast this with the usual approach, Here one solves $0 = y - N_f(x, s)$ and $0 = y - N_s(x)$ to obtain $(\tilde{x}(s), \tilde{y}(s))$. As previously (see (8.2)), we have

$$\begin{aligned}
-\frac{s}{\tau}\frac{dx}{ds} &= y - N_f(x, s), \\
\frac{s}{\epsilon\tau}\frac{dy}{ds} &= y - N_s(x),
\end{aligned}$$

and we now linearize about $(\tilde{x}(s) + u(s), \tilde{y}(s) + v(s))$ to obtain, after cancellations,

$$\begin{aligned}
-\frac{s}{\tau}\frac{du}{ds} &= \frac{s}{\tau}\frac{d\tilde{x}}{ds} + v - u(\partial N_f(\tilde{x}, s)/\partial x), \\
\frac{s}{\epsilon\tau}\frac{dv}{ds} &= -\frac{s}{\epsilon\tau}\frac{d\tilde{y}}{ds} + v - udN_s(\tilde{x})/dx.
\end{aligned}$$

We can apply the same decomposition of $(u(s), v(s)) = (u_1(s), v_1(s)) + (\tilde{u}(s), \tilde{v}(s))$ as above. However, if we again assume that $\epsilon\tau_{syn} = O(1)$, then we will have $(u_1, v_1) = O(1)$ from the $d\tilde{y}/ds$ term, and an $O(1)$ error can result from calculation with the equation of variations.

Acknowledgment. Thanks to F. Diener for providing material that would have otherwise been unobtainable.

REFERENCES

[1] S. M. Baer, T. Erneux, and J. Rinzel, *The slow passage through a Hopf bifurcation: Delay, memory effects, and resonance*, SIAM J. Appl. Math., 49 (1989), pp. 55–71.

[2] A. Compte, N. Brunel, P. S. Goldman-Rakic, and X. J. Wang, *Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model*, Cerebral Cortex, 10 (2000), pp. 910–923.

[3] M. Diener, *The canard unchained or how fast/slow dynamical systems bifurcate*, Math. Intell., 6 (1984), pp. 38–49.

[4] F. Diener and M. Diener, *Maximal delay,* in Dynamic Bifurcations, É. Benoît, ed., Lecture Notes in Math. 1493, Springer, New York, 1993, pp. 71–86.

[5] S. Doi and S. Kumagai, *Nonlinear dynamics of small scale biophysical neural networks*, in Biophysical Neural Networks, Mary Ann Liebert Inc., Larchmont, NY, 2001, pp. 261–297.

[6] B. Ermentrout, *Type* I *membranes, phase resetting curves, and synchrony*, Neural Comput., 8 (1996), pp. 979–1001.
[7] B. Ermentrout, *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*, Software Environ. Tools 14, SIAM, Philadelphia, 2002.
[8] J. Guckenheimer, R. Harris-Warrick, J. Peck, and A. Willms, *Bifurcation, bursting, and spike frequency adaptation*, J. Comp. Neurosci., 4 (1997), pp. 257–277.
[9] J. Guckenheimer and A. Willms, *Asymptotic analysis of subcritical Hopf-homoclinic bifurcation*, Phys. D, 139 (2000), pp. 195–216.
[10] B. S. Gutkin, C. R. Laing, C. L. Colby, C. C. Chow, and G. B. Ermentrout, *Turning on and off with excitation: The role of spike-timing asynchrony and synchrony in sustained neural activity*, J. Comput. Neurosci., 11 (2001), pp. 121–134.
[11] D. Hansel, G. Mato, and C. Meunier, *Synchrony in excitatory neural networks*, Neural Comput., 7 (1995), pp. 307–337.
[12] D. Hansel and G. Mato, *Existence and stability of persistent states in large neuronal networks*, Phys. Rev. Lett., 86 (2001), pp. 4175–4178.
[13] A. L. Hodgkin and A. F. Huxley, *A quantitative description of the membrane current and its application to conduction and excitation in nerves*, J. Physiol. (Lond.), 117 (1952), pp. 500–544.
[14] F. C. Hoppensteadt and E. M. Izhikevich, *Weakly Connected Neural Networks*, Springer, New York, 1997.
[15] A. I. Neishtadt, *Prolongation of the loss of stability in the case of dynamic bifurcations.* I, Differential Equations, 23 (1987), pp. 1385–1390.
[16] A. I. Neishtadt, *Prolongation of the loss of stability in the case of dynamic bifurcations.* II, Differential Equations, 24 (1988), pp. 171–176.
[17] D. J. Pinto and G. B. Ermentrout, *Spatially structured activity in synaptically coupled neuronal networks:* I. *Traveling fronts and pulses*, SIAM J. Appl Math., 62 (2001), pp. 206–225.
[18] J. Rinzel, *Excitation dynamics: Insights from simplified membrane models*, Fed. Proc., 44 (1985), pp. 2944–2946.
[19] J. Rinzel, *A formal classification of bursting mechanisms in excitable systems*, in Proceedings of the International Congress of Mathematicians, A. M. Gleason, ed., AMS, Providence, RI, 1987, pp. 1578–1593.
[20] D. Somers and N. Kopell, *Rapid synchronization through fast threshold modulation*, Biol. Cybern., 68 (1993), pp. 393–407.
[21] J. Su, J. Rubin, and D. Terman, *Effects of noise on elliptic bursters*, Nonlinearity, 17 (2004), pp. 133–157.
[22] P. Szmolyan and M. Wechselberger, *Canards in* $\mathbf{R}^3$, J. Differential Equations, 177 (2001), pp. 419–453.
[23] R. D. Traub and R. Miles, *Neuronal Networks of the Hippocampus*, Cambridge University Press, Cambridge, UK, 1991.
[24] X.-J. Wang and J. Rinzel, *Oscillatory and bursting properties of neurons*, in Handbook of Brain Theory and Neural Networks, M. A. Arbib, ed., MIT Press, Cambridge, MA, 1995, pp. 689–691.

SIAM J. APPL. MATH.
Vol. 65, No. 1, pp. 93–112

COMPUTATION OF THE EFFECTIVE DIFFUSIVITY TENSOR FOR TRANSPORT OF A PASSIVE SCALAR IN A TURBULENT INCOMPRESSIBLE FLOW*

T. KOMOROWSKI†‡ AND P. WIDELSKI‡

Abstract. We consider the passive scalar transport in an incompressible random flow. Our basic result is a proof of the convergence of a certain numerical scheme for the computation of the eddy diffusivity tensor. The scheme leads to the formula for the diffusivity expressed in terms of an infinite series. We give a rigorous proof of the geometric bounds on the magnitude of the terms of the series, provided that the Eulerian field is Markovian and Gaussian and its temporal dynamics has a sufficiently large spectral gap. The principal tools used in the proofs are the decomposition of the space of square integrable fields formed over the possible realizations of the Eulerian velocity field in the Gaussian chaos and the hypercontractivity properties of Gaussian measures.

Key words. Gaussian Ornstein–Uhlenbeck field, mixing, convection-diffusion equation, passive scalar

AMS subject classifications. Primary, 60F17, 35B27; Secondary, 60G44

DOI. 10.1137/S003613990343519X

1. Introduction.

The transport of a passive scalar field $T(\cdot,\cdot)$ in a turbulent flow can be modeled by the convection-diffusion equation with a random drift,

$$\begin{cases} \partial_t T(t,\mathbf{x}) + \mathbf{u}\,(t,\mathbf{x}) \cdot \nabla_{\mathbf{x}} T(t,\mathbf{x}) = \kappa \Delta_{\mathbf{x}} T(t,\mathbf{x}), \\ T(0,\mathbf{x}) = T_0(\mathbf{x}). \end{cases} \tag{1.1}$$

Here $\mathbf{u} = (u_1, \ldots, u_d) : \mathbb{R} \times \mathbb{R}^d \times \Omega \to \mathbb{R}^d$ is a d-dimensional, $d \geq 2$, random field, usually called *the Eulerian velocity*, given over a certain probability space $\mathcal{T} := (\Omega, \mathcal{V}, \mathbb{P})$, and $T_0(\cdot)$ is a deterministic initial condition. The drift models the turbulent convection by a flow of a certain fluid. It is therefore assumed to be time-space homogeneous, ergodic, centered, and incompressible, i.e., $\nabla_{\mathbf{x}} \cdot \mathbf{u}(t,\mathbf{x}) := \sum_{i=1}^{d} \partial_{x_i} u_i(t,\mathbf{x}) \equiv 0$. The parameter $\kappa > 0$, called *the molecular diffusivity*, describes the strength of the diffusive dispersion of the medium.

The passage from the microscopic to macroscopic description of transport is obtained by an appropriate change of scales. For example, under the diffusive scaling the macroscopic coordinates $(t', \mathbf{x}')$ are given by $t \sim t'/\varepsilon^2$ and $\mathbf{x} \sim \mathbf{x}'/\varepsilon$, where $\varepsilon \ll 1$ is a certain small parameter. Suppose also that the initial data varies on the macroscopic scale, so it is of the form $T_0(\varepsilon \mathbf{x})$. Then, in the macroscopic coordinates the rescaled field $T_\varepsilon(t,\mathbf{x}) = T(t/\varepsilon^2, \mathbf{x}/\varepsilon)$ (we omit primes here) satisfies

$$\begin{cases} \partial_t T_\varepsilon(t,\mathbf{x}) + \frac{1}{\varepsilon} \mathbf{u}\left(\frac{t}{\varepsilon^2}, \frac{\mathbf{x}}{\varepsilon}\right) \cdot \nabla_{\mathbf{x}} T_\varepsilon(t,\mathbf{x}) = \kappa \Delta_{\mathbf{x}} T_\varepsilon(t,\mathbf{x}), \\ T_\varepsilon(0,\mathbf{x}) = T_0(\mathbf{x}). \end{cases} \tag{1.2}$$

One feature that can be proved about the scaled solution, under appropriate assumptions on the statistics of the drift, is the self-averaging property of the scaled scalar

*Received by the editors September 24, 2003; accepted for publication (in revised form) April 26, 2004; published electronically September 24, 2004.
http://www.siam.org/journals/siap/65-1/43519.html

†Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland. The research of this author was partially supported by KBN grant 2PO3A 031 23.

‡Institute of Mathematics, UMCS, pl. Marii Curie-Skłodowskiej 1, 20-031 Lublin, Poland (komorow@hektor.umcs.lublin.pl, pwidel@golem.umcs.lublin.pl).

field. In the weak form it can be stated as follows:

$$\lim_{\varepsilon\to 0+}\left\langle\left[\int_{\mathbb{R}^d} T_\varepsilon(t,\mathbf{x})\varphi(\mathbf{x})d\mathbf{x}-\int_{\mathbb{R}^d} T^*(t,\mathbf{x})\varphi(\mathbf{x})d\mathbf{x}\right]^2\right\rangle=0\quad\forall\,\varphi\in C_0^\infty(\mathbb{R}^d). \tag{1.3}$$

Here $\langle\cdot\rangle$ denotes the averaging over the realizations of the drift and T^* is a (deterministic) solution of a constant coefficient heat equation

$$\begin{cases}\partial_t T^*(t,\mathbf{x})=\sum_{i,j=1}^d K^*_{i,j}\partial^2_{x_i,x_j}T^*(t,\mathbf{x}),\\ T^*(0,\mathbf{x})=T_0(\mathbf{x}).\end{cases} \tag{1.4}$$

$\mathbf{K}^*=[K^*_{i,j}]$—*the effective diffusivity* tensor—is a constant matrix. Since the procedure described above eliminates the inhomogeneity that appears in the convection-diffusion equation on the microscopic level and leads to a space-time homogeneous equation (1.4), it is sometimes referred to as *homogenization*. To substantiate the averaging property claimed in (1.3) one could first rewrite (1.2) in the divergence form. Since the drift $\mathbf{u}(t,\mathbf{x})$ is divergence-free, there exists an antisymmetric tensor-valued potential $\mathbf{H}(t,\mathbf{x})=[H_{p,q}(t,\mathbf{x})]$, $H_{p,q}(t,\mathbf{x})=-H_{q,p}(t,\mathbf{x})$ such that $\mathbf{u}(t,\mathbf{x})=\nabla_{\mathbf{x}}\cdot\mathbf{H}(t,\mathbf{x})$ and the equation in question takes the form $\partial_t T_\varepsilon(t,\mathbf{x})-\nabla_{\mathbf{x}}\cdot(a(t/\varepsilon^2,\mathbf{x}/\varepsilon)\nabla_{\mathbf{x}}T_\varepsilon(t,\mathbf{x}))=0$, where the diffusivity matrix $a(t,\mathbf{x})=[a_{p,q}(t,\mathbf{x})]$ is given by $a_{p,q}(t,\mathbf{x})=\kappa\delta_{p,q}-H_{p,q}(t,\mathbf{x})$. A rigorous proof of self-averaging for parabolic operators of this form, when the stream matrix $\mathbf{H}(t,\mathbf{x})$ is an L^∞ bounded, time-space homogeneous, and ergodic random field, follows from the results obtained by Zhikov, Kozlov, and Olejnik in [13, Theorem 1, p. 187]. However, if $\mathbf{H}(t,\mathbf{x})$ admits unbounded realizations but possesses an absolute pth moment, where $p>d+2$, as is the case in the present paper, then the averaging in the sense of (1.2) can be concluded from the quenched version of the invariance principle for random characteristics of (1.2) and has been shown in [1, 2]. The aforementioned homogenization results hold in both time dependent and static (time independent) cases. In [7] self-averaging is also shown for Gaussian time dependent drifts, for which the stream matrix need not exist. The temporal dynamics of the field is assumed to be Markovian and uniformly mixing on all spatial scales, i.e., possesses *a spectral gap*; see [7, Theorem 1, p. 528].

Because the proof of the existence of the effective diffusivity matrix is a result of an application of an appropriate ergodic theorem, what is usually left unanswered by the homogenization theorems is how to calculate the effective diffusivity tensor from the statistics of the Eulerian velocity. In this paper we take up the task of providing a formula for computing the effective diffusivity. We consider random drifts that are space-time homogeneous, Gaussian, and Markov and whose spectral dynamics possesses sufficiently strong spectral gap.

For the family of the Eulerian velocity fields described above we present a rigorous scheme for calculation of the effective diffusivity matrix that results in an infinite series expansion; see (3.18) below. In addition, we provide a geometric bound for the nth term of this series (see Theorem 3.1 below), which results in the control of the series tails (see Corollary 3.2). This control holds provided the spectral gap of the Eulerian field dynamics is sufficiently large. The precise estimate of the size of the spectral gap is possible thanks to formula (3.6).

Another interesting question pertaining to the model with a Gaussian drift is how to relate $\mathbf{K}^*$ to the autocovariance tensor

$$\mathbf{R}(t-s,\mathbf{x}-\mathbf{y}):=[\langle u_i(t,\mathbf{x})u_j(s,\mathbf{y})\rangle], \tag{1.5}$$

which, as is well known, characterizes the Gaussian drift $\mathbf{u}(\cdot,\cdot)$. To simplify our calculations we suppose further that the field is spatially isotropic and its spectral gap is identical for all spatial scales. The specific form of the spectrum of the covariance matrix is presented in (2.1) below. However, as becomes apparent during the course of the argument, our proofs do not depend on isotropy of the drift. We could also admit fields whose mixing rates vary on different scales, so long as they are bounded away from zero by a certain constant that is not too small. In section 4 we give a formula for the nth term of the series for the effective diffusivity in terms of the spectrum of the velocity field; see (4.10). Due to the fact that this term is in principle, i.e., discarding some possible cancellations, a sum of $n!$ terms, the computational value of this formula would be severely limited. Thanks to Theorem 3.1, however, we are able to control the size of a particular term of the series.

Let us describe briefly the main ideas used in the derivation of the formula for the effective diffusivity. The computation is contingent on finding, in an appropriate space, the solution of the cell problem (2.14); in the literature of the subject it is called *the corrector field*. In section 3 below we propose a numerical scheme for computing this field; see the formulas (3.1) and (3.5) for the definition of the scheme. The corrector field is then given by (3.17), and the eddy diffusivity can be calculated using (3.18). To gain appropriate estimates on the L^2-norm of the nth term ψ_n appearing in the scheme (see (3.4) for its definition), we use the decomposition of the space of square integrable functionals formed over the Eulerian velocity field at the given snapshot of time (say, when $t = 0$) into the Gaussian chaos. This together with the hypercontractivity property of Gaussian measure and the spectral gap estimate for the dynamics of the Eulerian field produce geometric bounds for the respective term; see Proposition 3.1 for the precise statement of the bounds.

Finally, in section 4 we provide an explicit formula for the nth term of the series (3.18); see Proposition 4.3. The formula is stated using the language of Feynman graphs. As we have already mentioned, in practical calculations, this formula should be coupled with the estimate (3.19) on the tails of the series expansion for the effective diffusivity.

2. The description of the model.

2.1. Homogeneous Gaussian random drifts. We suppose the following:

(V1) $\mathbf{u} : \mathbb{R} \times \mathbb{R}^d \times \Omega \to \mathbb{R}^d$ is a zero mean, space-time homogeneous, spatially isotropic, Gaussian random field over the probability space $\mathcal{T}$.

(V2) The autocovariance matrix of the field (cf. (1.5)) is given by

$$\mathbf{R}(t-s, \mathbf{x}-\mathbf{y}) = b \int_{\mathbb{R}^d} \cos((\mathbf{x}-\mathbf{y})\cdot\mathbf{k})\, e^{-a|t-s|} \mathcal{E}(|\mathbf{k}|)|\mathbf{k}|^{1-d} \hat{\Gamma}(\mathbf{k})\, d\mathbf{k}. \tag{2.1}$$

Here $a, b > 0$, $\hat{\Gamma}(\mathbf{k}) = [\hat{\Gamma}_{i,j}(\mathbf{k})]$, with $\hat{\Gamma}_{i,j}(\mathbf{k}) := \delta_{i,j} - k_i k_j |\mathbf{k}|^{-2}$. We assume that the power energy spectrum satisfies the power law

$$\mathcal{E}(k) := \mathbf{1}_{[0,K_0]}(k) k^{1-2\alpha}, \tag{2.2}$$

where $K_0 > 0$ is fixed and $\alpha < 1$, to ensure the L^2-integrability of the field.

Remark 2.1. It is well known that, thanks to (2.2), such a random field possesses a modification that is $\mathbb{P}$ a.s. jointly locally Hölder continuous in $(t, \mathbf{x}) \in \mathbb{R} \times \mathbb{R}^d$ and C^∞ smooth in $\mathbf{x}$ for any fixed $t \in \mathbb{R}$.

Remark 2.2. A direct calculation yields

$$\langle|\mathbf{u}(0,\mathbf{0})|^2\rangle = b(d-1)\omega_{d-1}\int_0^{K_0}\frac{dk}{k^{2\alpha-1}} = \frac{b}{2}(d-1)\omega_{d-1}K_0^{2(1-\alpha)}(1-\alpha)^{-1}, \tag{2.3}$$

where ω_{d-1} denotes the surface measure of $\mathbb{S}^{d-1}$—the unit sphere in $\mathbb{R}^d$.

Remark 2.3. The presence of the factor $\hat{\Gamma}(\cdot)$ in the formula ensures that the spatial realizations of the field are incompressible. The parameter $a > 0$, called *the spectral gap*, controls the rate at which the field decorrelates in the temporal variable. In our model we assume that this rate is constant on all spatial scales. Let us mention here one particular case that has been widely studied in the literature. When $a = b$ and $a \to +\infty$, the autocovariance matrix converges (in the distribution sense) to the autocovariance matrix of a δ-correlated velocity field, the so-called Kraichnan model (see [10]) given by

$$\mathbf{R}(t-s,\mathbf{x}-\mathbf{y}) = \delta(t-s)\int_{\mathbb{R}^d}\cos((\mathbf{x}-\mathbf{y})\cdot\mathbf{k})\,\mathcal{E}(|\mathbf{k}|)|\mathbf{k}|^{1-d}\,\hat{\Gamma}(\mathbf{k})\,d\mathbf{k}. \tag{2.4}$$

The effective diffusivity in this case is given explicitly and equals

$$\kappa_* = \kappa + \frac{1}{2}\int_{|\mathbf{k}|\le K_0}\mathcal{E}(|\mathbf{k}|)|\mathbf{k}|^{1-d}\hat{\Gamma}_{11}(\mathbf{k})d\mathbf{k} = \kappa + \frac{1}{4}\omega_{d-1}\left(1-\frac{1}{d}\right)K_0^{2(1-\alpha)}(1-\alpha)^{-1}.$$

2.2. Formula for effective diffusivity. An abstract cell problem. Thanks to isotropy of the Eulerian velocity field $\mathbf{u}(\cdot,\cdot)$, the effective diffusivity tensor must commute with any rotation. Hence it is of the form

$$\mathbf{K}^* = \kappa_*\mathbf{I}, \text{ where } \kappa_* = \kappa + d_* \tag{2.5}$$

and d_* is called *eddy diffusivity*. In order to determine eddy diffusivity, one needs to solve an auxiliary *cell problem for the corrector*; see, e.g., [7, (4.16), p. 537]. To formulate this problem, we need to introduce an appropriate functional space that is big enough to contain all possible spatial realizations of the velocity field at any given time instant; see also [8].

Suppose that m is a positive integer and $\vartheta_\rho(\mathbf{x}) := (1+|\mathbf{x}|^2)^{-\rho}$, $\mathbf{x}\in\mathbb{R}^d$, where $\rho > d/2$. Let $\mathcal{H}$ be the Hilbert space of d-dimensional incompressible vector fields that is the completion of $C^\infty_{0,div} := \{f\in C^\infty_0(\mathbb{R}^d;\mathbb{R}^d) : \nabla_{\mathbf{x}}\cdot f = 0\}$ w.r.t. the norm

$$||f||^2_{\mathcal{H}} := \int_{\mathbb{R}^d}(|f(\mathbf{x})|^2 + |\nabla_{\mathbf{x}}f(\mathbf{x})|^2 + \cdots + |\nabla^m_{\mathbf{x}}f(\mathbf{x})|^2)\vartheta_\rho(\mathbf{x})\,d\mathbf{x}.$$

We can always assume that m is big enough (e.g., $m > d/2+1$), so, thanks to the Sobolev embedding, any $f\in\mathcal{H}$ is of the C^1 class of regularity. The presence of a weight $\vartheta_\rho(\cdot)$ follows from the fact that for a given t the spatial realizations of the Gaussian field $\mathbf{u}(t,\cdot)$ grow, as $C\log^{1/2}|\mathbf{x}|$, for $|\mathbf{x}|\gg 1$; see, e.g., [12].

Let μ be the law in $\mathcal{H}$ of the Gaussian velocity field $\mathbf{u}(0,\cdot)$. Denote $L^2 := L^2(\mu)$, L^2_0 its subspace consisting of F such that $\int F\,d\mu = 0$. The measure μ is Gaussian of zero mean; i.e., $\int f(\mathbf{0})\mu(df) = \mathbf{0}$, with autocovariance

$$\int f(\mathbf{x})\otimes f(\mathbf{y})\mu(df) = b\int_{\mathbb{R}^d}\cos((\mathbf{x}-\mathbf{y})\cdot\mathbf{k})\,\frac{\mathcal{E}(|\mathbf{k}|)}{|\mathbf{k}|^{d-1}}\,\hat{\Gamma}(\mathbf{k})\,d\mathbf{k}. \tag{2.6}$$

For any two vectors $\mathbf{a} = (a_1, \dots, a_d)$, $\mathbf{b} = (b_1, \dots, b_d)$ the symbol $\mathbf{a} \otimes \mathbf{b}$ denotes a $d \times d$ matrix $[a_i b_j]$. Equation (2.6) implies homogeneity of μ, i.e., $\mu \tau_{\mathbf{x}} = \mu \ \forall \mathbf{x} \in \mathbb{R}^d$, where $\tau_{\mathbf{x}} : \mathcal{H} \to \mathcal{H}$ is given by $\tau_{\mathbf{x}} f(\cdot) := f(\mathbf{x} + \cdot)$. We define *the abstract spatial gradient operator* ∇ as follows. Let

$$D_p F(f) := \partial_{x_p | \mathbf{x}=\mathbf{0}} F(\tau_{\mathbf{x}}(f)), \quad p = 1, \dots, d, \tag{2.7}$$

for $F \in L^2$, such that all the partial derivatives on the right-hand side of (2.7) exist in the L^2 sense. Let $\nabla F := (D_1 F, \dots, D_d F)$. The abstract Laplace operator Δ can be defined as

$$\Delta F = \sum_{p=1}^{d} D_p^2 F$$

for those F for which the second partials exist in the L^2 sense.

Remark 2.4. Observe that, due to incompressibility of $\mathbf{u}$, we have

$$(\tilde{\mathbf{u}} \cdot \nabla F, G)_{L^2} = -(\tilde{\mathbf{u}} \cdot \nabla G, F)_{L^2}. \tag{2.8}$$

Here

$$\tilde{\mathbf{u}}(f) = (\tilde{u}_1(f), \dots, \tilde{u}_d(f)) := f(\mathbf{0}). \tag{2.9}$$

Remark 2.5. The operator $\kappa\Delta$ is the generator of a semigroup of symmetric Markov contractions on L^2, given by

$$S(t)F(f) = \int_{\mathbb{R}^d} r_\kappa(t, \mathbf{y}) F(\tau_{\mathbf{y}} f) d\mathbf{y}, \quad F \in L^2, \tag{2.10}$$

where

$$r_\kappa(t, \mathbf{x}) := \frac{1}{(4\kappa\pi t)^{d/2}} \exp\left\{ -\frac{|\mathbf{x}|^2}{4\kappa t} \right\}. \tag{2.11}$$

It is therefore a self-adjoint, negative definite operator.

To introduce the temporal derivative $D_0 F$ we need to describe in more detail the dynamics of the $\mathcal{H}$-valued stochastic process $(\mathbf{u}(t, \cdot))_{t \geq 0}$. This process can be thought of as the time stationary solution of an $\mathcal{H}$-valued linear stochastic differential equation

$$du(t) = -au(t)dt + \sqrt{2a} B dW(t), \tag{2.12}$$

with $u(0) = \mathbf{u}(0, \cdot)$. Here $W(\cdot)$ is a cylindrical Wiener process on $L^2_{div}(\mathbb{R}^d, \mathbb{R}^d)$—the space of all square integrable incompressible d-dimensional vector fields—defined over the probability space $\mathcal{T}$, and $B : L^2_{div}(\mathbb{R}^d, \mathbb{R}^d) \to \mathcal{H}$ is a Hilbert–Schmidt operator given by

$$\widehat{B\psi}(\mathbf{k}) = \sqrt{\mathcal{E}\left(|\mathbf{k}|\right)} |\mathbf{k}|^{(1-d)/2} \hat{\psi}(\mathbf{k}). \tag{2.13}$$

Both here and below $\hat{\psi}$ denotes the Fourier transform of a given function ψ. We refer the reader to [7] for details of this construction.

The Eulerian velocity field $\mathbf{u}(t, \mathbf{x})$ can be then identified with $u(t)(\mathbf{x})$. Let $D_0 : D(D_0) \to L^2$ be the generator of $u(\cdot)$ and $(P(t))_{t \geq 0}$ the corresponding semigroup of

Markov operators. It can easily be shown that the dynamics described by (2.12) and (2.13) is *reversible*; i.e., each $P(t)$ is self-adjoint (see [7, (3.4), p. 530]).

The abstract cell problem can be formulated as

$$\kappa\Delta\chi + D_0\chi - \tilde{\mathbf{u}}\cdot\nabla\chi = \tilde{u}; \tag{2.14}$$

see [7, (4.16), p. 537]. Here $\tilde{\mathbf{u}}$ is given by (2.9) and $\tilde{u}$ denotes one of the components of $\tilde{\mathbf{u}}$. To fix our attention we shall admit $\tilde{u} := \tilde{u}_1$. The unknown field χ is called a *corrector*. Thanks to the assumed isotropy of the field, *the eddy diffusivity* d_* is given by

$$d_* := -(\tilde{u}, \chi)_{L^2}; \tag{2.15}$$

see [7, (5.10), p. 544]. Unfortunately, (2.14) does not need to have a solution in L^2. This is, for instance, a typical situation in the case of static fields. Then, one can only guarantee the existence of χ in a distribution sense. However, under the assumption that the dynamics of $\mathbf{u}(\cdot)$ possesses a spectral gap, i.e., its spectral measure is of the form (2.1) with $a > 0$, one can show that there exists a unique $\chi \in L^2$ solving (2.14) and satisfying $\int \chi d\mu = 0$; see [7, (4.16), p. 537].

It is immensely difficult in general to perform explicit calculations of eddy diffusivity with the help of formula (2.15). This is due to the fact that the cell problem (2.14) is formulated for functionals defined over an infinite dimensional Hilbert space and it is very seldom possible to solve it explicitly. However, in the Gaussian case it is possible to give a numerical scheme for calculating eddy diffusivity with the help of a decomposition of L^2 into Gaussian chaos.

2.3. The decomposition of L^2 using Gaussian chaos. Let $\mathcal{P}_n$ be the L^2 closure of the linear space spanned by the monomials $f \mapsto \langle\varphi_1, f\rangle \dots \langle\varphi_m, f\rangle$, where $m \le n$, $\varphi_1, \dots, \varphi_m \in \mathcal{S}_d$. Here $\mathcal{S}_d$ is the space of incompressible fields belonging to the Schwartz class and $\langle g, f\rangle := \int_{\mathbb{R}^d} f(\mathbf{x})\cdot g(\mathbf{x}) d\mathbf{x}$ for any $f \in \mathcal{S}_d$, $g \in \mathcal{H}$. $H_n := \mathcal{P}_n \ominus \mathcal{P}_{n-1}$ is called *the space of nth degree Hermite polynomials*. We denote by Π_n the orthogonal projection of L^2 onto H_n. It is well known (see, e.g., [6, Theorem 2.6, p. 16]) that $L^2 = \bigoplus_{n\ge 0} H_n$. Thanks to the fact that the group induced by shift $U^{\mathbf{x}}F := F\tau_{\mathbf{x}}$ leaves each H_n invariant, we see from (2.10) that $S(t)(H_n) = H_n$ for each n.

The Hermite polynomials also provide for a neat description of the operator D_0. Namely, for any $F \in H_n$ we have $F \in D(D_0)$ and

$$D_0 F = -anF. \tag{2.16}$$

Equation (2.16) can be seen by using the formula for the generator of $(\mathbf{u}(t,\cdot))_{t\ge 0}$ contained on p. 207 of [9]. A simple consequence of (2.16) is the following *spectral gap estimate* of D_0:

$$-(D_0F, F)_{L^2} \ge a\|F\|^2_{L^2} \quad \text{for any } F \in L^2 \text{ s.t. } \int F d\mu = 0. \tag{2.17}$$

Another consequence of (2.16) and the fact that each H_n is invariant under semigroup $S(\cdot)$ is commutation relation $P(t)S(s) - S(s)P(t) = 0$ for arbitrary $t, s \ge 0$. Hence $R(t) := S(t)P(t)$, $t \ge 0$, defines a semigroup of self-adjoint operators, whose generator equals $\kappa\Delta + D_0$ on a sufficiently large subspace $\mathcal{C}$ that constitutes a core for both $\kappa\Delta$, D_0. One can, for instance, take

$$\mathcal{C} := \bigcap_{p\in(2,\infty)} W^{2,p} \cap D(D_0),$$

where $W^{m,p}$ denotes the closure of those F, for which $\mathbf{x} \mapsto F\tau_{\mathbf{x}}$ possesses m derivatives at $\mathbf{0}$ in the norm given by $\|F\|^p_{m,p} := \sum_{i_1+\cdots+i_d\le m} \|D_1^{i_1} \dots D_d^{i_d} F\|^p_{L^p}$.

2.4. The integral representation of the corrector. Let $\mathcal{L}F := \kappa\Delta F + D_0 F - \tilde{\mathbf{u}} \cdot \nabla F$, $F \in \mathcal{C}$. It can be shown (see [7, Proposition 3, p. 536]) that the closure of $\mathcal{L}$ is a generator of a semigroup $(Q(t))_{t\ge 0}$ of Markov operators on L^2 that leaves $\mathcal{C}$ invariant. This semigroup is exponentially stable with $\|Q(t)F\|_{L^2} \le e^{-at}\|F\|_{L^2}$ for any F satisfying $\int F d\mu = 0$ (see [7, (4.5), p. 536]). Using the semigroup, we can write the unique zero mean solution to the Poisson equation (2.14) (cf. [7, (4.16), p. 537]):

$$\chi := -\int_0^\infty Q(t)\tilde{u}dt. \tag{2.18}$$

Since $\tilde{u} \in \mathcal{C}$ we conclude that also $\chi \in \mathcal{C}$. From (2.14) and the spectral gap estimate (2.17) we can also infer that for a nontrivial Eulerian velocity field

$$d_* = -(\mathcal{S}\chi, \chi)_{L^2} = \kappa\|\nabla\chi\|^2_{L^2} - (D_0\chi, \chi)_{L^2} > \kappa\|\nabla\chi\|^2_{L^2} > 0,$$

and in consequence $\kappa_* > \kappa$. The latter inequality highlights a well-known physical fact that in an incompressible medium a turbulent convection enhances diffusive properties of the medium.

Remark 2.6. Note that in our case the "time derivation operator" D_0 is self-adjoint and satisfies the spectral gap estimate (2.17). A differentiation w.r.t. to time operator D_0 can be in fact introduced for environments that are only stationary in t (with no additional assumption on the structure of their temporal dynamics); cf. [13, p. 193]. However, when D_0 is defined in such a general way, discarding additional information about the dissipative properties of the evolution of the environment in time, it is an anti-self-adjoint operator (conditioning on the environment up to a given moment of time becomes averaging w.r.t. a trivial, deterministic, probability measure), and the spectral gap inequality (2.17) fails to be true. In fact, then $(D_0F, F)_{L^2} = 0 \ \forall\ F$. In consequence, the definition of χ via (2.18) would not make sense in L^2 under such general circumstances, but one can make sense of it in an appropriate distribution space (see [13, p. 194]).

3. Numerical scheme for calculation of the effective diffusivity. Since $\tilde{u} \in \mathcal{P}_1$, $\int \tilde{u}d\mu = 0$, and the spectrum of $\kappa\Delta + D_0$ restricted to L^2_0 has a gap of size at least $a > 0$, we can find a unique $\psi_1 \in \mathcal{P}_1$ that is the solution of

$$(\kappa\Delta + D_0)\psi_1 = \tilde{u} \tag{3.1}$$

and satisfies $\int \psi_1 d\mu = 0$. In fact

$$\psi_1 := -\int_0^\infty R(t)\tilde{u}dt. \tag{3.2}$$

Suppose we have already defined $\psi_n \in \mathcal{P}_n$, $n = 1, \dots, N$, for a certain N. Set

$$u_N := \tilde{\mathbf{u}} \cdot \nabla\psi_N \in \mathcal{P}_{N+1}. \tag{3.3}$$

Note that

$$\int u_N d\mu = \int \nabla \cdot (\tilde{\mathbf{u}}\psi_N)d\mu = -\int \nabla\mathbf{1} \cdot \tilde{\mathbf{u}}\psi_N d\mu = 0,$$

so

$$\psi_{N+1} := -\int_0^\infty R(t)u_N dt \in \mathcal{P}_{N+1} \tag{3.4}$$

is the unique zero mean solution of

$$(D_0 + \kappa\Delta)\psi_{N+1} = u_N. \tag{3.5}$$

THEOREM 3.1. *Let*

$$q := \left(\frac{6}{a\kappa}\right)^{1/2} \|\tilde{\mathbf{u}}\|_{L^2}. \tag{3.6}$$

Then,

$$\|\psi_N\|_{L^2} \le q^{N-1}\|\psi_1\|_{L^2} \quad \forall\, N \ge 1. \tag{3.7}$$

Proof. Denote $\psi_{N,k} := \Pi_k\psi_N$ (the projection onto the space of kth degree Hermite polynomials) and

$$|\!|\!|F|\!|\!|^2 := \sum_{k=1}^{+\infty} k\|\Pi_k F\|_{L^2}^2.$$

Obviously $\psi_{N,k} = 0$ for $k > N$. Projecting both sides of (3.5) onto H_k and taking the scalar product against $\psi_{N+1,k}$, we obtain

$$ak\|\psi_{N+1,k}\|_{L^2}^2 + \kappa\|\nabla\psi_{N+1}\|_{L^2}^2 \tag{3.8}$$

$$= -(\tilde{\mathbf{u}}\cdot\nabla\psi_{N,k+1}, \psi_{N+1,k})_{L^2} - (\tilde{\mathbf{u}}\cdot\nabla\psi_{N,k-1}, \psi_{N+1,k})_{L^2}.$$

The last equality is a consequence of the following two observations. First, note that $\Pi_k(\tilde{\mathbf{u}}\cdot\nabla\psi_{N,l}) \neq 0$ only when $l = k-1, k, k+1$. In addition, thanks to (2.8),

$$(\Pi_k(\tilde{\mathbf{u}}\cdot\nabla\psi_{N,k}), \psi_{N,k})_{L^2} = (\tilde{\mathbf{u}}\cdot\nabla\psi_{N,k}, \psi_{N,k})_{L^2} = 0.$$

To estimate the right-hand side of (3.8), we consider two cases. First, when $k = 1$, the second term on the right-hand side vanishes. On the other hand, thanks to incompressibility of $\mathbf{u}(\cdot,\cdot)$, the first term equals $(\tilde{\mathbf{u}}\cdot\nabla\psi_{N+1,1}, \psi_{N,2})_{L^2}$. Its absolute value is, by virtue of the Cauchy–Schwarz inequality, less than or equal to

$$\begin{aligned}
&\|\psi_{N,2}\|_{L^2}\left[\sum_{i,j=1}^d \int \tilde{u}_i^2 (D_j\psi_{N+1,1})^2 d\mu\right]^{1/2} \\
&\le \|\psi_{N,2}\|_{L^2}\left[\sum_{i,j=1}^d \left(\int \tilde{u}_i^4 d\mu\right)^{1/2}\left(\int (D_j\psi_{N+1,1})^4 d\mu\right)^{1/2}\right]^{1/2} \\
&\le \sqrt{3}\|\psi_{N,2}\|_{L^2}\|\mathbf{u}\|_{L^2}\|\nabla\psi_{N+1,1}\|_{L^2}.
\end{aligned}$$

In the last inequality we used the fact that, for any centered Gaussian random variable X, its fourth moment $EX^4 = 3(EX^2)^2$.

For $k \geq 2$ we estimate as follows. The first term on the right-hand side of (3.8) equals $(\tilde{\mathbf{u}} \cdot \nabla \psi_{N+1,k}, \psi_{N,k+1})_{L^2}$, and its absolute value can be estimated from above by

$$\|\nabla \psi_{N+1,k}\|_{L^2} \left[\int (|\tilde{\mathbf{u}}||\psi_{N,k+1}|)^2 d\mu \right]^{1/2} \tag{3.9}$$

$$\leq \sqrt{d}\, \|\nabla \psi_{N+1,k}\|_{L^2} \|\tilde{u}\|_{L^{2m}} \|\psi_{N,k+1}\|_{L^{2m/(m-1)}}$$

for an arbitrary $m > 1$. The hypercontractivity property of Gaussian measures on L^p spaces (see [6, Theorem 5.10, p. 62]) allows us to estimate

$$\|\psi_{N,k+1}\|_{L^{2m/(m-1)}} \leq \left(\frac{2m}{m-1} - 1 \right)^{(k+1)/2} \|\psi_{N,k+1}\|_{L^2} \tag{3.10}$$

$$= \left(\frac{m+1}{m-1} \right)^{(k+1)/2} \|\psi_{N,k+1}\|_{L^2}.$$

On the other hand, $\tilde{u}$ is a Gaussian random variable under μ; hence

$$\|\tilde{u}\|_{L^{2m}} = [(2m-1)!!]^{1/(2m)} \|\tilde{u}\|_{L^2}. \tag{3.11}$$

Here $(2m-1)!! := 1 \cdot 3 \cdot \cdots \cdot (2m-1)$. Using Stirling's formula (see, e.g., [5, paragraph 406]), $n! = \sqrt{2\pi n}(ne^{-1})^n e^{\theta/(12n)}$ for some $\theta \in (0,1)$, we conclude from (3.11) that

$$[(2m-1)!!]^{1/(2m)} = \left[\frac{(2m-1)!}{2^{m-1}(m-1)!} \right]^{1/(2m)} \tag{3.12}$$

$$\leq (2e)^{-1/2} (2m-1)(m-1)^{-1/2} 2^{1/(4m)} \left(1 - \frac{1}{2m-1} \right)^{1/(4m)} \exp\left\{ [24m(2m-1)]^{-1} \right\}.$$

Summarizing, from (3.9)–(3.12) we conclude that

$$|(\tilde{\mathbf{u}} \cdot \nabla \psi_{N,k+1}, \psi_{N+1,k})_{L^2}| \leq C\left(\frac{k+1}{2}, m \right) \|\tilde{\mathbf{u}}\|_{L^2} (k+1)^{1/2} \|\psi_{N,k+1}\|_{L^2} \|\nabla \psi_{N+1,k}\|_{L^2}$$

$$\leq \frac{1}{2\kappa} C^2 \left(\frac{k+1}{2}, m \right) \|\tilde{\mathbf{u}}\|_{L^2}^2 (k+1) \|\psi_{N,k+1}\|_{L^2}^2 + \frac{\kappa}{2} \|\nabla \psi_{N+1,k}\|_{L^2}^2. \tag{3.13}$$

Here

$$C(p,m) := (2ep)^{-1/2} (2m-1)(m-1)^{-1/2} \left(\frac{m+1}{m-1} \right)^{p/2} \tag{3.14}$$

$$\times\ 2^{1/(4m)} \left(1 - \frac{1}{2m-1} \right)^{1/(4m)} \exp\left\{ [24m(2m-1)]^{-1} \right\}.$$

Likewise,

$$|(\tilde{\mathbf{u}} \cdot \nabla \psi_{N,k-1}, \psi_{N+1,k})_{L^2}| \tag{3.15}$$

$$\leq \frac{1}{2\kappa} C^2 \left(\frac{k-1}{2}, m' \right) \|\tilde{\mathbf{u}}\|_{L^2}^2 (k-1) \|\psi_{N,k-1}\|_{L^2}^2 + \frac{\kappa}{2} \|\nabla \psi_{N+1,k}\|_{L^2}^2.$$

Note that for $m = 2k + 1$ we have

$$C((k+1)/2, 2k+1) \le 2^{1/12} e^{-1/2+1/360} \left(2 + \frac{1}{2k}\right) \left(1 + \frac{1}{k}\right)^{k/2} < 1.0595 \left(2 + \frac{1}{2k}\right);$$

hence $C^2((k+1)/2, 2k+1) < 6$ for $k \ge 2$. A direct calculation also shows that $C^2(1,3) \approx 5.1620$. Likewise, for $m' = 2k-1$ we have $C^2((k-1)/2, 2k-1) < 6$, $k \ge 2$. Summing up (3.8) over k and using (3.13), (3.16), we conclude therefore that

$$a \sum_{k=1}^{N+1} k \|\psi_{N+1,k}\|_{L^2}^2 + \kappa \|\nabla \psi_{N+1}\|_{L^2}^2 \tag{3.16}$$

$$\le \kappa \|\nabla \psi_{N+1}\|_{L^2}^2 + \frac{6}{\kappa} \|\tilde{\mathbf{u}}\|_{L^2}^2 \sum_{k=1}^{N+1} k \|\psi_{N,k+1}\|_{L^2}^2.$$

Hence, from (3.16) we get

$$a |\!|\!| \psi_{N+1} |\!|\!|^2 \le \frac{6}{\kappa} \|\tilde{\mathbf{u}}\|_{L^2}^2 |\!|\!| \psi_N |\!|\!|^2,$$

and (3.7) follows. □

The solution to (2.14) is given by

$$\chi := \sum_{n=1}^{+\infty} \psi_n. \tag{3.17}$$

The series in (3.17) converges in the L^2 sense provided that $q < 1$; cf. (3.6). Also, as we shall show in Proposition 4.1, we have $(\psi_n, \tilde{u})_{L^2} = 0$ if n is even. From (2.15) we can write therefore that

$$d_* = -\sum_{n=0}^{+\infty} (\psi_{2n+1}, \tilde{u})_{L^2}; \tag{3.18}$$

thus, using (3.7), we conclude the following.

Corollary 3.2.

$$\left| d_* + \sum_{n=0}^{M} (\psi_{2n+1}, \tilde{u})_{L^2} \right| \le q^{2M+2} (1 - q^2)^{-1} \|\psi_1\|_{L^2} \|\tilde{u}\|_{L^2}. \tag{3.19}$$

To calculate $\|\psi_1\|_{L^2}$ we use the spectral representation of $\tilde{u}$,

$$\tilde{u}(\tau_{\mathbf{x}} f) = \int e^{i\mathbf{k}\cdot\mathbf{x}} \hat{u}(d\mathbf{k}; f),$$

where $\hat{u}(\cdot)$ is its spectral measure that satisfies $\hat{u}^*(d\mathbf{k}) = \hat{u}(-d\mathbf{k})$ (because $\tilde{u}$ is real-valued) and

$$\langle \hat{u}^*(d\mathbf{k}) \hat{u}(d\mathbf{k}') \rangle = b\delta(\mathbf{k} - \mathbf{k}') \frac{\mathcal{E}(|\mathbf{k}|)}{|\mathbf{k}|^{d-1}} \hat{\Gamma}_{11}(\mathbf{k}) d\mathbf{k} d\mathbf{k}'.$$

Also, because $\mathbf{x} \mapsto \hat{u}(\tau_{\mathbf{x}} f)$ is a real Gaussian field, we have that (Re $\hat{u}(\cdot)$, Im $\hat{u}(\cdot)$) is jointly Gaussian. Hence,

$$\psi_1 = (\kappa\Delta + D_0)^{-1}\tilde{u} = -\int \frac{\hat{u}(d\mathbf{k})}{\kappa|\mathbf{k}|^2 + a}$$

and

$$\|\psi_1\|_{L^2}^2 = \int\int \frac{\langle \hat{u}^*(d\mathbf{k})\hat{u}(d\mathbf{k}')\rangle}{(\kappa|\mathbf{k}|^2 + a)(\kappa|\mathbf{k}'|^2 + a)} = b\omega_{d-1}\left(1 - \frac{1}{d}\right)\int_0^{K_0} \frac{dk}{(\kappa k^2 + a)^2 k^{2\alpha-1}}. \tag{3.20}$$

4. Calculation of $(\psi_n, \tilde{u})_{L^2}$. We start with some auxiliary notation. For any function $F(t_1, \dots, t_n, \mathbf{x}_1, \dots, \mathbf{x}_n)$ of n temporal and spatial variables we define

$$DF(t_1, \dots, t_{2n}, \mathbf{x}_1, \dots, \mathbf{x}_n) := \nabla_{\mathbf{y}|\,\mathbf{y}=\mathbf{0}} F(t_1, \dots, t_n, \mathbf{x}_1 + \mathbf{y}, \dots, \mathbf{x}_n + \mathbf{y}).$$

$\mathbf{W}_n(\cdot)$ is defined inductively by

$$\mathbf{W}_0(s_1, \mathbf{x}_1) := \mathbf{u}(s_1, \mathbf{x}_1), \tag{4.1}$$

$$\mathbf{W}_n(s_1, \dots, s_{n+1}, \mathbf{x}_1, \dots, \mathbf{x}_{n+1}) := \mathbf{u}(s_{n+1}, \mathbf{x}_{n+1}) \cdot D\mathbf{W}_{n-1}(s_1, \dots, s_n, \mathbf{x}_1, \dots, \mathbf{x}_n). \tag{4.2}$$

Let also $\Delta_n := [(s_1, \dots, s_n) : s_1 \geq \cdots \geq s_n \geq 0]$.

PROPOSITION 4.1. *We have*

$$-(\psi_n, \tilde{u})_{L^2} = \int_{\Delta_n}\dots\int \int_{(\mathbb{R}^d)^n}\dots\int \langle W_{n-1,1}(\mathbf{s}, \mathbf{x}) u_1(0, \mathbf{0})\rangle R_{n-1}(\mathbf{s}, \mathbf{x}) d\mathbf{s} d\mathbf{x}. \tag{4.3}$$

Here $\mathbf{s}$, $\mathbf{x}$ *stand for the abbreviations of the ensemble of variables* $\mathbf{s} = (s_1, \dots, s_n)$, $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, $d\mathbf{s} = ds_1 \dots ds_n$, $d\mathbf{x} = d\mathbf{x}_1 \dots d\mathbf{x}_n$,

$$R_{n-1}(\mathbf{s}, \mathbf{x}) := \prod_{m=1}^{n} r_\kappa(s_m - s_{m+1}, \mathbf{x}_m - \mathbf{x}_{m+1}), \tag{4.4}$$

where $r_\kappa(\cdot, \cdot)$ *is the heat kernel defined in* (2.11) *and* $s_{n+1} := 0$, $\mathbf{x}_{n+1} := \mathbf{0}$.

Proof. To show (4.3) we prove that

$$\psi_n = -\int_{\Delta_n}\dots\int \int_{(\mathbb{R}^d)^n}\dots\int \mathbb{E}_0[W_{n-1,1}(\mathbf{s}, \mathbf{x})] R_{n-1}(\mathbf{s}, \mathbf{x}) d\mathbf{s} d\mathbf{x}, \tag{4.5}$$

where $\mathbb{E}_t$ denotes the conditional expectation w.r.t. σ-algebra $\mathcal{U}_t$ generated by $\mathbf{u}(s, \cdot)$, $s \leq t$. We use an induction argument. For $n = 1$, (4.5) is a consequence of (3.2) and the definition of the semigroup $R(t)$.

Suppose that we have established (4.5) for a certain $n \geq 1$. Note that

$$\psi_{n+1} = -\int_0^{+\infty} R(t) u_n dt. \tag{4.6}$$

Recalling the definition of u_n (see (3.3)), we get

$$u_n = -\int\!\!\dots\!\!\int_{\Delta_n} \int\!\!\dots\!\!\int_{(\mathbb{R}^d)^n} \mathbf{u}(0,\mathbf{0}) \cdot \mathbb{E}_0[DW_{n-1,1}(\mathbf{s},\mathbf{x})] R_{n-1}(\mathbf{s},\mathbf{x}) d\mathbf{s} d\mathbf{x},$$

and in consequence

$$R(t)u_n = -\int\!\!\dots\!\!\int_{\Delta_n} \int\!\!\dots\!\!\int_{(\mathbb{R}^d)^{n+1}} \mathbb{E}_0[\mathbf{u}(t,\mathbf{z}) \cdot DW_{n-1,1}(\mathbf{s}+t,\mathbf{x}+\mathbf{z})] R_{n-1}(\mathbf{s},\mathbf{x}) r_\kappa(t,\mathbf{z}) d\mathbf{s} d\mathbf{x} d\mathbf{z}.$$

Here $\mathbf{s}+t := (s_1+t,\dots,s_n+t)$, $\mathbf{x}+\mathbf{z} := (\mathbf{x}_1+\mathbf{z},\dots,\mathbf{x}_n+\mathbf{z})$. Setting $s'_{n+1} := t$, $s'_i := s_i + t$, $\mathbf{x}'_{n+1} := \mathbf{z}$, $\mathbf{x}'_i := \mathbf{x}_i$, $i = 1,\dots,n$, it is clear from (4.6) and definitions (4.3), (4.4) that formula (4.5) holds for $n+1$. □

Remark 4.2. From (4.3) and elementary properties of Gaussian variables we conclude that $(\psi_{2n}, \tilde{u})_{L^2} = 0 \; \forall\, n$.

To calculate $\langle W_{2n,1}(\mathbf{s},\mathbf{x}) u_1(0,\mathbf{0})\rangle$ appearing in the formula for $(\psi_{2n+1}, \tilde{u})_{L^2}$ (cf. Proposition 4.1), we need to compute the mathematical expectation of a multiple product of Gaussian random variables (cf. [4]). For that purpose it is convenient to use a graphical representation, borrowed from the quantum field theory; see, e.g., [6]. A *Feynman diagram* $\mathcal{G}$ of order $n \geq 0$ and rank $r \geq 0$ is a graph consisting of n vertices made of elements of $\mathbb{Z}_n := \{1,\dots,n\}$ that are positive integers and a set $E(\mathcal{G})$ of $r \leq n/2$ edges without common endpoints. So there are r pairs of vertices, each joined by an edge, and $n-2r$ unpaired vertices, called *free vertices*. An edge whose endpoints are $m, m' \in B$ is represented by $\widehat{mm'}$; we always assume that $m < m'$. Denote the set of all free vertices by $A(\mathcal{G})$. The diagram is *complete* if $A(\mathcal{G})$ is empty and *incomplete* otherwise. Denote by $\mathfrak{F}_n$ the family of all complete Feynman diagrams based on $\mathbb{Z}_n$ (n must then be even). For a given $\mathcal{G} \in \mathfrak{F}_n$ and any $l \leq n$ we denote by $V_l(\mathcal{G})$ all those vertices $m \leq l$ for which there is $n > l$ such that $\widehat{mn} \in E(\mathcal{G})$.

Suppose that $\mathcal{G}$, $\mathcal{G}'$ are Feynman diagrams based on $\mathbb{Z}_n$, $\mathbb{Z}_{n-1}$, respectively. We call $\mathcal{G}'$ *an immediate predecessor* of $\mathcal{G}$ and denote this $\mathcal{G}' \hookrightarrow \mathcal{G}$ if $E(\mathcal{G}') \subseteq E(\mathcal{G})$. Diagram $\mathcal{G}$ is called *admissible* if $A(\mathcal{G}) \neq \emptyset$. For $n \geq 1$ we define a class $\mathfrak{S}_n$ that consists of all sequences $\mathcal{F} := (\mathcal{F}_k)_{k=1}^n$ of Feynman diagrams $\mathcal{F}_1 \hookrightarrow \mathcal{F}_2 \dots \hookrightarrow \mathcal{F}_n$, such that each $\mathcal{F}_k$, based on $\mathbb{Z}_k$, is admissible. For n even we denote by $\mathfrak{S}_n^c$ the class of those sequences $\mathcal{F}_1 \hookrightarrow \mathcal{F}_2 \dots \hookrightarrow \mathcal{F}_n$ for which $(\mathcal{F}_k)_{k=1}^{n-1} \in \mathfrak{S}_{n-1}$ and $\mathcal{F}_n \in \mathfrak{F}_n$. Let $\mathcal{F} \in \mathfrak{S}_n$ or $\mathfrak{S}_n^c$. Denote by $A_k(\mathcal{F}) := A(\mathcal{F}_k)$, $E_k(\mathcal{F}) := E(\mathcal{F}_k)$, $k = 1,2,\dots,n$. We also let $a_k(\mathcal{F})$ denote the cardinality of $A_k(\mathcal{F})$. Let $e_1(\mathcal{F}) := 0$ and $e_{k+1}(\mathcal{F}) := 1/2[a_k(\mathcal{F}) + 1 - a_{k+1}(\mathcal{F})]$, $k = 1,\dots,n-1$.

LEMMA 4.1. *Let $n \geq 0$ and $\mathbf{s} = (s_1,\dots,s_{2n+1}) \in \Delta_{2n+1}$, $\mathbf{x} = (\mathbf{x}_1,\dots,\mathbf{x}_{2n+1})$, $\mathbf{k} = (\mathbf{k}_1,\dots,\mathbf{k}_{2n+1})$. We have then*

$$\langle W_{2n,1}(\mathbf{s},\mathbf{x}) u_1(0,\mathbf{0})\rangle \tag{4.7}$$

$$= (-1)^n b^{n+1} \sum \int\!\!\dots\!\!\int \exp\left\{ \mathrm{i} \sum_{m=1}^{2n+1} \mathbf{k}_m \cdot \mathbf{x}_m \right\} P_{\mathbf{j}}(\mathbf{s},\mathbf{k};\mathcal{F}) Q_{\mathbf{j}}(\mathbf{k};\mathcal{F}_{2n+2}) d\mathbf{k}.$$

The summation extends over all integer-valued multi-indices $\mathbf{j} = (j_1,\dots,j_{2n+2})$ of length $2n+2$, such that $j_1 = j_{2n+2} = 1$ and all sequences $\mathcal{F} \in \mathfrak{S}_{2n+2}^c$. Also, for a given $\mathcal{F}$

$$P_{\mathbf{j}}(\mathbf{s},\mathbf{k};\mathcal{F}) = \exp\{-aa_{2n+1}(\mathcal{F})s_{2n+1}\} \tag{4.8}$$

$$\times \prod_{l=1}^{2n} \left\{ \left(\sum_{m\in A_l(\mathcal{F})} k_{m,j_{l+1}} \right) \exp\{-aa_l(\mathcal{F})(s_l - s_{l+1})\}[1-\exp\{-2a(s_l-s_{l+1})\}]^{e_l(\mathcal{F})} \right\},$$

$$Q_{\mathbf{j}}(\mathbf{k};\mathcal{F}_{2n+2}) = \prod_{\widehat{mm'}\in E(\mathcal{F}_{2n+2})} \frac{\mathcal{E}(|\mathbf{k}_m|)}{|\mathbf{k}_m|^{d-1}} \hat{\Gamma}_{j_m,j_{m'}}(\mathbf{k}_m)\delta(\mathbf{k}_m+\mathbf{k}_{m'}). \tag{4.9}$$

Here, for abbreviation's sake, $d\mathbf{k} := d\mathbf{k}_1 \dots d\mathbf{k}_{2n+2}$.

This lemma follows from an argument analogous to that used in the proof of Lemma 1 of [4]. For the reader's convenience we present it in the appendix below.

Using Proposition 4.1 and Lemma 4.1, contained in the following section we have the following formula.

PROPOSITION 4.3. *We have*

$$(\psi_{2n+1},\tilde{u})_{L^2} = (-1)^{n+1}b^{n+1}(2a)^{-2n-1}\sum \prod_{l=1}^{2n+1} e_l(\mathcal{F})! \int\dots\int \prod_{l=1}^{2n} \left(\sum_{m\in V_l(\mathcal{F}_{2n+2})} k_{m,j_{l+1}} \right) \tag{4.10}$$

$$\times \prod_{l=1}^{2n+1} \left[\prod_{p=0}^{e_l(\mathcal{F})} \left(\frac{1}{2}a_l(\mathcal{F}) + p + \frac{\kappa}{2a} \left| \sum_{m\in V_l(\mathcal{F}_{2n+2})} \mathbf{k}_m \right|^2 \right) \right]^{-1} Q_{\mathbf{j}}(\mathbf{k}_1,\dots,\mathbf{k}_{2n+2};\mathcal{F}_{2n+2})d\mathbf{k}.$$

The summation range is the same as in Lemma 4.1.

Proof. Using the Fourier transform, we can write

$$R_{2n}(\mathbf{s},\mathbf{x}) = \int\dots\int \prod_{m=1}^{2n+1} \exp\left\{-\kappa|\mathbf{q}_m|^2(s_m-s_{m+1}) + \mathrm{i}\mathbf{q}_m\cdot(\mathbf{x}_m-\mathbf{x}_{m+1})\right\} d\mathbf{q},$$

with $d\mathbf{q} = d\mathbf{q}_1\dots d\mathbf{q}_{n+1}$. Applying formula (4.3) to represent $(\psi_{2n+1},\tilde{u})_{L^2}$ and substituting from (4.7) for $\langle W_{2n,1}(\mathbf{s},\mathbf{x})u_1(0,\mathbf{0})\rangle$, we obtain that

$$(\psi_{2n+1},\tilde{u})_{L^2} = (-1)^{n+1}b^{n+1}\sum \int_{\Delta_n}\dots\int\int\dots\int \prod_{m=1}^{2n+1} \exp\left\{-\kappa|\mathbf{q}_m|^2(s_m-s_{m+1})\right\} \tag{4.11}$$

$$\times \prod_{m=1}^{2n+1} \exp\left\{\,\mathrm{i}(\mathbf{k}_m+\mathbf{q}_m-\mathbf{q}_{m-1})\cdot\mathbf{x}_m\right\} P_{\mathbf{j}}(\mathbf{k}_1,\dots,\mathbf{k}_{2n+1};\mathcal{F})$$

$$\times Q_{\mathbf{j}}(\mathbf{k}_1,\dots,\mathbf{k}_{2n+2};\mathcal{F}_{2n+2})d\mathbf{s}d\mathbf{x}d\mathbf{k}d\mathbf{q};$$

here $\mathbf{q}_0 := \mathbf{0}$ and the summation range is the same as in Lemma 4.1. Performing the integration over $\mathbf{x}$ variables yields the expression $\prod_{m=1}^{2n+1}\delta(\mathbf{k}_m+\mathbf{q}_m-\mathbf{q}_{m-1})$, which in turn implies that $\mathbf{q}_1 = -\mathbf{k}_1$, $\mathbf{q}_m = \mathbf{q}_{m-1}-\mathbf{k}_m$. Hence $\mathbf{q}_l = -\sum_{m=1}^{l}\mathbf{k}_m$. Changing variables $s'_m := s_m - s_{m+1}$, $m = 1,\dots,2n+1$ ($s_{2n+2}=0$), and substituting for $P_{\mathbf{j}}(\mathbf{k}_1,\dots,\mathbf{k}_{2n+1};\mathcal{F})$ from (4.8), we get

$$(\psi_{2n+1},\tilde{u})_{L^2}=(-1)^{n+1}b^{n+1}\sum\int\dots\int\prod_{l=1}^{2n}\left(\sum_{m\in A_l(\mathcal{F})}k_{m,j_{l+1}}\right)Q_{\mathbf{j}}(\mathbf{k}_1,\dots,\mathbf{k}_{2n+2};\mathcal{F}_{2n+2})$$

$$\times\prod_{l=1}^{2n+1}\left[\int_0^{+\infty}\exp\left\{-\left(aa_l(\mathcal{F})+\kappa\left|\sum_{m=1}^{l}\mathbf{k}_m\right|^2\right)s_l'\right\}\left(1-e^{-2as_l'}\right)^{e_l(\mathcal{F})}ds_l'\right]d\mathbf{k}.$$

Substituting $t_l:=e^{-2as_l'}$ into the integral w.r.t. s_l', we conclude that

$$(\psi_{2n+1},\tilde{u})_{L^2}=(-1)^{n+1}b^{n+1}(2a)^{-2n-1}\sum\int\dots\int\prod_{l=1}^{2n}\left(\sum_{m\in A_l(\mathcal{F})}k_{m,j_{l+1}}\right)$$
$$\times Q_{\mathbf{j}}(\mathbf{k}_1,\dots,\mathbf{k}_{2n+2};\mathcal{F}_{2n+2})$$

$$\times\prod_{l=1}^{2n+1}B\left(\frac{1}{2}a_l(\mathcal{F})+\frac{\kappa}{2a}\left|\sum_{m=1}^{l}\mathbf{k}_m\right|^2,e_l(\mathcal{F})+1\right)d\mathbf{k}.$$

Here $B(\cdot,\cdot)$ denotes the Euler beta function. Note that for a fixed $l\le 2n+1$ for any $m,m'\le l$ such that $\widehat{mm'}\in E(\mathcal{F}_{2n+2})$ we necessarily have $\mathbf{k}_m+\mathbf{k}_{m'}=\mathbf{0}$; thus the summation range in the sum of $\mathbf{k}$'s reduces to $V_l(\mathcal{F}_{2n+2})$. Using the classical formula

$$B(x,n+1)=n!\left[\prod_{p=0}^{n}(x+p)\right]^{-1}$$

valid for any positive integer n, we conclude (4.10). □

Computations of the terms appearing in (4.10). For $n=0$ the class $\mathfrak{S}_2^c$ consists of a single diagram sequence $1\hookrightarrow\widehat{12}$. From (4.10), after a simple calculation, we obtain

$$(\psi_1,\tilde{u})_{L^2}=-b\int_{|\mathbf{k}|\le K_0}\left(1-\frac{k_1^2}{|\mathbf{k}|^2}\right)\frac{1}{a+\kappa|\mathbf{k}|^2}\times\frac{d\mathbf{k}}{|\mathbf{k}|^{2\alpha+d-2}} \tag{4.12}$$

$$=-\omega_{d-1}b\left(1-\frac{1}{d}\right)\int_0^{K_0}\frac{dk}{(a+\kappa k^2)k^{2\alpha-1}}.$$

The second equality in (4.12) is a consequence of isotropy. For $n=1$ there are only four sequences of diagrams from $\mathfrak{S}_4^c$ for which the corresponding terms of (4.10) are nonvanishing. They are (A) $1\hookrightarrow 12\hookrightarrow 123\hookrightarrow\widehat{14}\,\widehat{23}$, (B) $1\hookrightarrow 12\hookrightarrow 123\hookrightarrow\widehat{13}\,\widehat{24}$, (C) $1\hookrightarrow 12\hookrightarrow\widehat{13}\,2\hookrightarrow\widehat{13}\,\widehat{24}$, and (D) $1\hookrightarrow 12\hookrightarrow\widehat{23}\,1\hookrightarrow\widehat{23}\,\widehat{14}$. The corresponding terms of the sum appearing on the right-hand side of (4.10), after using isotropy, can be calculated as follows:

$$A=b^2\left(1-\frac{1}{d}\right)\int\int_{|\mathbf{k}_1|,|\mathbf{k}_2|\le K_0}|\mathbf{k}_1|^2\left[1-\left(\frac{\mathbf{k}_1\cdot\mathbf{k}_2}{|\mathbf{k}_1||\mathbf{k}_2|}\right)^2\right]$$
$$\times\frac{1}{a+\kappa|\mathbf{k}_1|^2}\times\frac{1}{2a+\kappa|\mathbf{k}_1+\mathbf{k}_2|^2}\times\frac{1}{3a+\kappa|\mathbf{k}_1|^2}\times\frac{d\mathbf{k}_1d\mathbf{k}_2}{(|\mathbf{k}_1||\mathbf{k}_2|)^{2\alpha+d-2}},$$

$$B = \frac{b^2}{d} \int\int_{|\mathbf{k}_1|,|\mathbf{k}_2| \le K_0} \mathbf{k}_1 \cdot \mathbf{k}_2 \left[\left(\frac{\mathbf{k}_1 \cdot \mathbf{k}_2}{|\mathbf{k}_1||\mathbf{k}_2|} \right)^2 - 1 \right] \times \frac{1}{a + \kappa |\mathbf{k}_1|^2} \times \frac{1}{2a + \kappa |\mathbf{k}_1 + \mathbf{k}_2|^2} \times \frac{1}{3a + \kappa |\mathbf{k}_2|^2} \times \frac{d\mathbf{k}_1 d\mathbf{k}_2}{(|\mathbf{k}_1||\mathbf{k}_2|)^{2\alpha + d - 2}},$$

$$C = \frac{2ab^2}{d} \int\int_{|\mathbf{k}_1|,|\mathbf{k}_2| \le K_0} \mathbf{k}_1 \cdot \mathbf{k}_2 \left[\left(\frac{\mathbf{k}_1 \cdot \mathbf{k}_2}{|\mathbf{k}_1||\mathbf{k}_2|} \right)^2 - 1 \right] \times \frac{1}{a + \kappa |\mathbf{k}_1|^2} \times \frac{1}{a + \kappa |\mathbf{k}_2|^2} \times \frac{1}{2a + \kappa |\mathbf{k}_1 + \mathbf{k}_2|^2} \times \frac{1}{3a + \kappa |\mathbf{k}_2|^2} \times \frac{d\mathbf{k}_1 d\mathbf{k}_2}{(|\mathbf{k}_1||\mathbf{k}_2|)^{2\alpha + d - 2}},$$

$$D = 2ab^2 \left(1 - \frac{1}{d} \right) \int\int_{|\mathbf{k}_1|,|\mathbf{k}_2| \le K_0} |\mathbf{k}_1|^2 \left[1 - \left(\frac{\mathbf{k}_1 \cdot \mathbf{k}_2}{|\mathbf{k}_1||\mathbf{k}_2|} \right)^2 \right] \times \frac{1}{(a + \kappa |\mathbf{k}_1|^2)^2} \times \frac{1}{2a + \kappa |\mathbf{k}_1 + \mathbf{k}_2|^2} \times \frac{1}{3a + \kappa |\mathbf{k}_1|^2} \times \frac{d\mathbf{k}_1 d\mathbf{k}_2}{(|\mathbf{k}_1||\mathbf{k}_2|)^{2\alpha + d - 2}}.$$

Adding all these terms, we conclude that

$$(\psi_3, \tilde{u})_{L^2} = \frac{b^2}{d} \int\int_{|\mathbf{k}_1|,|\mathbf{k}_2| \le K_0} \left[1 - \left(\frac{\mathbf{k}_1 \cdot \mathbf{k}_2}{|\mathbf{k}_1||\mathbf{k}_2|} \right)^2 \right] \left[\frac{(d-1)|\mathbf{k}_1|^2}{a + \kappa |\mathbf{k}_1|^2} - \frac{\mathbf{k}_1 \cdot \mathbf{k}_2}{a + \kappa |\mathbf{k}_2|^2} \right] \times \frac{1}{a + \kappa |\mathbf{k}_1|^2} \times \frac{1}{2a + \kappa |\mathbf{k}_1 + \mathbf{k}_2|^2} \times \frac{d\mathbf{k}_1 d\mathbf{k}_2}{(|\mathbf{k}_1||\mathbf{k}_2|)^{2\alpha + d - 2}}.$$

A simple calculation shows that the approximation of the eddy diffusivity obtained by using (3.19), with $M = 1$, has an error bounded by $0.1 d^{-1/2} \|\mathbf{u}\|_{L^2}^{3/2} \|\psi_1\|_{L^2}$ provided that $a\kappa \approx 22$.

When $n \ge 2$ we try to transform formula (4.10) a bit in order to drop the summation over multi-indices $\mathbf{j}$ and replace the summation over sequences of Feynman diagrams by the sum over the complete diagrams of length $2n + 2$. Let us fix then a sequence of Feynman diagrams $\mathcal{F} \in \mathfrak{S}^c_{2n+2}$ and denote by $I(\mathcal{F})$ its corresponding term appearing on the right-hand side of (4.10). Let us consider two cases.

Case 1, *when bond* $\widehat{1, 2n+2} \in E(\mathcal{F}_{2n+2})$. Let $\widehat{m\, m'} \in E(\mathcal{F}_{2n+2})$ be such that $1 < m < m' < 2n + 2$. Performing the summation over the respective multi-indices, we get

$$\sum_{j_m, j_{m'}} \left(\sum_{p \in V_{m-1}(\mathcal{F}_{2n+2})} k_{p, j_m} \right) \left(\sum_{p' \in V_{m'-1}(\mathcal{F}_{2n+2})} k_{p', j_{m'}} \right) \Gamma_{j_m, j'_m}(\mathbf{k}_m)$$

$$= \sum_{\substack{p \in V_{m-1}(\mathcal{F}_{2n+2}) \\ p' \in V_{m'-1}(\mathcal{F}_{2n+2})}} \left(\mathbf{k}_p \cdot \mathbf{k}_{p'} - \frac{(\mathbf{k}_p \cdot \mathbf{k}_m)(\mathbf{k}_{p'} \cdot \mathbf{k}_m)}{|\mathbf{k}_m|^2} \right).$$

Denote by $\widehat{m(p), m'(p)}$, $p = 1, \dots, n+1$, all the edges of $E(\mathcal{F}_{2n+2})$ whose left endpoints are enumerated in increasing order. We can write then that

(4.13)
$$I(\mathcal{F}) = \frac{(-b)^{n+1}}{(2a)^{2n+1}} \prod_{l=1}^{2n+1} e_l(\mathcal{F})! \int\limits_{|\mathbf{k}_1|,\dots,|\mathbf{k}_{n+1}|\le K_0} \dots \int \left(1 - \frac{k_{1,1}^2}{|\mathbf{k}_1|^2}\right) \mathcal{J}(\mathbf{k}_1, \dots, \mathbf{k}_{n+1}; \mathcal{F}_{2n+2})$$
$$\times \prod_{l=1}^{2n+1} \left[\prod_{p=0}^{e_l(\mathcal{F})} \left(\frac{1}{2} a_l(\mathcal{F}) + p + \frac{\kappa}{2a} \left| \sum_{m(q)\in V_l(\mathcal{F}_{2n+2})} \mathbf{k}_q \right|^2 \right) \right]^{-1} \frac{d\mathbf{k}_1 \dots d\mathbf{k}_{n+1}}{(|\mathbf{k}_1| \dots |\mathbf{k}_{n+1}|)^{2\alpha+d-2}}.$$

Here for a given $\mathcal{G} \in \mathfrak{F}_{2n+2}$ such that $\widehat{1, 2n+2} \in E(\mathcal{G})$, we set

$$\mathcal{J}(\mathbf{k}_1, \dots, \mathbf{k}_{n+1}; \mathcal{G}) := \prod_{l=1}^{n+1} \sum_{\substack{m(p)\in V_{m(l)-1}(\mathcal{F}_{2n+2}) \\ m(p')\in V_{m'(l)-1}(\mathcal{F}_{2n+2})}} \left(\mathbf{k}_p \cdot \mathbf{k}_{p'} - \frac{(\mathbf{k}_p \cdot \mathbf{k}_l)(\mathbf{k}_{p'} \cdot \mathbf{k}_l)}{|\mathbf{k}_l|^2} \right). \tag{4.14}$$

As above we let $1 = m(1) < \dots < m(n+1)$ denote the left vertices of all the edges from $E(\mathcal{G})$.

Using isotropy, we can further simplify the formula and obtain that

$$I(\mathcal{F}) = \left(1 - \frac{1}{d}\right) \frac{(-b)^{n+1}}{(2a)^{2n+1}} \prod_{l=1}^{2n+1} e_l(\mathcal{F})! \int\limits_{|\mathbf{k}_1|,\dots,|\mathbf{k}_{n+1}|\le K_0} \dots \int \mathcal{J}(\mathbf{k}_1, \dots, \mathbf{k}_{n+1}; \mathcal{F}_{2n+2})$$
$$\times \prod_{l=1}^{2n+1} \left[\prod_{p=0}^{e_l(\mathcal{F})} \left(\frac{1}{2} a_l(\mathcal{F}) + p + \frac{\kappa}{2a} \left| \sum_{m(q)\in V_l(\mathcal{F}_{2n+2})} \mathbf{k}_q \right|^2 \right) \right]^{-1} \frac{d\mathbf{k}_1 \dots d\mathbf{k}_{n+1}}{(|\mathbf{k}_1| \dots |\mathbf{k}_{n+1}|)^{2\alpha+d-2}}.$$

Case 2, when bond $\widehat{1, 2n+2} \notin E(\mathcal{F}_{2n+2})$. Let $q > 1$ be such that $\widehat{m(q), 2n+2} \in E(\mathcal{F}_{2n+2})$. After a calculation similar to the one in the previous case, we obtain that

$$I(\mathcal{F}) = \frac{(-b)^{n+1}}{d(2a)^{2n+1}} \prod_{l=1}^{2n+1} e_l(\mathcal{F})! \int\limits_{|\mathbf{k}_1|,\dots,|\mathbf{k}_{n+1}|\le K_0} \dots \int \mathcal{J}(\mathbf{k}_1, \dots, \mathbf{k}_{n+1}; \mathcal{F}_{2n+2})$$
$$\times \prod_{l=1}^{2n+1} \left[\prod_{p=0}^{e_l(\mathcal{F})} \left(\frac{1}{2} a_l(\mathcal{F}) + p + \frac{\kappa}{2a} \left| \sum_{m(q)\in V_l(\mathcal{F}_{2n+2})} \mathbf{k}_q \right|^2 \right) \right]^{-1} \frac{d\mathbf{k}_1 \dots d\mathbf{k}_{n+1}}{(|\mathbf{k}_1| \dots |\mathbf{k}_{n+1}|)^{2\alpha+d-2}}.$$

Here for a given $\mathcal{G} \in \mathfrak{F}_{2n+2}$ such that $\widehat{m(q), 2n+2} \in E(\mathcal{G})$ for some $q > 1$, we set

$$\mathcal{J}(\mathbf{k}_1, \dots, \mathbf{k}_{n+1}; \mathcal{G}) := \prod_{l\notin\{1,q\}} \sum_{\substack{m(p)\in V_{m(l)-1}(\mathcal{G}) \\ m(p')\in V_{m'(l)-1}(\mathcal{G})}} \left(\mathbf{k}_p \cdot \mathbf{k}_{p'} - \frac{(\mathbf{k}_p \cdot \mathbf{k}_l)(\mathbf{k}_{p'} \cdot \mathbf{k}_l)}{|\mathbf{k}_l|^2} \right) \tag{4.15}$$

$$\times \sum_{\substack{m(p)\in V_{m'(1)-1}(\mathcal{G})\\ m(p')\in V_{m(q)-1}(\mathcal{G})}} \Bigg[\mathbf{k}_p\cdot\mathbf{k}_{p'} + \frac{(\mathbf{k}_q\cdot\mathbf{k}_{p'})(\mathbf{k}_1\cdot\mathbf{k}_p)(\mathbf{k}_1\cdot\mathbf{k}_q)}{(|\mathbf{k}_1||\mathbf{k}_q|)^2} - \frac{(\mathbf{k}_p\cdot\mathbf{k}_q)(\mathbf{k}_{p'}\cdot\mathbf{k}_q)}{|\mathbf{k}_q|^2} - \frac{(\mathbf{k}_p\cdot\mathbf{k}_1)(\mathbf{k}_{p'}\cdot\mathbf{k}_1)}{|\mathbf{k}_1|^2}\Bigg].$$

Fix a Feynman diagram $\mathcal{G}\in\mathfrak{F}_{2n+2}$ and a sequence $(e_l)_{l=1}^{2n+2}$ such that $e_l\ge 0$ and $\sum e_l = n+1$. Let $\mathcal{F}\in\mathfrak{S}^c_{2n+2}$ be such that

$$e_l(\mathcal{F}) = e_l,\quad l\in\{1,\dots,2n+1\},\ \text{and}\ \mathcal{F}_{2n+2}=\mathcal{G}. \tag{4.16}$$

Since $\mathcal{F}$ is admissible, $(e_l)_{l=1}^{2n+1}$ must belong to the set $\mathcal{S}_n$ of all sequences satisfying $e_1+\dots+e_{2n+1}\le n$, $e_1+\dots+e_{2p}\le p-1$ and $e_1+\dots+e_{2p-1}\le p-1$, $p\in\{1,\dots,n\}$. Let a_l be given by $a_1=1$ and $a_{l+1}:=a_l+1-2e_l$, $l\ge 1$. With a given $(e_l)_{l=1}^{2n+1}$ and $\mathcal{G}$ let us calculate the number of $\mathcal{F}$'s satisfying (4.16). The diagram $\mathcal{F}_{2n+1}$ may be obtained from $\mathcal{G}$ by removing the bond containing $2n+2$ and $e_{2n+2}-1$ out of the remaining n bonds. There are $\binom{n}{e_{2n+2}-1}$ ways of doing that. Out of the remaining $n+1-e_{2n+2}$ bonds we can remove e_{2n+1} in $\binom{n+1-e_{2n+2}}{e_{2n+1}}$ ways, etc. One can see therefore that the number of sequences $\mathcal{F}$ corresponding to a given $\mathcal{G}\in\mathfrak{F}_{2n+2}$ and $(e_l)_{l=1}^{2n+1}\in\mathcal{S}_n$ equals $n!/[e_1!\dots e_{2n+1}!(n-e_1-\dots-e_{2n+1})!]$. Summarizing the above, we can rewrite (4.10) in the form

$$(\psi_{2n+1},\tilde{u})_{L^2} = (-1)^{n+1}b^{n+1}(2a)^{-2n-1}\sum\frac{n!}{(n-\sum_{l=1}^{2n+1}e_l)!}\int\dots\int \mathcal{J}(\mathbf{k}_1,\dots,\mathbf{k}_{n+1};\mathcal{G}) \tag{4.17}$$

$$\times\prod_{l=1}^{2n+1}\left[\prod_{p=0}^{e_l(\mathcal{F})}\left(\frac{1}{2}a_l+p+\frac{\kappa}{2a}\left|\sum_{m(q)\in V_l(\mathcal{G})}\mathbf{k}_q\right|^2\right)\right]^{-1}\frac{d\mathbf{k}_1\dots d\mathbf{k}_{n+1}}{(|\mathbf{k}_1|\dots|\mathbf{k}_{n+1}|)^{2\alpha+d-2}}.$$

The summation extends over all $\mathcal{G}\in\mathfrak{F}_{2n+2}$ and sequences $(e_l)_{l=1}^{2n+1}\in\mathcal{S}_n$. The cardinality D_{2n-1} of the set $\mathcal{S}_n$ can be calculated using the counting method presented in [11]; see, in particular, Lemma 1 on p. 16. We have the following recursive formula, $D_0:=1$, and

$$D_k=\sum_{i=0}^{k-1}(-1)^i\binom{[(k-i+1)/2]+k}{i+1}D_{k-i-1},\quad k=1,\dots,2n-1. \tag{4.18}$$

The number of terms that may, in principle, appear in the sum on the left-hand side of (4.17) equals, therefore, $(2n+1)!!D_{2n-1}$; here $(2n+1)!!$ stands for the number of complete Feynman diagrams of length $2n+2$. Note, however, that in fact the number of nonvanishing terms is necessarily smaller than the one given above. All terms corresponding to Feynman diagrams for which at least one of the sets $V_l(\mathcal{G})$, $l=1,\dots,2n$, is empty are equal to 0 (cf. (4.10)). Any $\mathcal{G}$ having that property shall be called a null diagram.

For example, when $n=2$, one calculates from (4.18) that $D_3=7$. Out of 15 Feynman diagrams belonging to $\mathfrak{F}_6$, all three diagrams containing bond $\widehat{5\,6}$ are null

diagrams. In addition, one can see that $\widehat{1\,2}\,\widehat{3\,5}\,\widehat{4\,6}$, $\widehat{1\,2}\,\widehat{3\,6}\,\widehat{4\,5}$ are also null. The sum in (4.17) involves therefore at most $7 \times 10 = 70$ nonvanishing terms.

An estimate of D_{2n-1} for a general n can also be obtained from [11]. Namely, using the counting argument contained on p. 14 and the upper bound (1.41) of [11], one can see that $D_{2n-1} \le 5\binom{3n+1}{2n-1}/(3n+1)$.

Appendix. The proof of Lemma 4.1. Using the Fourier transform in the spatial variable, we can write that

$$\mathbf{u}(t,\mathbf{x}) = \int e^{\mathrm{i}\mathbf{x}\cdot\mathbf{k}}\hat{\mathbf{u}}(t,d\mathbf{k}), \tag{A.1}$$

where $\hat{\mathbf{u}}(t,d\mathbf{k}) = (\hat{u}_1(t,d\mathbf{k}),\dots,\hat{u}_d(t,d\mathbf{k}))$ is a real-valued Gaussian spectral measure with the structure function

$$\langle \hat{\mathbf{u}}^*(t,d\mathbf{k}) \otimes \hat{\mathbf{u}}(s,d\mathbf{k}')\rangle = be^{-a|t-s|}\,\frac{\mathcal{E}(|\mathbf{k}|)}{|\mathbf{k}|^{d-1}}\,\hat{\Gamma}(\mathbf{k})\delta(\mathbf{k}-\mathbf{k}')\,d\mathbf{k}d\mathbf{k}' \tag{A.2}$$

satisfying $\hat{\mathbf{u}}^*(t,d\mathbf{k}) = \hat{\mathbf{u}}(t,-d\mathbf{k})$.

To show the lemma it suffices only to prove that for arbitrary $\mathbf{s} = (s_1,\dots,s_{n+1}) \in \Delta_{n+1}$, $s_{n+2} \le s_{n+1}$, we have

$$\mathbb{E}_{s_{n+2}} W_{n,1}(\mathbf{s},\mathbf{x}) \tag{A.3}$$

$$= \mathrm{i}^n \sum \int\dots\int b^{f_{n+1}(\mathcal{F})} \exp\left\{\mathrm{i}\sum_{m=1}^{n+1}\mathbf{k}_m\cdot\mathbf{x}_m\right\} \widehat{P}_{\mathbf{j}}(\mathbf{s},\mathbf{k};s_{n+2},\mathcal{F})\widehat{Q}_{\mathbf{j}}(d\mathbf{k};s_{n+2},\mathcal{F}).$$

Here $f_m(\mathcal{F}) = \sum_{l\le m} e_l(\mathcal{F})$ denotes the cardinality of $E(\mathcal{F}_m)$. The summation extends over all integer-valued multi-indices $\mathbf{j} = (j_1,\dots,j_{n+1})$, such that $j_1 = 1$ and all Feynman diagrams $\mathcal{F} \in \mathfrak{S}_{n+1}$. As we recall, $\mathbb{E}_t$ is the conditional expectation w.r.t. the σ-algebra $\mathcal{U}_t$ generated by $\mathbf{u}(s,\cdot)$, $s \le t$, and

$$\widehat{P}_{\mathbf{j}}(\mathbf{s},\mathbf{k};s_{n+2},\mathcal{F}) := \exp\{-aa_{n+1}(\mathcal{F})(s_{n+1}-s_{n+2})\} \tag{A.4}$$

$$\times\prod_{l=1}^{n}\left\{\left(\sum_{m\in A_l(\mathcal{F})} k_{m,j_{l+1}}\right)\exp\{-aa_l(\mathcal{F})(s_l-s_{l+1})\}[1-\exp\{-2a(s_l-s_{l+1})\}]^{e_l(\mathcal{F})}\right\},$$

$$\widehat{Q}_{\mathbf{j}}(d\mathbf{k};s_{n+2},\mathcal{F}) \tag{A.5}$$

$$:= \prod_{\widehat{mm'}\in E_{n+1}(\mathcal{F})} \frac{\mathcal{E}(|\mathbf{k}_m|)}{|\mathbf{k}|^{d-1}}\hat{\Gamma}_{j_m,j_{m'}}(\mathbf{k}_m)\delta(\mathbf{k}_m+\mathbf{k}_{m'})d\mathbf{k}_m d\mathbf{k}_{m'} \prod_{m\in A_{n+1}(\mathcal{F})}\hat{u}_{j_m}(s_{n+2},d\mathbf{k}_m).$$

We use the induction argument. Formula (A.3) obviously holds for $n = 0$. Suppose that it holds for a certain n. Then

$$\mathbb{E}_{s_{n+3}} W_{n+1,1}(\mathbf{s},\mathbf{x}) = \mathbb{E}_{s_{n+3}}\left[\mathbf{u}(s_{n+2},\mathbf{x}_{n+2})\cdot D\mathbb{E}_{s_{n+2}} W_{n,1}(s_1,\dots,s_{n+1},\mathbf{x}_1,\dots,\mathbf{x}_{n+1})\right].$$

Calculate $\mathbb{E}_{s_{n+2}}\mathbf{W}_n(\cdot)$ using (A.3). Note that

$$\mathbb{E}_{s_{n+3}} W_{n+1,1}(\mathbf{s},\mathbf{x}) = \mathrm{i}^n \sum \int\dots\int b^{f_{n+1}(\mathcal{F})} \exp\left\{\mathrm{i}\sum_{m=1}^{n+2}\mathbf{k}_m\cdot\mathbf{x}_m\right\} \widehat{P}_{\mathbf{j}}(\mathbf{s},\mathbf{k};s_{n+2},\mathcal{F}) \tag{A.6}$$

$$\times\, \mathbb{E}_{s_{n+3}}\left[\hat{u}_{j_{n+2}}(s_{n+2},\mathbf{k}_{n+2})\frac{\partial}{\partial y_{j_{n+2}}}_{|\,\mathbf{y}=\mathbf{0}}\exp\left\{\,\mathrm{i}\sum_{m=1}^{n+1}\mathbf{k}_m\cdot\mathbf{y}\right\}\widehat{Q}_{\mathbf{j}}(d\mathbf{k};s_{n+2},\mathcal{F})\right].$$

The summation extends over all integer-valued multi-indices $\mathbf{j}=(j_1,\dots,j_{n+2})$, such that $j_1=1$ and all Feynman diagrams $\mathcal{F}\in\mathfrak{S}_{n+1}$. Differentiating w.r.t. $y_{j_{n+2}}$ and using the identification rule $\mathbf{k}_m=-\mathbf{k}_{m'}$, when $\widehat{mm'}\in E_{n+1}(\mathcal{F})$, we obtain that the right-hand side of (A.6) equals

$$\text{(A.7)}\qquad \mathrm{i}^{n+1}\sum\int\dots\int b^{f_{n+1}(\mathcal{F})}\exp\left\{\,\mathrm{i}\sum_{m=1}^{n+2}\mathbf{k}_m\cdot\mathbf{x}_m\right\}\widehat{P}_{\mathbf{j}}(\mathbf{s},\mathbf{k};s_{n+2},\mathcal{F})$$

$$\times\left(\sum_{m\in A_{n+1}(\mathcal{F})}k_{m,j_{n+2}}\right)\mathbb{E}_{s_{n+3}}\left[\hat{u}_{j_{n+2}}(s_{n+2},\mathbf{k}_{n+2})\widehat{Q}_{\mathbf{j}}(d\mathbf{k};s_{n+2},\mathcal{F})\right].$$

From elementary properties of Gaussian variables we conclude that

$$\hat{u}_j(s_{n+2},d\mathbf{k};s_{n+3}):=\mathbb{E}_{s_{n+3}}\hat{u}_j(s_{n+2},d\mathbf{k})=e^{-a(s_{n+2}-s_{n+3})}\hat{u}_j(s_{n+3},d\mathbf{k}).$$

The spectral measure

$$\hat{u}_j^{\perp}(s_{n+2},d\mathbf{k};s_{n+3}):=\hat{u}_j(s_{n+2},d\mathbf{k})-\mathbb{E}_{s_{n+3}}\hat{u}_j(s_{n+2},d\mathbf{k})$$

is independent of the σ-algebra $\mathcal{U}_{s_{n+2}}$. Moreover, note that

$$\langle\hat{u}_j^{\perp}(s_{n+2},d\mathbf{k};s_{n+3})\hat{u}_{j'}^{\perp}(s_{n+2},d\mathbf{k}';s_{n+3})\rangle=\left[1-e^{-2a(s_{n+2}-s_{n+3})}\right]\langle\hat{u}_j(0,d\mathbf{k})\hat{u}_{j'}(0,d\mathbf{k}')\rangle.$$

Replace each $\hat{u}_{j_m}(s_{n+2},d\mathbf{k}_m)$ appearing in (A.7), i.e., $\hat{u}_{j_{n+2}}(s_{n+2},\mathbf{k}_{n+2})$ and the spectral measures that occur in $\widehat{Q}_{\mathbf{j}}(d\mathbf{k};s_{n+2},\mathcal{F})$, by

$$e^{-a(s_{n+2}-s_{n+3})}\hat{u}_{j_m}(s_{n+3},d\mathbf{k}_m)+\hat{u}_{j_m}^{\perp}(s_{n+2},d\mathbf{k}_m;s_{n+3}).$$

The conditional expectation $\mathbb{E}_{s_{n+3}}$ can be expressed using the expectation of the product of the terms in the form $\hat{u}_{j_m}^{\perp}(s_{n+2},d\mathbf{k}_m;s_{n+3})$ times the product of $e^{-a(s_{n+2}-s_{n+3})}\times\hat{u}_{j_m}(s_{n+3},d\mathbf{k}_m)$. In order to finish the induction argument we apply the rules of calculating the expectation of products of Gaussian random variables using Feynman diagrams. To prove that out of the diagrams generated in that way we need only to take into account those belonging to $\mathfrak{S}_{n+2}$, it suffices only to show that

$$\text{(A.8)}\qquad \mathrm{i}^{n+1}\sum\int\dots\int b^{f_{n+1}(\mathcal{F})}\exp\left\{\,\mathrm{i}\sum_{m=1}^{n+2}\mathbf{k}_m\cdot\mathbf{x}_m\right\}\widehat{P}_{\mathbf{j}}(\mathbf{s},\mathbf{k};s_{n+2},\mathcal{F})$$

$$\times\left(\sum_{m\in A_{n+1}(\mathcal{F})}k_{m,j_{n+2}}\right)\left\langle\hat{u}_{j_{n+2}}^{\perp}(s_{n+2},\mathbf{k}_{n+2};s_{n+3})\widehat{Q}_{\mathbf{j}}^{\perp}(d\mathbf{k};s_{n+2},s_{n+3},\mathcal{F})\right\rangle=0.$$

Here

$$\widehat{Q}_{\mathbf{j}}^{\perp}(d\mathbf{k};s_{n+2},s_{n+3},\mathcal{F})$$

$$:=\prod_{\widehat{mm'}\in E_{n+1}(\mathcal{F})}\frac{\mathcal{E}(|\mathbf{k}_m|)}{|\mathbf{k}_m|^{d-1}}\hat{\Gamma}_{j_m,j_{m'}}(\mathbf{k}_m)\delta(\mathbf{k}_m+\mathbf{k}_{m'})d\mathbf{k}_m d\mathbf{k}_{m'}\prod_{m\in A_{n+1}(\mathcal{F})}\hat{u}_{j_m}^{\perp}(s_{n+2},d\mathbf{k}_m;s_{n+3}).$$

To do so note that since $\sum_{j_{n+2}} k_{n+2,j_{n+2}} \hat{u}^{\perp}_{j_{n+2}}(s_{n+2}, d\mathbf{k}_{n+2}; s_{n+3}) = 0$, the expression appearing on the left-hand side of (A.8) equals

$$\text{(A.9)} \qquad \mathrm{i}^n \sum b^{f_{n+1}(\mathcal{F})} \widehat{P}_{\mathbf{j}}(\mathbf{s}, \mathbf{k}; s_{n+2}, \mathcal{F}) \left\langle \hat{u}^{\perp}_{j_{n+2}}(s_{n+2}, d\mathbf{k}_{n+2}; s_{n+3}) \right.$$

$$\left. \times \frac{\partial}{\partial y_{j_{n+2}}}{}_{|\mathbf{y}=0} \left[\int \dots \int \exp\left\{ \mathrm{i} \sum_{m=1}^{n+2} \mathbf{k}_m \cdot (\mathbf{x}_m + \mathbf{y}) \right\} \widehat{Q}^{\perp}_{\mathbf{j}}(d\mathbf{k}; s_{n+2}, s_{n+3}, \mathcal{F}) \right] \right\rangle$$

$$= \mathrm{i}^n \sum b^{f_{n+1}(\mathcal{F})} \widehat{P}_{\mathbf{j}}(\mathbf{s}, \mathbf{k}; s_{n+2}, \mathcal{F}) \frac{\partial}{\partial y_{j_{n+2}}}{}_{|\mathbf{y}=0} \left\langle \int \dots \int \exp\left\{ \mathrm{i} \sum_{m=1}^{n+2} \mathbf{k}_m \cdot \mathbf{x}_m \right\} \right.$$

$$\left. \times \hat{u}^{\perp}_{j_{n+2}}(s_{n+2}, d\mathbf{k}_{n+2}; s_{n+3}) \widehat{Q}^{\perp}_{\mathbf{j}}(d\mathbf{k}; s_{n+2}, s_{n+3}, \mathcal{F}) \right\rangle = 0.$$

The last equality in (A.9) is a consequence of spatial stationarity of the field.

REFERENCES

[1] A. Fannjiang and T. Komorowski, *A Martingale approach to homogenization of unbounded random flows,* Ann. Probab., 25 (1997), pp. 1872–1894.

[2] A. Fannjiang and T. Komorowski, *An invariance principle for diffusion in turbulence,* Ann. Probab., 27 (1999), pp. 751–781.

[3] A. Fannjiang and T. Komorowski, *Turbulent diffusion in Markovian flows,* Ann. Appl. Probab., 9 (1999), pp. 591–610.

[4] A. Fannjiang and T. Komorowski, *Frozen path approximation for turbulent diffusion and fractional Brownian motion in random flows*, SIAM J. Appl. Math., 63 (2003), pp. 2042–2062.

[5] G. M. Fichtenholtz, *A Course of Differential and Integral Calculus, Vol.* 2, Fizmatgiz, Moscow, 1959 (in Russian).

[6] S. Janson, *Gaussian Hilbert Spaces*, Cambridge Tracts in Math. 129, Cambridge University Press, Cambridge, UK, 1997.

[7] T. Komorowski, *Diffusion approximation for the convection diffusion equation with random drift.*, Probab. Theory Related Fields, 121 (2001), pp. 525–550.

[8] T. Komorowski and S. Olla, *On the superdiffusive behavior of passive tracer with a Gaussian drift*, J. Statist. Phys., 108 (2002), pp. 647–668.

[9] T. Komorowski and S. Olla, *On the sector condition and homogenization of diffusions with a Gaussian drift*, J. Funct. Anal., 197 (2003), pp. 179–211.

[10] R. H. Kraichnan, *Small-scale structure of a scalar field convected by turbulence*, Phys. Fluids, 11 (1968), pp. 945–953.

[11] S. G. Mohanty, *Lattice Path Counting and Applications*, Academic Press, New York, 1979.

[12] P. K. Pathak and C. Qualls, *A law of iterated logarithm for stationary Gaussian processes*, Trans. Amer. Math. Soc., 181 (1973), pp. 185–193.

[13] V. V. Zhikov, S. M. Kozlov, and O. A. Olejnik, *Averaging of parabolic operators*, Trudy Moskov. Mat. Obshch., 45 (1982), pp. 182–236 (in Russian).

SIAM J. APPL. MATH.
Vol. 65, No. 1, pp. 113–130

APPLICATION OF SYMMETRY ANALYSIS TO A PDE ARISING IN THE CAR WINDSHIELD DESIGN*

NICOLETA BÎLĂ†

Abstract. A new approach to parameter identification problems from the point of view of symmetry analysis theory is given. A mathematical model that arises in the design of car windshield represented by a linear second order mixed type PDE is considered. Following a particular case of the direct method (due to Clarkson and Kruskal), we introduce a method to study the group invariance between the parameter and the data. The equivalence transformations associated with this inverse problem are also found. As a consequence, the symmetry reductions relate the inverse and the direct problem and lead us to a reduced order model.

Key words. symmetry reductions, parameter identification problems

AMS subject classifications. 58J70, 70G65, 35R30, 35R35

DOI. 10.1137/S0036139903434031

1. Introduction.

Symmetry analysis theory links differential geometry to PDEs theory [18], symbolic computation [9], and, more recently, to numerical analysis theory [3], [6]. The notion of continuous transformation groups was introduced by Sophus Lie [14], who also applied them to differential equations. Over the years, Lie's method has been proven to be a powerful tool for studying a remarkable number of PDEs arising in mathematical physics (more details can be found for example in [2], [10], and [21]). In the last several years a variety of methods have been developed in order to find special classes of solutions of PDEs, which cannot be determined by applying the classical Lie method. Olver and Rosenau [20] showed that the common theme of all these methods has been the appearance of some form of group invariance. On the other hand, parameter identification problems arising in the inverse problems theory are concerned with the identification of physical parameters from observations of the evolution of a system. In general, these are ill-posed problems, in the sense that they do not fulfill Hadamard's postulates for all admissible data: a solution exists, the solution is unique, and the solution depends continuously on the given data. Arbitrary small changes in data may lead to arbitrary large changes in the solution. The iterative approach of studying parameter identification problems is a functional-analytic setup with a special emphasis on iterative regularization methods [8].

The aim of this paper is to show how parameter identification problems can be analyzed with the tools of group analysis theory. This is a new direction of research in the theory of inverse problems, although the symmetry analysis theory is a common approach for studying PDEs. We restrict ourselves to the case of a parameter identification problem modeled by a PDE of the form

$$F(x, w^{(m)}, E^{(n)}) = 0, \tag{1.1}$$

where the unknown function $E = E(x)$ is called *parameter*, and, respectively, the arbitrary function $w = w(x)$ is called *data*, with $x = (x_1, \ldots, x_p) \in \Omega \subset R^p$ a given

*Received by the editors September 4, 2003; accepted for publication (in revised form) May 4, 2004; published electronically September 24, 2004. This work was supported by the Austrian Science Foundation FWF, Project SFB 1308 "Large Scale Inverse Problems."

http://www.siam.org/journals/siap/65-1/43403.html

†Institute for Industrial Mathematics, Johannes Kepler University, 69 Altenbergerstrasse, Linz, A-4040, Austria (bila@indmath.uni-linz.ac.at).

domain (here $w^{(m)}$ denotes the function w together with its partial derivatives up to order m). Assume that the parameters and the data are analytical functions. The PDE (1.1) sometimes augmented with certain boundary conditions is called *the inverse problem* associated with a *direct problem*. The direct problem is the same equation but the unknown function is the data, for which certain boundary conditions are required.

The *classical Lie method* allows us to find the *symmetry group* related to a PDE. This is a (local) Lie group of transformations acting on the space of the independent variables and the space of the dependent variables of the equation with the property that it leaves the set of all analytical solutions invariant. Knowledge of these *classical symmetries* allows us to reduce the order of the studied PDE and to determine *group-invariant solutions* (or *similarity solutions*) which are invariant under certain subgroups of the full symmetry group (for more details see [18]). Bluman and Cole [1] introduced the *nonclassical method* that allows one to find the *conditional symmetries* (also called *nonclassical symmetries*) associated with a PDE. These are transformations that leave only a subset of the set of all analytical solutions invariant. Note that any classical symmetry is a nonclassical symmetry but not conversely. Another procedure for finding symmetry reductions is the *direct method* (due to Clarkson and Kruskal [5]). The relation between these last two methods has been studied by Olver [19]. Moreover, for a PDE with coefficients depending on an arbitrary function, Ovsiannikov [21] introduced the notion of *equivalence transformations*, which are (local) Lie group of transformations acting on the space of the independent variables, the space of the dependent variables and the space of the arbitrary functions that leave the equation unchanged. Notice that these techniques based on group theory do not take into account the boundary conditions attached to a PDE.

To find symmetry reductions associated with the parameter identification problem (1.1) one can seek classical and nonclassical symmetries related to this equation. Two cases can occur when applying the classical Lie method or the nonclassical method, depending if the data w is known or not. From the symbolic computation point of view, the task of finding symmetry reductions for a PDE depending on an arbitrary function might be a difficult one, due to the lack of the symbolic manipulation programs that can handle these kind of equations. Another method to determine symmetry reductions for (1.1) might be a particular case of the direct method, which has been applied by Zhdanov [24] to certain multidimensional PDEs arising in mathematical physics. Based on this method and taking into account that (1.1) depends on an arbitrary function, we introduce a procedure to find the relation between the data and the parameter in terms of a similarity variable (see section 2). As a consequence, the equivalence transformations related to (1.1) must be considered as well. These final symmetry reductions are found by using any symbolic manipulation program designed to determine classical symmetries for a PDE system—now both the data and the parameter are unknown functions in (1.1). The equivalence transformations relate the direct problem and the inverse problem. Moreover, one can find special classes of data and parameters, respectively, written in terms of the invariants of the group action, the order of the studied PDE can be reduced at least by one, and analytical solutions of (1.1) can be found.

At the first step, the group approach of the free boundary problem related to (1.1) can be considered and, afterwards, the invariance of the boundary conditions under particular group actions has to be analyzed (see [2]). In the case of parameter identification problems we sometimes have to deal with two pairs of boundary conditions, for data and the parameter as well, otherwise we might only know the boundary

conditions for the data. Thus, the problem of finding symmetry reductions for a given data can be more complicated. At least by finding the equivalence transformations related to the problem, the invariants of the group actions can be used to establish suitable domains Ω on which the order of the model can be reduced.

In this paper we consider a mathematical model arising in the car windshield design. Let us briefly explain the *gravity sag bending process*, one of the main industrial processes used in the manufacture of car windshields. A piece of glass is placed over a rigid frame, with the desired edge curvature and heated from below. The glass becomes viscous due to the temperature rise and sags under its own weight. The final shape depends on the viscosity distribution of the glass obtained from varying the temperature. It has been shown that the sag bending process can also be controlled (in a first approximation) in the terms of Young's modulus E, a spatially varying glass material parameter, and the displacement of the glass w can be described by the thin linear elastic plate theory (see [11], [16], and [17] and references from there). The model is based on the linear plate equation

(1.2)
$$\left(E\left(w_{xx}+\nu w_{yy}\right)\right)_{xx}+2(1-\nu)\left(Ew_{xy}\right)_{xy}+\left(E\left(w_{yy}+\nu w_{xx}\right)\right)_{yy}=\tfrac{12(1-\nu^2)f}{h^3} \quad \text{on} \quad \Omega,$$

where $w=w(x,y)$ represents the displacement of the glass sheet (the target shape) occupying a domain $\Omega\subset R^2$, $E=E(x,y)$ is Young's modulus, a positive function that can be influenced by adjusting the temperature in the process of heating the glass, f is the gravitational force, $\nu\in\left(0,\frac{1}{2}\right]$ is the glass Poisson ratio, and h is thickness of the plate. The *direct problem* (or the *forward problem*) is the following: for a given Young modulus E, find the displacement w of a glass sheet occupying a domain Ω before the heating process. Note that the PDE (1.2) is an elliptic fourth order linear PDE for the function w. Until now, two problems related to (1.2) have been studied: the *clamped plate case* and the *simply supported plate case* (more details can be found for example in [15]). In this paper we consider the clamped case, in which the following boundary conditions are required: the plate is placed over a rigid frame, i.e.,

$$w(x,y)|_{\partial\Omega}=0, \tag{1.3}$$

and, respectively,

$$\frac{\partial w}{\partial n}|_{\partial\Omega}=0, \tag{1.4}$$

which means the (outward) normal derivative of w must be zero, i.e., the sheet of glass is not allowed to freely rotate around the tangent to $\partial\Omega$. The associated *inverse problem* consists of finding Young's modulus E for a given data w in (1.2). This is a linear second order PDE for Young's modulus that can be written as

$$\begin{aligned}(w_{xx}+\nu w_{yy})E_{xx}+2(1-\nu)w_{xy}E_{xy}+(w_{yy}+\nu w_{xx})E_{yy}\\ +2(\Delta w)_xE_x+2(\Delta w)_yE_y+(\Delta^2 w)E=1\end{aligned} \tag{1.5}$$

after the scaling transformations $w\to\frac{1}{k}w$ or $E\to\frac{1}{k}E$, with $k=\frac{12(1-\nu^2)f}{h^3}$. In (1.5), Δ denotes the Laplace operator. The main problem in the car windshield design is that the prescribed target shape w is frequent such that the discriminant

$$D=(1-\nu)^2w_{xy}^2-(w_{xx}+\nu w_{yy})(w_{yy}+\nu w_{xx})$$

of (1.5) changes sign in the domain Ω, so that we get a mixed type PDE. This is one of the reasons for which optical defects might occur during the process. Note that (1.5) would naturally call for boundaries conditions for E on $\partial\Omega$ in the purely elliptic case (when $D < 0$), and Cauchy data on a suitable (noncharacteristic part) $\Gamma \subset \partial\Omega$ in the purely hyperbolic part (for $D > 0$). There is a recent interest in studying this inverse problem (see, e.g., [13]). It is known [15] that a constant Young's modulus corresponds to a data which satisfies the nonhomogeneous biharmonic equation (2.29). A survey on this subject can be found in [23]. Salazar and Westbrook [22] studied the case when the data and the parameter are given by radial functions; Kügler [12] used a derivative free iterative regularization method for analyzing the problem on rectangular frames; and a simplified model for the inverse problem on circular domains was considered by Engl and Kügler [7].

So far it is not obvious which shapes can be made by using this technique. Hence, we try to answer this question by finding out the symmetry reductions related to the PDE (1.5) hidden by the nonlinearity that occurs between the data and the parameter. In this sense, we determine (see section 3) the group of transformations that leave the equation unchanged, and so, its mixed type form. Knowledge of the invariants of these group actions allows us to write the target shape and the parameter in terms of them, and, therefore, to reduce the order of the studied equation. We find again the obvious result that a Young's modulus constant corresponds to data which is a solution of a nonhomogeneous biharmonic equation. The circular case problem considered by Salazar and Westbrook is, in fact, a particular case of our study. We show that other target shapes which are not radial functions can be considered. We prove that (1.5) is invariant under scaling transformations. It follows that target shapes modeled by homogeneous functions can be analyzed as well. In particular, we are interested in target shapes modeled by homogeneous polynomials defined on elliptical domains or square domains with rounded corners.

The paper is structured as follows. To reduce the order of the PDE (1.5) we propose in section 2 a method for studying the relation between the data and the parameter in terms of the similarity variables. The equivalence transformations related to this equation are given in section 3. The symbolic manipulation program DESOLV, authors Carminati and Vu [4] has been used for this purpose. Table 1 contains a complete classification of these symmetry reductions. In the last section, we discuss the PDE (1.5) augmented with the boundary conditions (1.3) and (1.4), namely, how to use the invariants of the group actions (on suitable bounded domains Ω) in order to incorporate the boundary conditions. In this sense, certain examples of exact and of numerical solutions of the reduced ODEs are given.

2. Conditional symmetries. The direct method approach to a second order PDE

$$\mathcal{F}(x, y, E^{(2)}) = 0$$

consists of seeking solutions written in the form

$$E(x, y) = \Phi(x, y, F(z)), \quad \text{where} \quad z = z(x, y), \quad (x, y) \in \Omega. \tag{2.1}$$

In this case the function z is called *similarity variable* and its level sets $\{z = k\}$ are named *similarity curves*. After substituting (2.1) into the studied second order PDE, we require that the result to be an ODE for the arbitrary function $F = F(z)$. Hence, certain conditions are imposed upon the functions Φ, z and their partial derivatives.

The particular case

$$E(x,y) = F(z(x,y)) \tag{2.2}$$

consists of looking for solutions depending only on the similarity variable z. If z is an invariant of the group action then the solutions of the form (2.2) are as well. Assume that the similarity variable is such that $\|\nabla z\| \neq 0$ on $\bar{\Omega}$.

In this section we apply this particular approach to (1.5) in order to study if the parameter and the data are functionally independent, which means whether or not they can depend on the same similarity variable. Assume that Young's modulus takes the form (2.2). In this case we get the relation

$$\begin{aligned} &F''(z)\left[z_x^2(w_{xx}+\nu w_{yy})+2z_xz_y(1-\nu)w_{xy}+z_y^2(w_{yy}+\nu w_{xx})\right] \\ &\quad + F'(z)\left[z_{xx}(w_{xx}+\nu w_{yy})+2(1-\nu)z_{xy}w_{xy}++z_{yy}(w_{yy}+\nu w_{xx})\right. \\ &\qquad \left. +2z_x(\Delta w)_x+2z_y(\Delta w)_y\right]+F(z)(\Delta^2 w)=1, \end{aligned} \tag{2.3}$$

which must be an ODE for the unknown function $F = F(z)$. This condition is satisfied if the coefficients of the partial derivatives of F are function of z only (note that these coefficients are also invariant under the same group action). Denote them by

$$\begin{aligned} \Gamma_1(z) &= z_x^2(w_{xx}+\nu w_{yy})+2z_xz_y(1-\nu)w_{xy}+z_y^2(w_{yy}+\nu w_{xx}), \\ \Gamma_2(z) &= z_{xx}(w_{xx}+\nu w_{yy})+2(1-\nu)z_{xy}w_{xy}+z_{yy}(w_{yy}+\nu w_{xx}) \\ &\quad +2z_x(\Delta w)_x+2z_y(\Delta w)_y, \\ \Gamma_3(z) &= \Delta^2 w. \end{aligned} \tag{2.4}$$

If these relations hold, then the PDE (1.5) is reduced to the second order linear ODE

$$\Gamma_1(z)F''(z)+\Gamma_2(z)F'(z)+\Gamma_3(z)F(z)=1. \tag{2.5}$$

2.1. Data and parameter invariant under the same group. If the target shape is invariant under the same group action as Young's modulus, then

$$w(x,y) = G(z(x,y)), \tag{2.6}$$

where $G = G(z)$. Substituting (2.6) into the relations (2.4) we get

$$\begin{aligned} \Gamma_1 &= G''(z_x^2+z_y^2)^2+G'\left[(z_x^2+\nu z_y^2)z_{xx}+2(1-\nu)z_xz_yz_{xy}+(z_y^2+\nu z_x^2)z_{yy}\right], \\ \Gamma_2 &= 2G'''(z_x^2+z_y^2)^2+G''\left\{[7z_x^2+(\nu+2)z_y^2]z_{xx}+2(5-\nu)z_xz_yz_{xy}\right. \\ &\quad \left.+[7z_y^2+(\nu+2)z_x^2]z_{yy}\right\}+G'\left\{(\Delta z)^2+2(1-\nu)(z_{xy}^2-z_{xx}z_{yy})\right. \\ &\quad \left.+2\left[z_x(\Delta z)_x+z_y(\Delta z)_y\right]\right\}, \\ \Gamma_3 &= G''''(z_x^2+z_y^2)^2+2G'''\left[(3z_x^2+z_y^2)z_{xx}+4z_xz_yz_{xy}+(z_x^2+3z_y^2)z_{yy}\right. \\ &\quad \left.+G''\left\{3(\Delta z)^2+4(z_{xy}^2-z_{xx}z_{yy})+4\left[z_x(\Delta z)_x+z_y(\Delta z)_y\right]\right\}+G'\Delta^2 z.\right. \end{aligned} \tag{2.7}$$

Next, the coefficients of the partial derivatives of the function G, denoted by Γ_i, must depend only on z, i.e.,

$$\begin{aligned} \Gamma_1 &= \alpha^4 G''+a_1G', \\ \Gamma_2 &= 2\alpha^4 G'''+a_2G''+a_3G', \\ \Gamma_3 &= \alpha^4 G''''+2a_4G'''+a_5G''+a_6G', \end{aligned}$$

where

$$
\begin{aligned}
&\alpha^2(z) = z_x^2 + z_y^2,\\
&a_1(z) = (z_x^2 + \nu z_y^2) z_{xx} + 2(1-\nu) z_x z_y z_{xy} + (z_y^2 + \nu z_x^2) z_{yy},\\
&a_2(z) = \left[7z_x^2 + (\nu+2) z_y^2\right] z_{xx} + 2(5-\nu) z_x z_y z_{xy} + \left[7z_y^2 + (\nu+2) z_x^2\right] z_{yy},\\
&a_3(z) = (\Delta z)^2 + 2(1-\nu)(z_{xy}^2 - z_{xx} z_{yy}) + 2\left[z_x (\Delta z)_x + z_y (\Delta z)_y\right],\\
&a_4(z) = (3z_x^2 + z_y^2) z_{xx} + 4 z_x z_y z_{xy} + (z_x^2 + 3z_y^2) z_{yy},\\
&a_5(z) = 3(\Delta z)^2 + 4(z_{xy}^2 - z_{xx} z_{yy}) + 4\left[z_x (\Delta z)_x + z_y (\Delta z)_y\right],\\
&a_6(z) = \Delta^2 z.
\end{aligned}
\tag{2.8}
$$

The first relation in (2.8) is a two-dimensional (2D) eikonal equation. From this we get

$$
\begin{aligned}
&z_x^2 z_{xx} + 2 z_x z_y z_{xy} + z_y^2 z_{yy} = \alpha^3(z)\alpha'(z),\\
&z_{xx} = \alpha(z)\alpha'(z) - \tfrac{z_y}{z_x} z_{xy},\\
&z_{yy} = \alpha(z)\alpha'(z) - \tfrac{z_x}{z_y} z_{xy}.
\end{aligned}
$$

The last two equations imply

$$
z_y^2 z_{xx} - 2 z_x z_y z_{xy} + z_x^2 z_{yy} = \alpha^3(z)\alpha'(z) - \alpha^4(z)\frac{z_{xy}}{z_x z_y}. \tag{2.9}
$$

Assume that there is a function $\beta = \beta(z)$ such that

$$
z_{xy} = \beta(z) z_x z_y. \tag{2.10}
$$

Indeed, since the left-hand side in (2.9) depends only on z, one can easily check if z satisfies both the 2D eikonal equation in (2.8) and (2.10), then all the functions $a_i = a_i(z)$ defined by (2.8) are written in terms of α and β. Therefore, the problem of finding the similarity variable z is reduced to that of integrating the 2D eikonal equation and the PDE system

$$
\begin{cases}
z_{xx} = \alpha\alpha' - \beta z_y^2,\\
z_{xy} = \beta z_x z_y,\\
z_{yy} = \alpha\alpha' - \beta z_x^2.
\end{cases}
\tag{2.11}
$$

The system (2.11) is compatible if the following relation holds:

$$
\alpha\alpha'' + \alpha'^2 - 3\beta\alpha\alpha' + \alpha^2\left(\beta^2 - \beta'\right) = 0.
$$

Denote $\mu = \frac{1}{2}\alpha^2$. In this case, the above compatibility condition can be written as

$$
\mu'' - 3\beta\mu' + 2\mu\left(\beta^2 - \beta'\right) = 0. \tag{2.12}
$$

On the other hand, if the function β is given by

$$
\beta(z) = -\frac{\lambda''(z)}{\lambda'(z)}, \tag{2.13}
$$

where λ is a nonconstant function, then (2.10) turns into

$$(\lambda(z))_{xy} = 0.$$

The general solution of this equation is given by

$$\lambda(z(x,y)) = a(x) + b(y), \tag{2.14}$$

with a and b being arbitrary functions. Substituting β from (2.13) into the compatibility condition (2.12) and after integrating once, we get

$$\mu'\lambda' + 2\mu\lambda'' = k, \tag{2.15}$$

where k is an arbitrary constant.

Case 1. If $k \neq 0$, then after integrating (2.15) and substituting back $\mu = \frac{1}{2}\alpha^2$, we get

$$\alpha^2(z) = \frac{2k\lambda(z) + C_1}{\lambda'^2(z)}. \tag{2.16}$$

The relation (2.14) implies $\lambda'(z)z_x = a'(x)$, and $\lambda'(z)z_y = b'(y)$. We substitute these relations, (2.14) and (2.16), into the 2D eikonal equation (see (2.8)). It follows that the functions $a = a(x)$ and $b = b(y)$ are solutions of the following respective ODEs:

$$a'^2(x) - 2ka(x) = C_2 \quad \text{and} \quad b'^2(y) - 2kb(y) = C_3,$$

with $C_2 + C_3 = C_1$ (here C_i are real constants). The above ODEs admit the nonconstant solutions

$$a(x) = \frac{1}{2k}\left[k^2(x - C_4)^2 - C_2\right] \quad \text{and} \quad b(y) = \frac{1}{2k}\left[k^2(y - C_5)^2 - C_3\right],$$

and so (2.14) takes the form

$$\lambda(z(x,y)) = \frac{k}{2}\left[(x - C_4)^2 + (y - C_5)^2\right] - \frac{C_1}{2k}. \tag{2.17}$$

Notice that $\frac{1}{k_1}\lambda$ or $\lambda + k_2$ defines the same function β as the function λ does. Moreover, since the PDE (1.5) is invariant under translations in the (x, y)-space, we can consider

$$\lambda(z(x,y)) = x^2 + y^2. \tag{2.18}$$

If $\sqrt{\lambda}$ is a bijective function on a suitable interval, and if we denote by $\Phi = (\sqrt{\lambda})^{-1}$ its inverse function, then the similarity variable written in the polar coordinates (r, θ) (where $x = r\cos(\theta)$, $y = r\sin(\theta)$) is given by

$$z(x,y) = \Phi(r). \tag{2.19}$$

For simplicity, we consider $\Phi = \mathrm{Id}$, and from that we get

$$E = F(r) \quad \text{and} \quad w = G(r), \quad \text{where} \quad z(x,y) = r. \tag{2.20}$$

Hence, the ODE (2.5) turns into

$$\left(G'' + \frac{\nu}{r}G'\right)F'' + \left(2G''' + \frac{\nu+2}{r}G'' - \frac{1}{r^2}G'\right)F' + \left(G'''' + \frac{2}{r}G''' - \frac{1}{r^2}G'' + \frac{1}{r^3}G'\right)F = 1, \tag{2.21}$$

which can be reduced to the first order ODE

$$\left(G'' + \frac{\nu}{r}G'\right)F' + \left(G''' + \frac{1}{r}G'' - \frac{1}{r^2}G'\right)F = \frac{r^2 - r_0^2}{2r} + \frac{\gamma}{r}, \tag{2.22}$$

where $r_0 \in [0,1]$ with the property that

$$\gamma = \left[\left(rG'' + \nu G'\right)F' + \left(rG''' + G'' - \frac{1}{r}G'\right)F\right]|_{r=r_0}$$

is finite. The smoothness condition $G'(0) = 0$ implies that (2.22) can be written as [15]

$$\left(G'' + \frac{\nu}{r}G'\right)F' + \left(G''' + \frac{1}{r}G'' - \frac{1}{r^2}G'\right)F = \frac{r}{2}. \tag{2.23}$$

Case 2. If $k = 0$, similarly we get

$$z(x,y) = \Phi(k_1 x + k_2 y), \tag{2.24}$$

where k_1 and k_2 are real constants such that $k_1^2 + k_2^2 > 0$. In this case, for $\Phi = \mathrm{Id}$, the parameter and the data are written as

$$E = F(z) \quad \text{and} \quad w = G(z), \quad \text{where} \quad z(x,y) = k_1 x + k_2 y, \tag{2.25}$$

and the ODE (2.5) turns into

$$G''(z)F''(z) + 2G'''(z)F'(z) + G''''(z)F(z) = \frac{1}{(k_1^2 + k_2^2)^2}, \tag{2.26}$$

with $\{z|G''(z) = 0\}$ the associated set of singularities. Integrating the above ODE on the set $\{z|G''(z) \neq 0\}$ we obtain that Young's modulus is given by

$$E(x,y) = \frac{(k_1 x + k_2 y)^2 + C_1(k_1 x + k_2 y) + C_2}{2(k_1^2 + k_2^2)^2 G''(k_1 x + k_2 y)},$$

where C_i are arbitrary constants.

2.2. Data and parameter invariant under different groups. Consider two functionally independent functions on Ω, say, $z = z(x,y)$ and $v = v(x,y)$, and let

$$w = H(v(x,y)) \tag{2.27}$$

be the target shape. In this case, the data and the parameter do not share the same invariance. Similar to the above, substituting (2.27) into the relations (2.4) we get

(2.28)

$$\begin{aligned}
\Gamma_1 &= H''\left[(z_x v_x + z_y v_y)^2 + \nu(z_y v_x - z_x v_y)^2\right] \\
&\quad + H'\left[z_x^2 v_{xx} + 2z_x z_y v_{xy} + z_y^2 v_{yy} + \nu\left(z_x^2 v_{yy} - 2z_x z_y v_{xy} + z_y^2 v_{xx}\right)\right], \\
\Gamma_2 &= H'''(v_x^2 + v_y^2)(z_x v_x + z_y v_y) + H''\left[v_x^2 z_{xx} + 2v_x v_y z_{xy} + v_y^2 z_{yy}\right. \\
&\quad \left. + \nu\left(v_y^2 z_{xx} - 2v_x v_y z_{xy} + v_x^2 z_{yy}\right) + 2z_x v_x v_{xx} + 2(z_x v_y + z_y v_x)v_{xy} + 2z_y v_y v_{yy}\right. \\
&\quad \left. + (z_x v_x + z_y v_y)(\Delta v)\right] + H'\left[z_{xx} v_{xx} + 2z_{xy} v_{xy} + z_{yy} v_{yy} + \nu\left(z_{xx} v_{yy} - 2z_{xy} v_{xy}\right.\right. \\
&\quad \left.\left. + z_{yy} v_{xx}\right) + z_x(\Delta v)_x + z_y(\Delta v)_y\right], \\
\Gamma_3 &= H''''(v_x^2 + v_y^2)^2 + 2H'''\left[(3v_x^2 + v_y^2)v_{xx} + 4v_x v_y v_{xy} + (v_x^2 + 3v_y^2)v_{yy}\right] \\
&\quad + H''\left[3v_{xx}^2 + 4v_{xy}^2 + 3v_{yy}^2 + 2v_{xx}v_{yy} + 4v_x(\Delta v)_x + 4v_y(\Delta v)_y\right] + H'\Delta^2 v.
\end{aligned}$$

Recall that Γ_i's are functions of $z = z(x, y)$ only. Since each right-hand side in the above relations contains the function $H = H(v)$ and its derivatives, we require that the coefficients of the derivatives of H to be functions of v. It follows that Γ_i must be constant and denote them by γ_i. Therefore, the last condition in (2.28) becomes

$$\Delta^2(w) = \gamma_3, \tag{2.29}$$

which is the biharmonic equation. According to the above assumption, we seek solutions of (2.29) that are functions of v only. Similar to section 2.1, we get

$$v(x, y) = \Psi(r), \quad \text{or} \quad v(x, y) = \Psi(k_1 x + k_2 y), \tag{2.30}$$

and thus, for $\Psi = \mathrm{Id}$, the target shape is written as

$$w(x, y) = H(r), \quad \text{or} \quad w(x, y) = H(k_1 x + k_2 y). \tag{2.31}$$

Since $z = z(x, y)$ and $v = v(x, y)$ are functionally independent, we get

$$z(x, y) = k_1 x + k_2 y, \quad v(x, y) = \sqrt{x^2 + y^2} \tag{2.32}$$

or

$$z(x, y) = \sqrt{x^2 + y^2}, \quad v(x, y) = k_1 x + k_2 y. \tag{2.33}$$

One can prove that if the coefficients γ_i are constant, and if z and v are given by (2.32) or (2.33), respectively, then $\gamma_1 = \gamma_2 = 0$, and $\gamma_3 \neq 0$. On the other hand, the solutions of the biharmonic equation (2.29) of the form (2.31) are the following:

$$w(x, y) = \frac{\gamma_3}{64} z^4 + C_1 z^2 + C_2 \ln(z) + C_3 z^2 \ln(z) + C_4 \quad \text{for} \quad z = \sqrt{x^2 + y^2},$$

and, respectively,

$$w(x, y) = \frac{\gamma_3}{24(k_1^2 + k_2^2)^2} v^4 + C_1 v^3 + C_2 v^2 + C_3 v + C_4 \quad \text{for} \quad v = k_1 x + k_2 y,$$

and these correspond to the constant Young's modulus

$$E(x, y) = \frac{1}{\gamma_3}. \tag{2.34}$$

Notice that only particular solutions of the biharmonic equation have been found in this case (i.e., solutions invariant under rotations and translations). Since this PDE is also invariant under scaling transformations, which act not only on the space of the independent variables but on the data space as well, it is obvious to extend our study and to seek other types of symmetry reductions.

3. Equivalence transformations. Consider a one-parameter Lie group of transformations acting on an open set $\mathcal{D} \subset \Omega \times \mathcal{W} \times \mathcal{E}$, where $\mathcal{W}$ is the space of the data functions, and $\mathcal{E}$ is the space of the parameter functions, given by

$$\begin{cases} x^* = x + \varepsilon \zeta(x, y, w, E) + \mathcal{O}(\varepsilon^2), \\ y^* = y + \varepsilon \eta(x, y, w, E) + \mathcal{O}(\varepsilon^2), \\ w^* = w + \varepsilon \phi(x, y, w, E) + \mathcal{O}(\varepsilon^2), \\ E^* = E + \varepsilon \psi(x, y, w, E) + \mathcal{O}(\varepsilon^2), \end{cases} \tag{3.1}$$

where ε is the group parameter. Let

$$V=\zeta(x,y,w,E)\partial_x+\eta(x,y,w,E)\partial_y+\phi(x,y,w,E)\partial_w+\psi(x,y,w,E)\partial_E \tag{3.2}$$

be its associated general infinitesimal generator. The group of transformations (3.1) is called an *equivalence transformation* associated to the PDE (1.5) if this leaves the equation invariant. This means that the form of the equation in the new coordinates remains unchanged and the set of the analytical solutions is invariant under this transformation. The equivalence transformations can be found by applying the classical Lie method to (1.5), with E and w both considered as unknown functions (for more details see [10] and [21]). Following this method we obtain

$$\begin{cases}\zeta(x,y,w,E)=k_1+k_5x-k_4y,\\ \eta(x,y,w,E)=k_2+k_4x+k_5y,\\ \phi(x,y,w,E)=k_3+k_7x+k_6y+(4k_5-k_8)w,\\ \psi(x,y,w,E)=k_8E,\end{cases} \tag{3.3}$$

where k_i are real constants. The vector field (3.2) is written as $V=\sum_{i=1}^{8}k_iV_i$, where

$$\begin{gathered}V_1=\partial_x,\quad V_2=\partial_y,\quad V_3=\partial_w,\quad V_4=-y\partial_x+x\partial_y,\quad V_5=x\partial_x+y\partial_y+4w\partial_w,\\ V_6=y\partial_w,\quad V_7=x\partial_w,\quad V_8=-w\partial_w+E\partial_E.\end{gathered} \tag{3.4}$$

PROPOSITION 3.1. *The equivalence transformations related to the PDE* (1.5) *are generated by the infinitesimal generators* (3.4)*. Thus, the equation is invariant under translations in the x-space, y-space, w-space, rotations in the space of the independent variables (x,y), scaling transformations in the (x,y,w)-space, Galilean transformations in the (y,w) and (x,w) spaces, and scaling transformations in the (w,E)-space, respectively.*

Notice that the conditional symmetries found in section 2 represent particular cases of the equivalence transformations. Since each one-parameter group of transformations generated by V_i is a symmetry group, if ($w=G(x,y)$, $E=F(x,y)$) is a pair of known solutions of (1.5), so are the following:

$$\begin{aligned}
&w^{(1)}=G(x-\varepsilon_1,y), && E^{(1)}=F(x-\varepsilon_1,y),\\
&w^{(2)}=G(x,y-\varepsilon_2), && E^{(2)}=F(x,y-\varepsilon_2),\\
&w^{(3)}=G(x,y)+\varepsilon_3, && E^{(3)}=F(x,y),\\
&w^{(4)}=G(\tilde{x},\tilde{y}), && E^{(4)}=F(\tilde{x},\tilde{y}),\\
&w^{(5)}=e^{4\varepsilon_5}G(e^{-\varepsilon_5}x,e^{-\varepsilon_5}y), && E^{(5)}=F(e^{-\varepsilon_5}x,e^{-\varepsilon_5}y),\\
&w^{(6)}=G(x,y)+\varepsilon_6y, && E^{(6)}=F(x,y),\\
&w^{(7)}=G(x,y)+\varepsilon_7x, && E^{(7)}=F(x,y),\\
&w^{(8)}=e^{-\varepsilon_8}G(x,y), && E^{(8)}=e^{\varepsilon_8}F(x,y),
\end{aligned} \tag{3.5}$$

where $\tilde{x} = x\cos(\varepsilon_4) + y\sin(\varepsilon_4)$, $\tilde{y} = -x\sin(\varepsilon_4) + y\cos(\varepsilon_4)$, and ε_i are real constants. Moreover, the general solution of (1.5) constructed from a known one is given by

$$w(x,y) = e^{4\varepsilon_5-\varepsilon_8} G(e^{-\varepsilon_5}(\tilde{x}-\tilde{k}_1), e^{-\varepsilon_5}(\tilde{y}-\tilde{k}_2)) + e^{4\varepsilon_5-\varepsilon_8}\varepsilon_6 y + e^{4\varepsilon_5-\varepsilon_8}\varepsilon_7 x + e^{4\varepsilon_5-\varepsilon_8}\varepsilon_3,$$
$$E(x,y) = e^{\varepsilon_8} F(e^{-\varepsilon_5}(\tilde{x}-\tilde{k}_1), e^{-\varepsilon_5}(\tilde{y}-\tilde{k}_2)),$$

where $\tilde{k}_1 = \varepsilon_1\cos(\varepsilon_4) + \varepsilon_2\sin(\varepsilon_4)$, and $\tilde{k}_2 = \varepsilon_1\sin(\varepsilon_4) - \varepsilon_2\cos(\varepsilon_4)$.

The equivalence transformations form a Lie group $\mathcal{G}$ with an eight-dimensional associated Lie algebra $\mathcal{A}$. Using the adjoint representation of $\mathcal{G}$, one can find the optimal system of one-dimensional subalgebras of $\mathcal{A}$ (more details can be found in [18, pp. 203–209]). This optimal system is spanned by the vector fields given in Table 1. Denote by z, I, and J the invariants related to the one-parameter group of transformations generated by each vector field V_i. Here F and G are arbitrary functions, (r,θ) are the polar coordinates, and a, b, c are nonzero constants. To reduce the order of the PDE (1.5) one can also integrate the first order PDE system

$$\begin{cases} \zeta(x,y,w,E)w_x + \eta(x,y,w,E)w_y &= \phi(x,y,w,E), \\ \zeta(x,y,w,E)E_x + \eta(x,y,w,E)E_y &= \psi(x,y,w,E), \end{cases} \tag{3.6}$$

which defines the characteristics of the vector field (3.2). In Table 1, the associated reduced ODEs are listed. The invariance of (1.5) under the one-parameter groups of transformations generated by V_1, V_2, $V_1 + cV_6$, and $V_2 + cV_7$, respectively, leads us to the same ODE,

$$F''(z)G''(z) + 2F'(z)G'''(z) + F(z)G''''(z) = 1, \tag{3.7}$$

with the general solution

$$F(z) = \frac{z^2 + C_1 z + C_2}{2G''(z)} \tag{3.8}$$

on the set $\{z | G''(z) \neq 0\}$. The invariance under the scaling transformation generated by the vector field V_5 yields the reduced ODE

$$\begin{aligned} &\left[G''\left(z^2+1\right)^2 - 6z(z^2+1)G' + 12(z^2+\nu)G\right]F'' \\ &+ 2\left[\left(z^2+1\right)^2 G''' - 5z(z^2+1)G'' + 3(4z^2+\nu+1)G' - 12zG\right]F' \\ &+ \left[\left(z^2+1\right)^2 G'''' - 4z(z^2+1)G''' + 4(3z^2+1)G'' - 24zG' + 24G\right]F = 1. \end{aligned} \tag{3.9}$$

The ODE

$$\begin{aligned} &\left[\left(z^2+1\right)^2 G'' + 2(c-3)z(z^2+1)G' + (c-3)(c-4)(z^2+\nu)G\right]F'' \\ &+ \Big\{2\left(z^2+1\right)^2 G''' + 2(2c-5)z(z^2+1)G'' + 2(c-3)[z^2(c-4) + \nu(c-1) - 1]G' \\ &-2(c-3)(c-4)zG\Big\}F' + \Big\{\left(z^2+1\right)^2 G'''' + 2(c-2)z(z^2+1)G''' + \big[(c-3)(c-4)z^2 \\ &-2(c-2) + \nu c(c-1)\big]G'' - 2(c-4)(c-3)zG' + 2(c-4)(c-3)G\Big\}F = 1 \end{aligned} \tag{3.10}$$

TABLE 1

	Infintesimal generator	Invariants	$w = w(x,y)$	$E = E(x,y)$	ODE
1.	V_1	$z = y$ $I = w$ $J = E$	$w = G(z)$	$E = F(z)$	(3.7)
2.	V_2	$z = x$ $I = w$ $J = E$	$w = G(z)$	$E = F(z)$	(3.7)
3.	V_4	$z = r$ $I = w$ $J = E$	$w = G(z)$	$E = F(z)$	(2.21)
4.	V_5	$z = \frac{y}{x}$ $I = x^{-4}w$ $J = E$	$w = x^4 G(z)$	$E = F(z)$	(3.9)
5.	$cV_3 + V_4$	$z = r$ $I = w - c\theta$ $J = E$	$w = c\theta + G(z)$	$E = F(z)$	(2.21)
6.	$V_5 + cV_8$	$z = \frac{y}{x}$ $I = x^{c-4}w$ $J = x^{-c}E$	$w = x^{4-c}G(z)$	$E = x^c F(z)$	(3.10)
7.	$V_4 + cV_8$	$z = r$ $I = e^{c\theta}w$ $J = e^{-c\theta}E$	$w = e^{-c\theta}G(z)$	$E = e^{c\theta}F(z)$	(3.11)
8.	$V_4 + cV_5$	$z = re^{-c\theta}$ $I = r^{-4}w$ $J = E$	$w = r^4 G(z)$	$E = F(z)$	(3.13)
9.	$V_4 + cX_5 + bV_8$	$z = re^{-c\theta}$ $I = r^{\frac{b}{c}-4}w$ $J = r^{-\frac{b}{c}}E$	$w = r^{4-\frac{b}{c}}G(z)$	$E = r^{\frac{b}{c}}F(z)$	(3.14)
10.	$V_1 + cV_6$	$z = y$ $I = w - cxy$ $J = E$	$w = cxy + G(z)$	$E = F(z)$	(3.7)
11.	$V_2 + cV_7$	$z = x$ $I = w - cxy$ $J = E$	$w = cxy + G(z)$	$E = F(z)$	(3.7)
12.	$V_1 + cV_8$	$z = y$ $I = e^{cx}w$ $J = e^{-cx}E$	$w = e^{-cx}G(z)$	$E = e^{cx}F(z)$	(3.15)
13.	$V_2 + cV_8$	$z = x$ $I = e^{cy}w$ $J = e^{-cy}E$	$w = e^{-cy}G(z)$	$E = e^{cy}F(z)$	(3.15)

is obtained in case 6 of Table 1. The reduced equation

(3.11)
$$\left[G'' + \frac{\nu}{r}G' + \frac{\nu c^2}{r^2}G\right]F'' + \left[2G''' + \frac{\nu+2}{r}G'' + \frac{2\nu c^2 - 1}{r^2}G' - \frac{c^2(1+2\nu)}{r^3}G\right]F'$$

$$+ \left[G'''' + \frac{2}{r}G''' + \frac{c^2\nu - 1}{r^2}G'' + \frac{1 - c^2(2\nu+1)}{r^3}G' + \frac{2c^2(\nu+1)}{r^4}G\right]F = 1$$

is related to case 7. This can be written as the first order ODE

(3.12)
$$\left(G'' + \frac{\nu}{r}G' + \frac{\nu c^2}{r^2}G\right)F' + \left(G''' + \frac{1}{r}G'' + \frac{c^2\nu - 1}{r^2}G' - \frac{c^2(1+\nu)}{r^3}G\right)F = \frac{r^2 - r_0^2}{2r} + \frac{\gamma^*}{r},$$

where $r_0 \in [0,1]$ with the property that

$$\gamma^* = \left[F'\left(rG'' + \nu G' + \frac{\nu}{r}G\right) + F\left(rG''' + G'' + \frac{c^2\nu - 1}{r}G' - \frac{c^2(1+\nu)}{r^2}G\right)\right]|_{r=r_0}$$

is finite. In cases 8 and 9, after the change of the variable $z = \exp(t)$, the reduced ODEs are the following:

(3.13)
$$\begin{aligned}
&\left\{(c^2+1)^2G'' + (c^2+1)(\nu+7)G' + 4\left[\nu(3c^2+1) + c^2 + 3\right]G\right\}F'' \\
&+ \left\{2(c^2+1)^2G''' + (c^2+1)(\nu+19)G'' + 2[16 + (c^2+1)(3\nu+13)]G' + 8(\nu+7)G\right\}F' \\
&+ \left\{(c^2+1)^2G'''' + 12(c^2+1)G''' + 4(5c^2+13)G'' + 96G' + 64G\right\}F = 1,
\end{aligned}$$

and, respectively,

(3.14)
$$\begin{aligned}
&\left\{(c^2+1)^2G'' + \left(\frac{1}{c} + c\right)[c(\nu+7) - 2b]G' + \left(\frac{4}{c} - \frac{b}{c^2}\right)[c^3(1+3\nu) - c^2\nu b\right. \\
&\left. + c(\nu+3) - b]G\right\}F'' + \left\{2(c^2+1)^2G''' + \left(\frac{1}{c} + c\right)[c(\nu+19) - 4b]G''\right. \\
&+ 2\left[\frac{b^2}{c^2} + \nu b^2 + c^2(3\nu+13) - 4bc(\nu+1) - 12\frac{b}{c} + 3\nu + 29\right]G' \\
&\left. + \left(\frac{4}{c} - \frac{b}{c^2}\right)[2c(\nu+7) + b(\nu-5)]G\right\}F' + \left\{(c^2+1)^2G'''' \right. \\
&+ 2\left(c + \frac{1}{c}\right)(6c - b)G''' + \left[\frac{b^2}{c^2} + \frac{b}{c}(\nu - 17) + 20c^2 - bc(\nu+7) + \nu b^2 + 52\right]G'' \\
&\left. + \left(\frac{6}{c} - \frac{b}{c^2}\right)[16c + b(\nu-5)]G' + 2\left(\frac{4}{c} - \frac{b}{c^2}\right)[8c + b(\nu-3)]G\right\}F = 1.
\end{aligned}$$

In cases 12 and 13 we get the same equation,

$$\left(G'' + \nu c^2 G\right)F'' + 2\left(G''' + \nu c^2 G'\right)F' + \left(G'''' + \nu c^2 G''\right)F = 1, \tag{3.15}$$

with the general solution given by

$$F(z) = \frac{z^2 + C_1 z + C_2}{G''(z) + \nu c^2 G(z)} \tag{3.16}$$

on the set $\{z | G''(z) + \nu c^2 G(z) \neq 0\}$, where C_1 and C_2 are arbitrary real constants.

4. Conclusions. The data w is the function that models the target shape of a car windshield. Hence, we seek data with relevant physical and geometrical properties, such as smoothness and a positive curvature graph at least in the center of the bounded domain Ω, for which the boundary condition (1.3) is satisfied—which means that the sheet of glass is placed over a rigid frame. Moreover, if there is no free rotation of the plate around the tangent to $\partial\Omega$, then the condition (1.4) is required. Applying symmetry reductions theory to the PDE (1.5), we have shown that the data w and Young's modulus E can be expressed in terms of the invariants z, I, and J associated with a certain group action, i.e., $I = G(z)$ and $J = F(z)$, where w occurs in I, and E in J, respectively. Since now the technique of reducing the PDE (1.5) to an ODE has been applied only in the case of the radial functions ([15], [22], and [23]). Other symmetry reductions related to the studied model can be derived and these are listed in Table 1. The data given by homogeneous polynomials can be related to the invariance of the equation with respect to the scaling transformations (see cases 4 and 6, Table 1). The first two and last four cases in Table 1 allow us to construct other kind of data (see (3.5)). Since Ω must be bounded, the most interesting cases correspond to the rotational and the scaling symmetries. The problem of finding exact solutions of the reduced equations, which are second order linear nonhomogeneous ODEs, might be a difficult task depending on the form of the data. These equations are also, in general, ill-posed, as the initial problem is, and hence, regularization methods might be required in order to be studied, which is our current research.

One can make the following remarks: assume that $\partial\Omega = \{(x,y)|\ z(x,y) = k\}$ is the k-level set of the function z (here k being a nonzero constant) and $||\nabla z|| > 0$ on $\bar{\Omega}$. If the target shape is given by $w(x,y) = a(x,y)G(z(x,y))$, where $a = a(x,y)$ is a suitable function according to Table 1, then the boundary conditions (1.3) and (1.4) are equivalent to $G(k) = 0$ and $G'(k) = 0$. Therefore, the data might have the form $w(x,y) = a(x,y)\left(z(x,y) - k\right)^2 H(z(x,y))$. This corresponds to the case when the data and the bounded domain Ω are invariant under the same symmetry reduction. In our case, this can be applied to rotational symmetries. For scaling invariance, we have to incorporate the noninvariant boundary conditions in invariant solutions. As a consequence, we can extend the study of the problem on elliptical domains and on square domains with rounded corners. For instance, the class of target shapes of the form $w(x,y) = z^m(x,y) - k^m$, where $m \geq 1$ is a natural number, satisfies the boundary condition (1.3). In this case, the normal derivative of the data on the boundary is $\frac{\partial w}{\partial n}|_{\partial\Omega} = mk^{m-1}\|\nabla z\||_{\partial\Omega}$. If this quantity is small then the condition (1.4) is almost satisfied (i.e., there is a small free rotation of the plate around the tangent to $\partial\Omega$). In the following examples, we assume that the glass Poisson ration $\nu = 0.5$.

Example 1. *Rotational invariant data and parameter.* Consider the target shape of the form [23]

$$w(x,y) = G(r) = -\frac{1}{6}\left(r-1\right)^2\left(2r+1\right), \quad r = \sqrt{x^2+y^2},$$

defined on the unit disc (see Figure 4.1) which satisfies the boundary conditions (1.3) and (1.4). Since $G'(0) = 0$, the reduced ODE is (2.23) and this has a singularity at $r = \frac{3}{5}$. Since $E > 0$, we consider the constant of integration $C_1 = 1$, and so,

$$E(x,y) = F(r) = -\frac{1}{11}\left(r + \frac{1}{2}\right) + (5r-3)^{-\frac{6}{5}}.$$

The PDE (1.5) is elliptic for $r \in \left(\frac{3}{5}, \frac{3}{4}\right)$, hyperbolic for $r \in \left[0, \frac{3}{5}\right) \bigcup \left(\frac{3}{5}, 1\right]$, and parabolic if $r = \frac{3}{5}$ or $r = \frac{3}{4}$, respectively.

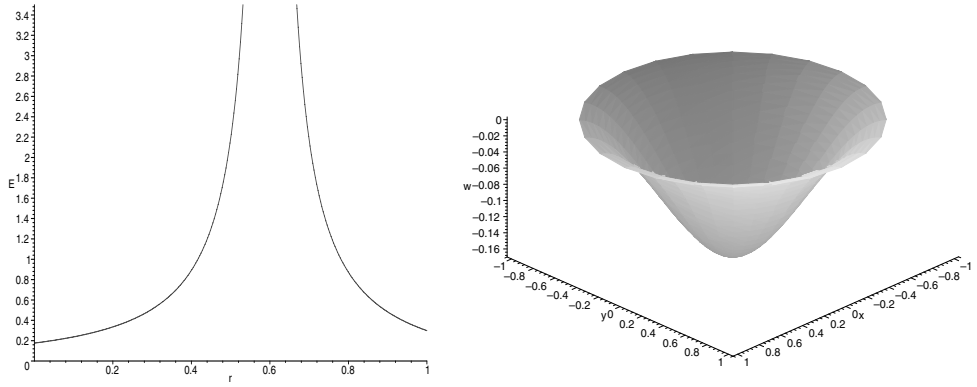


FIG. 4.1. *The parameter* $E(x,y) = F(r) = -\frac{1}{11}\left(r+\frac{1}{2}\right) + (5r-3)^{-\frac{6}{5}}$ *and the target shape* $w(x,y) = G(r) = -\frac{1}{6}(r-1)^2(2r+1)$, *with* $r = \sqrt{x^2+y^2}$, *defined on the unit disc.*

Example 2. *Particular target shapes on rounded square domains.*

(a) Suppose the Lamé oval $\partial\Omega = \{(x,y)|\ x^{2n} + y^{2n} = 1\}$ is the boundary of the domain (here $n \geq 2$ is a natural number). For a target shape of the form

$$w(x,y) = \left(x^{2n} + y^{2n}\right)^m - 1, \tag{4.1}$$

$m \geq 1$ being a natural number, (1.5) is elliptic on $\Omega - \{(0,0)\}$ and parabolic in $(0,0)$. These target shapes are invariant with respect to $V_5 + cV_8 + (4-c)V_3$, where $c = 4 - 2mn$. For $x > 0$ or $x < 0$, the functions (4.1) can be written as

$$w(x,y) = x^{2mn}G(z) - 1, \quad G(z) = \left(1 + z^{2n}\right)^m, \quad z = \frac{y}{x}.$$

According to case 6 in Table 1, the associated Young's modulus has the form

$$E(x,y) = x^{4-2mn}F(z), \quad z = \frac{y}{x}.$$

Since $w(x,y) = w(y,x) = w(-x,y) = w(x,-y) = w(-x,-y)$, Young's modulus also shares these discrete symmetries. Thus, the reduced ODE (3.10) can be integrated for $z \in [0,1]$. In particular, for $n = 2$ and $m = 1$, the data is a solution of the biharmonic equation and Young's modulus is $E = 48^{-1}$. For $n = 3$ and $m = 1$, the data and the numerical solution F satisfying $F(0) = 0.002$ and $F'(0) = 0$ are given in Figure 4.2.

(b) Assume that $\partial\Omega = \{(x,y)|\ x^{2n} + y^2 = 1\}$, where $n \geq 1$ is a natural number. Consider the class of target shapes

$$w(x,y) = x^{2n} + y^2 - 1,$$

invariant under the vector field $V_2 + 2V_6$. Hence, (1.5) is reduced to the ODE (3.7). For $n = 3$, the associated Young's modulus is given by

$$E(x,y) = F(x) = \frac{x^2 + C_1x + C_2}{2(30x^4+1)},$$

and since $E > 0$, we can set $C_1 = 0$ and $C_2 = 2$ (see Figure 4.3). Equation (1.5) is elliptic on $\Omega - Oy$ and parabolic on the y-axis.

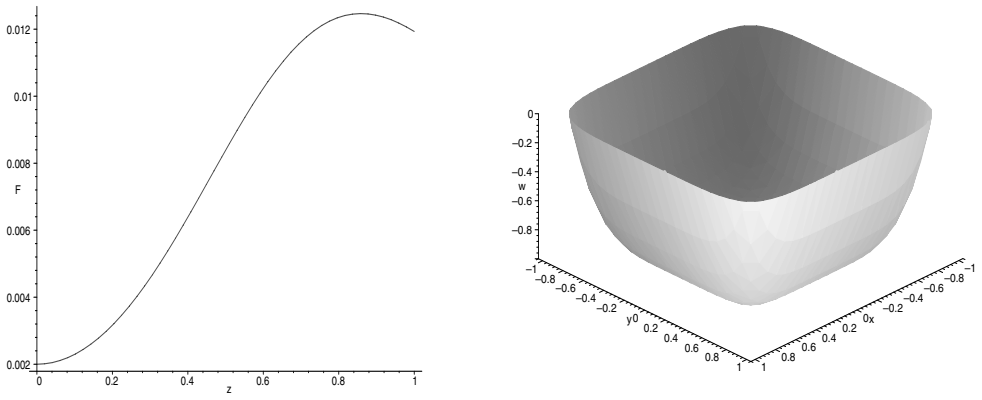


FIG. 4.2. *The parameter* $E(x,y) = x^{-2}F(z)$, *with* $z = \frac{y}{x}$, *and the data* $w(x,y) = x^6 + y^6 - 1$ *defined on the rounded square domain* $\{(x,y)|\ x^6 + y^6 < 1\}$.

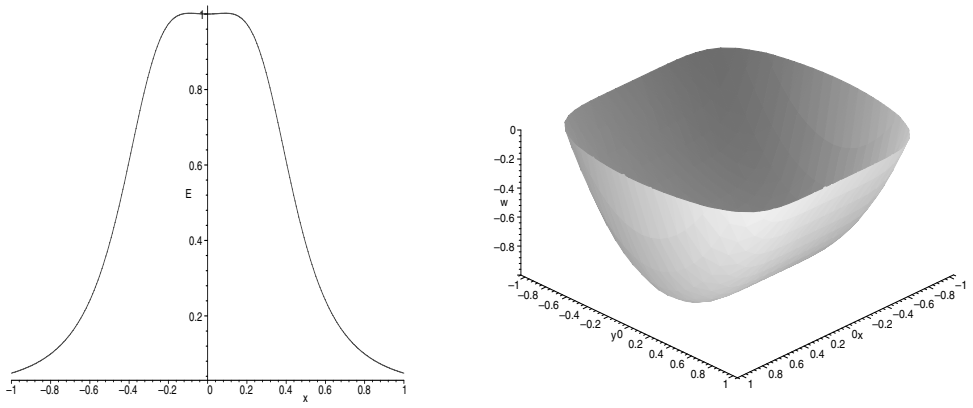


FIG. 4.3. *The parameter* $E(x,y) = \frac{x^2+2}{2(30x^4+1)}$ *and the data* $w(x,y) = x^6 + y^2 - 1$ *defined on the rounded square domain* $\{(x,y)|\ x^6 + y^2 < 1\}$.

Example 3. *Particular target shapes on elliptic domains.* Consider the data

$$w(x,y) = \left(\frac{x^2}{a^2} + \frac{y^2}{b^2}\right)^m - 1, \tag{4.2}$$

on the elliptic domain $\Omega = \{(x,y)|\ \frac{x^2}{a^2} + \frac{y^2}{b^2} < 1\}$, where $m \geq 1$ is a natural number. These target shapes are obtained from the invariance of the studied PDE with respect to $V_5 + cV_8 + (4-c)V_3$, where $c = 4 - 2m$. The PDE (1.5) is elliptic on $\Omega - \{(0,0)\}$ and parabolic in the origin. For $x > 0$ or $x < 0$, the functions (4.2) can be written as

$$w(x,y) = x^{2m}G(z) - 1, \quad G(z) = \left(\frac{1}{a^2} + \frac{z^2}{b^2}\right)^m, \quad z = \frac{y}{x}.$$

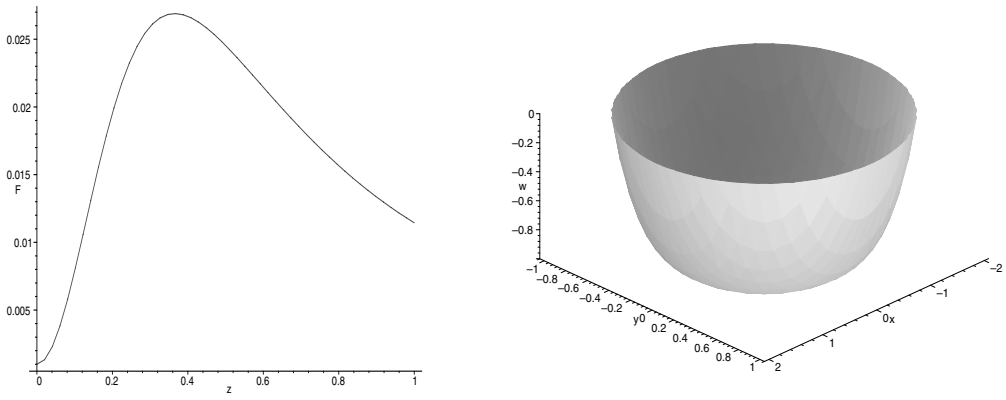


FIG. 4.4. *The parameter* $E(x,y) = x^{-2}F(z)$, *with* $z = \frac{y}{x}$, *and the data* $w(x,y) = \left(\frac{x^2}{4} + y^2\right)^3 - 1$ *defined on the elliptical domain* $\{(x,y) | \ \frac{x^2}{4} + y^2 = 1\}$.

In this case, we look for solutions to (1.5) of the form

$$E(x,y) = x^{4-2m}F(z), \quad z = \frac{y}{x}.$$

If $m = 2$, the data is a solution of the biharmonic equation, and the related Young's modulus is $E = 24\left(a^{-4} + b^{-4}\right) + 16a^{-2}b^{-2}$. If $m \geq 3$, the reduced ODE is (3.9). For $m = 3$, the data and the numerical solution F of (3.9) satisfying $F(0) = 0.001$ and $F'(0) = 0$ are plotted in Figure 4.4.

In brief, suppose that the target shape w on a domain Ω is given. In order to see if this is an invariant function with respect to the equivalence transformations related to the studied model, we should check if this is a solution of the first equation in (3.6), where the functions ζ, η, and ϕ are given by (3.3). Next, by integrating the second PDE in (3.6) we can determine the form of the parameter in terms of the similarity variables. The geometrical significance of the nonlinearity occurring between the data and the parameter in the inverse problem (1.1) is reflected by the group analysis tools. Investigating special groups of transformations connected to this equation, the order of the model can be reduced. The equation will be then written in terms of the invariants of the group actions. Another advantage of this approach is that of relating the direct and inverse problems through these symmetry reductions. It might be interesting for future study to link these results to the common approach of the inverse problems theory, especially in expressing the regularization methods in terms of the similarity variables. For other target shapes defined by functions which are not invariant under the listed symmetry reductions, the classical theory of the linear second order PDEs can be applied, but this might be quite difficult due to the form of the discriminant.

Acknowledgments. I would like to thank Prof. Heinz W. Engl and Dr. Philipp Kügler, Institute for Industrial Mathematics at Johannes Kepler University, for encouraging my research in applying the symmetry analysis to parameter identification problems, for interesting discussions, and for guiding me through this subject. I am grateful to Prof. Peter Olver, School of Mathematics, Institute of Technology at the University of Minnesota, for helpful comments which improved the presentation of this paper.

REFERENCES

[1] G. W. Bluman and J. D. Cole, *The general similarity solutions of the heat equation*, J. Math. Mech., 18 (1969), pp. 1025–1042.

[2] G. W. Bluman and S. Kumei, *Symmetries and Differential Equations*, Appl. Math. Sci. 81, Springer-Verlag, New York, 1989.

[3] C. J. Budd and M. D. Piggott, *Geometric integration and its applications*, in Handbook of Numerical Analysis, XI, North–Holland, Amsterdam, 2003, pp. 35–139.

[4] J. Carminati and K. Vu, *Symbolic computation and differential equations: Lie symmetries*, J. Symbolic Comput., 29 (2000), pp. 95–116.

[5] P. A. Clarkson and M. Kruskal, *New similarity reductions of the Boussinesq equation*, J. Math. Phys., 30 (1989), pp. 2201–2213.

[6] V. A. Dorodnitsyn, *Finite difference models entirely inheriting symmetry of original differential equations*, in Modern Group Analysis: Advanced Analytical and Computational Methods in Mathematical Physics, N. H. Ibragimov, M. Torrisi, and A. Valenti, eds., Kluwer, Dordrecht, The Netherlands, 1993, pp. 191–201.

[7] H. W. Engl and P. Kügler, *The influence of the equation type on iterative parameter identification problems which are elliptic or hyperbolic in the parameter*, European J. Appl. Math., 14 (2003), pp. 129–163.

[8] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, The Netherlands, 1996.

[9] W. Hereman, *Review of symbolic software for the computation of Lie symmetries of differential equations*, Euromath Bull., 1 (1994), pp. 45–82.

[10] N. H. Ibragimov, *CRC Handbook of Lie Group Analysis of Differential Equations, Vol.* 1: *Symmetries, Exact Solutions and Conservation Laws*, CRC Press, Boca Raton, 1994.

[11] D. Krause and H. Loch, *Mathematical Simulation in Glass Technology*, Springer-Verlag, Berlin, Heidelberg, New York, 2002.

[12] P. Kügler, *A Derivative Free Landweber Method for Parameter Identification in Elliptic Partial Differential Equations with Application to the Manufacture of Car Windshields*, Ph.D. thesis, Institute for Industrial Mathematics, Johannes Kepler University, Linz, Austria, 2003.

[13] P. Kügler, *A parameter identification problem of mixed type related to the manufacture of car windshields*, SIAM J. Appl. Math., 64 (2004), pp. 858–877.

[14] S. Lie, *Gesammelte Abhandlungen*, Band 4, B. G. Teubner, Leipzig, Germany, 1929, pp. 320–384.

[15] E. H. Mansfield, *The Bending and Stretching of Plates*, 2nd ed., Cambridge University Press, Cambridge, UK, 1989.

[16] P. S. Manservisi, *Control and optimization of the sag bending process in glass windshield design*, in Progress in Industrial Mathematics at ECMI98, L. Arkeryd, J. Bergh, P. Brenner, and R. Petterson, eds., Teubner, Stuttgart, Germany, 1999, pp. 97–105.

[17] O. S. Narayanaswamy, *Stress and structural relaxation in tempering glass*, J. Am. Ceramic Soc., 61 (1978), pp. 146–152.

[18] P. J. Olver, *Applications of Lie Groups to Differential Equations*, Grad. Texts in Math. 107, Springer-Verlag, New York, 1986.

[19] P. J. Olver, *Direct reduction and differential constraints*, Proc. Roy. Soc. London Ser. A, 444 (1994), pp. 509–523.

[20] P. J. Olver and P. Rosenau, *The construction of special solutions to partial differential equations*, Phys. Lett. A, 114 (1986), pp. 107–112.

[21] L. V. Ovsiannikov, *Group Analysis of Differential Equations*, W. F. Ames, trans., Academic Press, New York, 1982.

[22] D. Salazar and R. Westbrook, *Inverse problems of mixed type in linear plate theory*, European J. Appl. Math., to appear.

[23] D. Temple, *An Inverse System: An Analysis Arising from Windshield Manufacture*, M.Sc. thesis, Department of Mathematics, Oxford University, Oxford, UK, 2002.

[24] R. Z. Zhdanov, *On conditional symmetries of multidimensional nonlinear equations of quantum field theory*, in Symmetry in Nonlinear Mathematical Physics, Vol. 1, M. Shkil, A. Nikitin, and V. Boyko, eds., Natl. Acad. Sci. Ukraine, Inst. Math., Kiev, 1997, pp. 53–61.

SIAM J. APPL. MATH.
Vol. 65, No. 1, pp. 131–151

FRONT BIFURCATIONS IN AN EXCITATORY NEURAL NETWORK*

PAUL C. BRESSLOFF† AND STEFANOS E. FOLIAS†

Abstract. We show how a one-dimensional excitatory neural network can exhibit a symmetry breaking front bifurcation analogous to that found in reaction diffusion systems. This occurs in a homogeneous network when a stationary front undergoes a pitchfork bifurcation leading to bidirectional wave propagation. We analyze the dynamics in a neighborhood of the front bifurcation using perturbation methods, and we establish that a weak input inhomogeneity can induce a Hopf instability of the stationary front, leading to the formation of an oscillatory front or breather. We then carry out a stability analysis of stationary fronts in an exactly solvable model and use this to derive conditions for oscillatory fronts beyond the weak input regime. In particular, we show how wave propagation failure occurs in the presence of a large stationary input due to the pinning of a stationary front; a subsequent reduction in the strength of the input then generates a breather via a Hopf instability of the front. Finally, we derive conditions for the locking of a traveling front to a moving input, and we show how locking depends on both the amplitude and velocity of the input.

Key words. traveling waves, neural networks, cortical models, front bifurcations, inhomogeneous media

AMS subject classification. 92C20

DOI. 10.1137/S0036139903434481

1. Introduction.

Nonlinear integro-differential equations of the form

$$\tau_s \frac{\partial u(x,t)}{\partial t} = -u(x,t) + \int_{-\infty}^{\infty} w(x-x') f(u(x',t)) dx' - \beta v(x,t) + I(x),$$

$$\frac{1}{\varepsilon}\frac{\partial v(x,t)}{\partial t} = -v(x,t) + u(x,t) \tag{1.1}$$

have arisen as continuum models of one-dimensional cortical tissue [1, 12], in which $u(x,t)$ is a neural field that represents the local activity of a population of excitatory neurons at position $x \in \mathbf{R}$, $I(x)$ is an external input current, τ_s is a synaptic time constant (assuming first-order synapses), $f(u)$ denotes the output firing rate function, and $w(x-x')$ is the strength of connections from neurons at x' to neurons at x. The distribution $w(x)$ is taken to be a positive, even function of x. The neural field $v(x,t)$ represents some form of negative feedback mechanism such as spike frequency adaptation or synaptic depression, with β, ε determining the relative strength and rate of feedback. If additional nonlocal terms in v are introduced, then v represents instead the activity of a population of inhibitory neurons [17, 1]. The nonlinear function f is usually taken to be a smooth sigmoid function

$$f(u) = \frac{1}{1+\mathrm{e}^{-\gamma(u-\kappa)}} \tag{1.2}$$

with gain γ and threshold κ. The units of time are fixed by setting $\tau_s = 1$; a typical value of τ_s is 10 msec. It can be shown [12] that there is a direct link between the above model and experimental studies of wave propagation in cortical slices where synaptic inhibition is pharmacologically blocked [4, 7, 18]. Since there is strong vertical

*Received by the editors September 8, 2003; accepted for publication (in revised form) May 12, 2004; published electronically September 24, 2004. This research was supported by NSF grant DMS-0209824.

http://www.siam.org/journals/siap/65-1/43448.html

†Department of Mathematics, University of Utah, 155 South 1400 East 233 JWB, Salt Lake City, UT 84112 (bressloff@math.utah.edu, sfolias@math.utah.edu).

coupling between cortical layers, it is possible to treat a thin cortical slice as an effective one-dimensional medium. Analysis of the model provides valuable information regarding how the speed of a traveling wave, which is relatively straightforward to measure experimentally, depends on various features of the underlying cortical circuitry.

A number of previous studies have considered the existence and stability of traveling wave solutions of (1.1) in the case of a uniform input I, which is equivalent to a shift in the threshold κ. In particular, it has been shown that in the absence of any feedback ($\beta = 0$), the resulting scalar network can support the propagation of traveling fronts [5, 10], whereas traveling pulses tend to occur when there is significant negative feedback [17, 1, 12]. In this paper, we show that such feedback can also have a nontrivial effect on the propagation of traveling fronts. This is due to the occurrence of a symmetry breaking front bifurcation analogous to that found in reaction diffusion systems [14, 8, 16, 9, 2, 15, 13, 11]. We begin by deriving conditions for the existence of traveling wavefronts in the case of a homogeneous network (section 2). We then carry out a perturbation expansion in powers of the wavespeed c to show that a stationary front can undergo a supercritical pitchfork bifurcation at a critical rate of negative feedback, leading to bidirectional front propagation (section 3). As in the case of reaction diffusion systems, the front bifurcation acts as an organizing center for a variety of nontrivial dynamics including the formation of oscillatory fronts or breathers. We show how the latter can occur through a Hopf bifurcation from a stationary front in the presence of a weak stationary input inhomogeneity (section 4). Finally, we analyze the existence and stability of stationary fronts in an exactly solvable model, which is obtained by taking the high gain limit $\gamma \to \infty$ of the sigmoid function f such that $f(u) = H(u - \kappa)$, where H is the Heaviside function (section 5). As briefly reported elsewhere [3], the exactly solvable model allows us to study oscillatory fronts beyond the weak input regime. Rather than perturbing about the homogeneous case, we now consider a large input amplitude for which wave propagation failure occurs due to the pinning of a stationary front. A subsequent reduction in the amplitude of the input then induces a Hopf instability, leading to the formation of a breather. We conclude our analysis of the exactly solvable model by deriving conditions for the locking of a traveling front to a moving input, and we show how locking depends on both the amplitude and speed of the input.

The major advantage of the exactly solvable model is that it allows us to explicitly determine the existence and stability of stationary and traveling fronts in the presence of external inputs, without any restrictions on the size of the input. However, it has the disadvantage of restricting the nonlinear function f to be a step function. This is less realistic than the smooth nonlinearity (1.2), which matches quite well the input–output characteristics of populations of neurons. The lack of smoothness also makes it difficult to carry out a nonlinear analysis in order to determine whether or not the Hopf instability is supercritical, for example. As we show in this paper, such an analysis can be carried out for smooth f provided that the input amplitude is sufficiently weak. The fact that the nonlocal integro-differential equation (1.1) exhibits behavior similar to a reaction–diffusion system might not be surprising, particularly given that for the exponential weight distribution $w(x) = \mathrm{e}^{-|x|}$, equation (1.1) can be reduced to a PDE of the reaction–diffusion type. It is important to emphasize, however, that our results hold for a more general class of weight distribution $w(x)$ for which a corresponding (finite-order) PDE cannot be constructed. Hence, the analysis is a nontrivial extension of known results for reaction–diffusion equations.

2. Traveling fronts in a homogeneous network. In this section we investigate the existence of traveling front solutions of (1.1) for homogeneous inputs by combining results on scalar networks [5] with an extension of the analysis of front bifurcations in nonscalar reaction–diffusion equations [8, 2].

2.1. The scalar case. The existence of traveling front solutions in scalar, homogeneous networks was previously analyzed by Ermentrout and McLeod [5]. Their analysis can be applied to a scalar version of (1.1) obtained by taking $\varepsilon \to \infty$ so that $v = u$ and setting $I(x) = -h$ with h a constant input. This leads to the scalar integro-differential equation

$$\frac{\partial u(x,t)}{\partial t} = -(1+\beta)u(x,t) + \int_{-\infty}^{\infty} w(x-x')f(u(x',t))dx' - h. \tag{2.1}$$

Without loss of generality we choose h such that $\kappa = 0$ in the sigmoid function (1.2). The weight distribution w is assumed to be a positive, even, continuously differentiable function of x with unit normalization $\int_{-\infty}^{\infty} w(y)dy = 1$. Suppose that the function

$$F_{h,\beta}(u) = f(u) - (1+\beta)u - h \tag{2.2}$$

has precisely three zeros at $u = U_{\pm}(h,\beta), U_0(h,\beta)$ with $U_- < U_0 < U_+$ and $F'_{h,\beta}(U_{\pm}) < 0$. It can then be shown that (modulo uniform translations) there exists a unique traveling front solution of (2.1) such that $u(x,t) = U(\xi)$, $\xi = x - ct$, with $U(\xi) \to U_{\pm}$ as $\xi \to \mp\infty$ [5]. Moreover, the speed of the wave satisfies

$$c = c(h,\beta) = \frac{\Gamma_{h,\beta}}{\int_{-\infty}^{\infty} u'^2 f'(u)d\xi}, \tag{2.3}$$

where

$$\Gamma_{h,\beta} = \int_{U_-}^{U_+} F_{h,\beta}(u)du. \tag{2.4}$$

Since the denominator of (2.3) is positive definite, the sign of c is determined by the sign of the coefficient $\Gamma_{h,\beta}$. In particular, suppose that $h = 0.5$ and f is given by the sigmoid function (1.2) so that $f(u) - h = \tanh(u/2\gamma)/2$. It follows that, for $0 < 1+\beta < \gamma/4$, there exists a pair of stable homogeneous fixed points with $U_- = -U_+$, which in turn implies that $\Gamma_{h,\beta} = 0$ and the front solution is stationary; see Figure 2.1. The corresponding function $F_{h,\beta}(u)$ has the inflection symmetry $F_{h,\beta}(-u) = -F_{h,\beta}(u)$. Note that the stationary solution of (2.1) is also an ε-independent solution of the full system (1.1) with $I(x) = -h$, but it is not necessarily the only solution (see below).

2.2. The regime $\varepsilon \gg 1$. In the large ε regime, the neural field v varies on a much faster time scale than u. Introducing the stretched time coordinate $\tau = t/\delta$ with $\delta = \varepsilon^{-1} \ll 1$, we have

$$\frac{\partial u(x,\tau)}{\partial \tau} = \delta\left(-u(x,\tau) + \int_{-\infty}^{\infty} w(x-x')f(u(x',\tau))dx' - \beta v(x,\tau) - h\right),$$

$$\frac{\partial v(x,\tau)}{\partial \tau} = -v(x,\tau) + u(x,\tau). \tag{2.5}$$

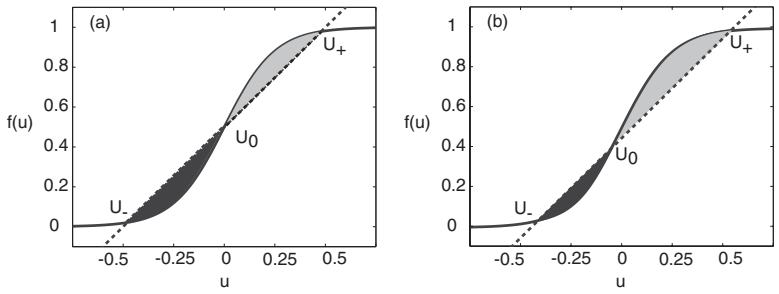


FIG. 2.1. *Balance condition for the speed of a traveling wavefront in a scalar excitatory network with* $u(x,t) = U(x-ct)$ *such that* $U(\mp\infty) = U_\pm$. *The solid curve is* $f(u) = 1/(1+\mathrm{e}^{-\gamma u})$ *with* $\gamma = 8$, *and the dashed line is* $g(u) = (1+\beta)u + h$. *The wavespeed* c *is positive (negative) if the gray shaded area is larger (smaller) than the black shaded area.* (a) $h = 0.5$, $\beta = 0.5$ *such that* $c = 0$. (b) $h = 0.4$, $\beta = 0.5$ *such that* $c > 0$.

To leading order in δ, u is independent of τ so that we can explicitly solve for v according to

$$v(x,t) = v_0(x)\mathrm{e}^{-\varepsilon t} + u(x,t)(1-\mathrm{e}^{-\varepsilon t}). \tag{2.6}$$

Thus after an initial transient of duration $t \sim \mathcal{O}(\varepsilon^{-1})$, the field v adiabatically follows the field u, with the latter evolving according to the scalar equation (2.1). It follows that in the large ε regime there exists a unique traveling wave solution of the full system with $(u(x,t), v(x,t)) = (U(x-ct), V(x-ct))$ such that $(U,V) \to (U_\pm, U_\pm)$ as $\xi \to \mp\infty$ and $c = c(h,\beta)$, $U_\pm = U_\pm(h,\beta)$. The front is stable in the large ε regime provided that the solution of the corresponding scalar equation is stable, which is found to be the case numerically. If $\Gamma_{h,\beta} = 0$, then the front is stationary and persists for all ε but may become unstable as ε is reduced.

2.3. The regime $0 < \varepsilon \ll 1$. In the small ε regime, additional front solutions can be constructed that connect the two fixed points $(u,v) = (U_\pm(h,\beta), U_\pm(h,\beta))$. This follows from the observation that the neural field v remains approximately constant on the length scale over which u varies, that is, within the transition layer of the front. Suppose that the system is prepared in the down state (U_-, U_-) and is perturbed on its left-hand side to induce a transition to the upper state (U_+, U_+). In this case $v \approx U_-$ within the transition layer, and this generates a front propagating to the right whose speed is approximately given by (2.3) with $h \to h + \beta U_-$, that is, $c = c(h + \beta U_-, 0)$. If, on the other hand, the system is prepared in the up state (U_+, U_+) and is perturbed on its right-hand side to induce a transition to the down state (U_-, U_-), then a left-propagating front is generated with $c = c(h + \beta U_+, 0)$. Note from (2.4) that

$$\Gamma_{h+\beta U_-,0} > \Gamma_{h,\beta} + \beta \int_{U_-}^{U_+} (u - U_-)du, \quad \Gamma_{h+\beta U_+,0} < \Gamma_{h,\beta} + \beta \int_{U_-}^{U_+} (u - U_+)du \tag{2.7}$$

so that $\Gamma_{h+\beta U_-,0} > \Gamma_{h,\beta} > \Gamma_{h+\beta U_+,0}$. Hence, the existence of fronts propagating in opposite directions clearly holds when h, β are chosen such that $\Gamma_{h,\beta} = 0$.

3. Front bifurcation. The above analysis suggests that if $\Gamma_{h,\beta} = 0$, then at some critical rate of feedback $\varepsilon = \varepsilon_c$, a pair of counterpropagating fronts bifurcate

from a stationary front. Moreover, all the front solutions have the same asymptotic behavior $(U(\xi),V(\xi)) \to (U_\pm,U_\pm)$ as $\xi \to \mp\infty$. Following along lines analogous to Hagberg and Meron [8], we carry out a perturbation expansion in powers of the speed c about this critical point, and we show that the stationary solution undergoes a pitchfork bifurcation.

First, set $I(x) = -h$ and $(u(x,t),v(x,t)) = (U(x-ct),V(x-ct))$ in (1.1) to obtain the pair of equations

$$\begin{aligned} -cU' &= -U + w * f(U) - \beta V, \\ -cV' &= \varepsilon[-V+U], \end{aligned} \tag{3.1}$$

where $U' = dU/d\xi$ and $*$ denotes the convolution operator,

$$w * U = \int_{-\infty}^{\infty} w(\xi-\xi')U(\xi')d\xi'. \tag{3.2}$$

Suppose that β and h are fixed such that $\Gamma_{h,\beta} = 0$, and denote the corresponding stationary solution by $(\overline{U},\overline{V})$. Expand the fields U,V as power series in c:

$$\begin{aligned} U(\xi) &= \overline{U}(\xi) + cU_1(\xi) + c^2U_2(\xi) + \cdots, \\ V(\xi) &= \overline{V}(\xi) + cV_1(\xi) + c^2V_2(\xi) + \cdots. \end{aligned} \tag{3.3}$$

Note that the higher order terms $U_n(\xi), V_n(\xi)$, $n \geq 1$, should all decay to zero as $\xi \to \pm\infty$, since the stationary solution already has the correct asymptotic behavior. Also expand ε according to

$$\varepsilon = \varepsilon_c + c\varepsilon_1 + c^2\varepsilon_2 + \cdots. \tag{3.4}$$

Substitute these expansions into (3.1) and Taylor expand the nonlinear function $f(U)$ about $\overline{U}$:

$$f(U) = f(\overline{U}) + \sum_{n\geq 1} \overline{f}_n(U-\overline{U})^n, \quad \overline{f}_n = \frac{1}{n!}\frac{d^nf}{dU^n}\bigg|_{U=\overline{U}}. \tag{3.5}$$

Collecting all terms at successive orders of c then generates a hierarchy of equations for the perturbative corrections U_n, V_n. The lowest order equation recovers the conditions for a stationary solution:

$$\begin{aligned} (1+\beta)\overline{U} + h &= w * f(\overline{U}), \\ \overline{V} &= \overline{U}. \end{aligned} \tag{3.6}$$

At order c we have

$$\begin{aligned} -\overline{U}' &= -U_1 + w * [\overline{f}_1U_1] - \beta V_1, \\ -\overline{V}' &= \varepsilon_c[-V_1+U_1] + \varepsilon_c[-\overline{V}+\overline{U}]. \end{aligned} \tag{3.7}$$

The term $-\beta V_1$ in the first line can be eliminated using the second. Since $\overline{V} = \overline{U}$, we thus find that

$$\mathcal{M}U_1 = \left(\frac{\beta}{\varepsilon_c} - 1\right)\overline{U}', \quad V_1 = U_1 + \frac{\overline{U}'}{\varepsilon_c}, \tag{3.8}$$

where $\mathcal{M}$ is the linear operator

$$\mathcal{M}U = -(1+\beta)U + w * [\overline{f}_1 U]. \tag{3.9}$$

Since the functions $U_n(\xi), V_n(\xi)$ decay to zero as $\xi \to \pm\infty$, we will assume that $\mathcal{M}$ acts on the space $L^2(\mathbf{R})$ and introduce the generalized inner product

$$\langle U|V\rangle = \int_{-\infty}^{\infty} f'(\overline{U}(\xi))U(\xi)V(\xi)d\xi \tag{3.10}$$

for all $U, V \in L^2(\mathbf{R})$. With respect to this space, $\mathcal{M}$ is self-adjoint and has the null vector $\overline{U}'$[1]:

$$\mathcal{M}\overline{U}' = \mathcal{M}^\dagger \overline{U}' = 0. \tag{3.11}$$

Applying the Fredholm alternative to (3.8) then gives the solvability condition

$$\langle \overline{U}'|\overline{U}'\rangle \left(\frac{\beta}{\varepsilon_c} - 1\right) = 0. \tag{3.12}$$

Since $f'(\overline{U}(\xi)) > 0$ for all ξ, it follows that $\langle \overline{U}'|\overline{U}'\rangle > 0$ and thus $\varepsilon_c = \beta$. This in turn means that $\mathcal{M}U_1 = 0$ and hence $U_1 = A\overline{U}'$ for some constant A. Since $\overline{U}'$ is the generator of uniform translations, we are free to choose the origin such that $A = 0$. Under this choice,

$$U_1 = 0, \quad V_1 = \frac{\overline{U}'}{\varepsilon_c}. \tag{3.13}$$

At order c^2 we obtain

$$\begin{aligned} -U_1' &= \mathcal{M}U_2 + \beta[-V_2 + U_2] + w * [\overline{f}_2 U_1^2], \\ -V_1' &= \varepsilon_c[-V_2 + U_2] + \varepsilon_1[-V_1 + U_1] + \varepsilon_2[-\overline{V} + \overline{U}]. \end{aligned} \tag{3.14}$$

Substituting for $-V_2 + U_2$ in the first line, taking $\overline{V} = \overline{U}$, $\beta = \varepsilon_c$, and using equation (3.13) then gives

$$\mathcal{M}U_2 = \frac{1}{\varepsilon_c}\left(\overline{U}'' - \varepsilon_1 \overline{U}'\right), \quad V_2 = U_2 + \frac{1}{\varepsilon_c^2}(\overline{U}'' - \varepsilon_1 \overline{U}'). \tag{3.15}$$

Applying the Fredholm alternative to (3.15) yields the solvability condition

$$\langle \overline{U}'|\overline{U}''\rangle = \varepsilon_1 \langle \overline{U}'|\overline{U}'\rangle. \tag{3.16}$$

In order to evaluate the inner product $\langle \overline{U}'|\overline{U}''\rangle$, we use the result

$$(1+\beta)\frac{d^2\overline{U}}{d\xi^2} = \int_{-\infty}^{\infty} w(\xi - \xi')\frac{d^2 f(\overline{U}(\xi'))}{d{\xi'}^2}d\xi', \tag{3.17}$$

[1] We could equally well proceed by taking the standard inner product $\langle U|V\rangle = \int_{-\infty}^{\infty} U(\xi)V(\xi)d\xi$. The adjoint of $\mathcal{M}$ is then given by $\mathcal{M}^\dagger U = -(1+\beta)U + \overline{f}_1 w * U$, which has the null vector $\overline{f}_1\overline{U}'$ where $\overline{f}_1 = f'(\overline{U})$.

which follows from differentiating (3.6) with respect to ξ and using the asymptotic properties of w. Then

$$\begin{aligned}\langle \overline{U}' | \overline{U}'' \rangle &= \int_{-\infty}^{\infty} f'(\overline{U}(\xi))\overline{U}'(\xi)\overline{U}''(\xi)d\xi \\ &= \int_{-\infty}^{\infty} \frac{df(\overline{U}(\xi))}{d\xi}\overline{U}''(\xi)d\xi \\ &= \frac{1}{1+\beta}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{df(\overline{U}(\xi))}{d\xi} w(\xi-\xi')\frac{d^2 f(\overline{U}(\xi'))}{d\xi'^2}d\xi' d\xi \\ &= \frac{1}{1+\beta}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{df(\overline{U}(\xi))}{d\xi} w'(\xi-\xi')\frac{df(\overline{U}(\xi'))}{d\xi'}d\xi' d\xi \\ &= 0, \end{aligned} \tag{3.18}$$

since $w'(\xi)$ is an odd function of ξ. Hence, $\varepsilon_1 = 0$ and

$$\mathcal{M}U_2 = \frac{\overline{U}''}{\varepsilon_c}, \quad V_2 = U_2 + \frac{\overline{U}''}{\varepsilon_c^2}. \tag{3.19}$$

At order c^3 we obtain

$$\begin{aligned} -U_2' &= \mathcal{M}U_3 + \beta[-V_3 + U_3] + 2w * [\overline{f}_2 U_1 U_2] + w * [\overline{f}_3 U_1^3], \\ -V_2' &= \varepsilon_c[-V_3 + U_3] + \varepsilon_1[-V_2 + U_2] + \varepsilon_2[-V_1 + U_1] + \varepsilon_3[-\overline{V} + \overline{U}]. \end{aligned} \tag{3.20}$$

Substituting for $-V_2 + U_2$ in the first line, taking $\overline{V} = \overline{U}$, $\beta = \varepsilon_c$, $\varepsilon_1 = 0$, and using (3.13) and (3.19) then gives

$$\mathcal{M}U_3 = \frac{1}{\varepsilon_c^2}\left(\overline{U}''' - \varepsilon_2\varepsilon_c\overline{U}'\right), \quad V_3 = U_3 + \frac{1}{\varepsilon_c^3}(\overline{U}''' + \varepsilon_c^2 U_2' - \varepsilon_2\varepsilon_c\overline{U}'). \tag{3.21}$$

Applying the Fredholm alternative to (3.21) yields the solvability condition

$$\varepsilon_2 = \frac{\langle \overline{U}' | \overline{U}''' \rangle}{\varepsilon_c \langle \overline{U}' | \overline{U}' \rangle} < 0. \tag{3.22}$$

The sign of ε_2 can be determined using (3.17),

$$\begin{aligned}\langle \overline{U}' | \overline{U}''' \rangle &= \int_{-\infty}^{\infty} f'(\overline{U}(\xi))\overline{U}'(\xi)\overline{U}'''(\xi)d\xi \\ &= \int_{-\infty}^{\infty} \frac{df(\overline{U}(\xi))}{d\xi}\overline{U}'''(\xi)d\xi \\ &= -\int_{-\infty}^{\infty} \frac{d^2 f(\overline{U}(\xi))}{d\xi^2}\overline{U}''(\xi)d\xi \\ &= -\frac{1}{1+\beta}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{d^2 f(\overline{U}(\xi))}{d\xi^2} w(\xi-\xi')\frac{d^2 f(\overline{U}(\xi'))}{d\xi'^2}d\xi' d\xi \\ &< 0, \end{aligned} \tag{3.23}$$

since $w(\xi)$ is an even, monotonically decreasing function of $|\xi|$. Hence $\varepsilon_2 < 0$.

Combining these various results, we find that

$$\begin{aligned} U(\xi) &= \overline{U}(\xi) + \mathcal{O}(c^2), \\ V(\xi) &= \overline{U}(\xi) + \frac{c}{\varepsilon_c}\overline{U}'(\xi) + \mathcal{O}(c^2), \end{aligned} \tag{3.24}$$

and

$$\varepsilon = \varepsilon_c + c^2\varepsilon_2 + \mathcal{O}(c^3). \tag{3.25}$$

Equation (3.25) implies that the stationary front undergoes a pitchfork bifurcation, which is supercritical since $\varepsilon_2 < 0$. (This assumes of course that the stationary front is stable for $\varepsilon > \varepsilon_c$. This can be confirmed numerically, and also proven analytically in the high gain limit; see section 5.) Close to the bifurcation point the shape of the propagating fronts is approximately the same as the stationary front, except that the recovery variable V is shifted relative to U by an amount proportional to the speed c, that is,

$$U(\xi) \approx \overline{U}(\xi), \quad V(\xi) \approx \overline{U}(\xi + c/\varepsilon_c). \tag{3.26}$$

An analogous result was previously obtained for reaction–diffusion equations [8]. It is important to emphasize that the occurrence of a pitchfork bifurcation from a stationary front does not require any underlying inflection symmetries of the nonlinear function f (see also [2]). We only require that the scalar equation (2.1) supports a stationary front for appropriate choices of h, β. The fact that the weight distribution $w(x)$ is even means that there must be a pitchfork bifurcation from a stationary solution rather than a transcritical bifurcation as in the case of a nonsymmetric w.

4. The effect of a weak input inhomogeneity. Now suppose that both ε and h are allowed to vary. We then expect a codimension 2 cusp bifurcation in which the pitchfork bifurcation unfolds into a saddle-node bifurcation, with the stationary front replaced by a traveling front in the large ε regime. More interestingly, as in the case of reaction–diffusion systems [16, 9, 2], the pitchfork bifurcation acts as an organizing center for a variety of dynamical phenomena, including the formation of breathers due to the presence of a weak input inhomogeneity or due to curvature (in the case of two spatial dimensions). These breathers consist of periodic reversals in propagation that can be understood in terms of a dynamic transition between the pair of counterpropagating fronts that is induced by the weak intrinsic perturbation. Such a transition involves an interaction between a translational degree of freedom and an order parameter that determines the direction of propagation. In order to unravel this interaction, it is necessary to extend the perturbation analysis of section 3 along lines analogous to previous treatments of reaction–diffusion systems [16, 9, 2].

Suppose that the system (1.1) undergoes a pitchfork bifurcation from a stationary state when $\varepsilon = \varepsilon_c = \beta$ and $I(x) = -h$. Introduce the small parameter δ according to $\varepsilon - \varepsilon_c = \delta^2\chi$ and introduce a weak input inhomogeneity by taking $I(x) = -h + \delta^3\eta(x)$. Since any fronts are slowly propagating, we rescale time according to $\tau = \delta t$ so that (1.1) becomes

$$\begin{aligned} \delta\frac{\partial u(x,\tau)}{\partial \tau} &= -u(x,t) + \int_{-\infty}^{\infty} w(x-x')f(u(x',\tau))dx' - \beta v(x,\tau) - h + \delta^3\eta(x), \\ \delta\frac{\partial v(x,\tau)}{\partial \tau} &= (\varepsilon_c + \delta^2\chi)\left[-v(x,\tau) + u(x,\tau)\right]. \end{aligned} \tag{4.1}$$

Motivated by (3.24), we introduce the ansatz that, sufficiently close to the pitchfork bifurcation, the solutions of (4.1) can be expanded in the form

$$\begin{aligned} u(x,\tau) &= \overline{U}(x-p(\tau)) + \delta^2 u_2(x,\tau) + \delta^3 u_3(x,\tau) + \cdots, \\ v(x,\tau) &= \overline{U}(x-p(\tau)) + \delta\frac{a(\delta\tau)}{\varepsilon_c}\overline{U}'(x-p(\tau)) + \delta^2 v_2(x,\tau) + \delta^3 v_3(x,\tau) + \cdots. \end{aligned} \tag{4.2}$$

Here p is identified with the translational degree of freedom, whereas a represents the order parameter associated with changes in propagation direction. Note that a is assumed to evolve on a slower time scale than p. We now substitute the ansatz (4.2) into (4.1) and expand in powers of δ along lines similar to the perturbation calculation of section 3.

At order δ we find that

$$p_\tau = a, \tag{4.3}$$

where $p_\tau = dp/d\tau$. At order δ^2 we obtain the pair of equations

$$\mathcal{M}u_2 = a^2\frac{\overline{U}''}{\varepsilon_c}, \qquad v_2 = u_2 + a^2\frac{\overline{U}''}{\varepsilon_c^2} \tag{4.4}$$

after setting $p_\tau = a$. The solvability condition for (4.4) is automatically satisfied. At order δ^3 we have

$$\begin{aligned} \frac{\partial u_2}{\partial\tau} &= \mathcal{M}u_3 + \beta[-v_3 + u_3] + \eta, \\ \frac{\partial v_2}{\partial\tau} + \frac{\overline{U}'a_{\hat\tau}}{\varepsilon_c} &= \varepsilon_c[-v_3 + u_3] - a\chi\frac{\overline{U}'}{\varepsilon_c} \end{aligned} \tag{4.5}$$

with $\hat\tau = \delta\tau$. Using (4.4), the following equation for u_3 is obtained:

$$\mathcal{M}u_3 = \frac{1}{\varepsilon_c^2}\left(a^3\overline{U}''' - a\chi\varepsilon_c\overline{U}' - a_{\hat\tau}\varepsilon_c\overline{U}'\right) - \eta. \tag{4.6}$$

Applying the Fredholm alternative to (4.6) yields an amplitude equation for a:

$$a_{\hat\tau} = -\chi a + a^3\frac{\langle\overline{U}'|\overline{U}'''\rangle}{\varepsilon_c\langle\overline{U}'|\overline{U}'\rangle} - \varepsilon_c\frac{\langle\overline{U}'|\eta\rangle}{\langle\overline{U}'|\overline{U}'\rangle}. \tag{4.7}$$

Finally, rescaling p, a, and η, we obtain the pair of equations

$$\begin{aligned} p_t &= a, \\ a_t &= (\varepsilon_c - \varepsilon)a + \frac{\langle\overline{U}'|\overline{U}'''\rangle}{\varepsilon_c\langle\overline{U}'|\overline{U}'\rangle}a^3 - \varepsilon_c\frac{\langle\overline{U}'|\eta\rangle}{\langle\overline{U}'|\overline{U}'\rangle}. \end{aligned} \tag{4.8}$$

Note that $\overline{U} = \overline{U}(x-p)$, so that the final coefficient on the right-hand side of (4.8) will be p-dependent in the case of an inhomogeneous input $\eta = \eta(x)$.

Cusp bifurcation for homogeneous inputs. It is clear from (4.8) that when $\eta = 0$ we recover the pitchfork bifurcation of a stationary front as found in section 3. In particular, for $\varepsilon < \varepsilon_c$ there are three constant speed solutions of (4.8) such that

$a_t = 0, P_t = a = c$, corresponding to an unstable stationary front and a pair of stable counterpropagating fronts with speeds

$$c = \pm\sqrt{(\varepsilon_c - \varepsilon)\varepsilon_c \frac{\langle \overline{U}' | \overline{U}' \rangle}{|\langle \overline{U}' | \overline{U}''' \rangle|}}. \tag{4.9}$$

If η is nonzero but constant, on the other hand, the final term on the right-hand side of (4.8) reduces to the constant coefficient $\varepsilon_c \eta (f(U_+) - f(U_-))/\langle \overline{U}' | \overline{U}' \rangle$, and the pitchfork bifurcation unfolds to a saddle-node bifurcation. There are two saddle-node lines in the (η, ε)-plane corresponding to the condition $dG(a)/da = 0$, where $a_t = G(a)$:

$$\eta_{sn} = \pm \frac{2}{3\sqrt{3}} \frac{(\varepsilon_c - \varepsilon)^{3/2}}{\varepsilon_c^{1/2}} \frac{\langle \overline{U}' | \overline{U}' \rangle^{3/2}}{(f(U_+) - f(U_-))|\langle \overline{U}' | \overline{U}''' \rangle|^{1/2}}, \tag{4.10}$$

and the corresponding speed along these lines is

$$c_{sn} = \pm\sqrt{(\varepsilon_c - \varepsilon)\varepsilon_c \frac{\langle \overline{U}' | \overline{U}' \rangle}{3|\langle \overline{U}' | \overline{U}''' \rangle|}}. \tag{4.11}$$

Hopf bifurcation for a weak inhomogeneity. The introduction of a weak input inhomogeneity can lead to a Hopf instability of the stationary front. We shall illustrate this by considering the particular example of the step inhomogeneity

$$\eta(x) = \begin{cases} s/2 & \text{if } x \leq 0, \\ -s/2 & \text{if } x > 0 \end{cases} \tag{4.12}$$

with $s > 0$. For such an input we find that

$$\langle \overline{U}' | \eta \rangle = \frac{s}{2}[2f(\overline{U}(-p)) - f(U_+) - f(U_-)]. \tag{4.13}$$

Recall from section 2 that when $h = 0.5$ the homogeneous network with f given by (1.2) supports a stationary front solution for which $U_\pm = \pm 0.5/(1+\beta)$, and $\overline{U}(0) = 0$ such that $f(U_+) + f(U_-) = 2f(0)$. Hence, (4.8) has a fixed point at $p = 0, a = 0$. Linearization about this fixed point shows that there is a Hopf bifurcation of the stationary front at $\varepsilon = \varepsilon_c$ with Hopf frequency

$$\omega_H = \sqrt{\frac{s\varepsilon_c f'(0)|\overline{U}'(0)|}{\langle \overline{U}' | \overline{U}' \rangle}}. \tag{4.14}$$

The supercritical or subcritical nature of the Hopf bifurcation can then be determined by evaluating higher order terms in a, p. However, this is complicated by the fact that we do not have an analytical expression for the stationary front solution $\overline{U}$, in contrast to the case of a reaction–diffusion equation with a cubic nonlinearity [2]. (Note that, as in the case of reaction–diffusion equations [2], one can develop a more intricate perturbation analysis that takes into account $\mathcal{O}(\delta^2)$ inhomogeneities and corresponding shifts in the Hopf bifurcation point. Here we have followed a simpler approach in order to illustrate the basic ideas underlying the perturbative treatment of the integro-differential equation (1.1).)

5. Exactly solvable model. We now consider the high gain limit $\gamma \to \infty$, for which (1.2) reduces to $f(u) = H(u-\kappa)$, where H is the Heaviside function $H(u) = 1$ if $u > 0$ and $H(u) = 0$ if $u \le 0$. The advantage of using a threshold nonlinearity is that explicit analytical expressions for front solutions can be obtained, which allows us to derive conditions for the Hopf instability of a stationary front without any restrictions on the size of the input inhomogeneity. Numerical simulations of the full system establish that the bifurcation is supercritical and that it generates an oscillatory modulation of the stationary front in the form of a breather [3]. (For a corresponding analysis of reaction–diffusion equations, see Prat and Li [13].)

5.1. Traveling fronts (homogeneous case). We begin by deriving exact traveling front solutions of (1.1) for $f(u) = H(u-\kappa)$ and a homogeneous input $I(x) = 0$. That is, we seek a solution of the form $u(x,t) = U(\xi)$, $\xi = x - ct$, $c > 0$, such that $U(0) = \kappa$, $U(\xi) < \kappa$ for $\xi > 0$ and $U(\xi) > \kappa$ for $\xi < 0$. Setting $v(x,t) = V(\xi)$, we then have

$$-cU'(\xi) + U(\xi) = \int_{-\infty}^{0} w(\xi - \xi')d\xi' - \beta V(\xi), \tag{5.1}$$

$$-\frac{c}{\varepsilon}V'(\xi) = -V(\xi) + U(\xi). \tag{5.2}$$

Differentiating the first equation and substituting into the second, we obtain a second-order ODE with boundary conditions at $\xi = 0$ and $\pm\infty$:

$$\begin{aligned} -c^2U''(\xi) + c[1+\varepsilon]U'(\xi) - \varepsilon[1+\beta]U(\xi) &= -cw(\xi) - \varepsilon W(\xi), \\ U(0) &= \kappa, \\ U(\mp\infty) &= U_{\pm}, \end{aligned} \tag{5.3}$$

where

$$W(\xi) = \int_{\xi}^{\infty} w(y)dy. \tag{5.4}$$

Here $U_{\pm}$ are the homogeneous fixed point solutions

$$U_+ = \frac{1}{1+\beta}, \quad U_- = 0. \tag{5.5}$$

We have used the fact that w has unit normalization, $W(-\infty) \equiv \int_{-\infty}^{\infty} w(y)dy = 1$. It follows that a necessary condition for the existence of a front solution is $\kappa < U_+$.

In order to establish the existence of a traveling front, we solve the boundary value problem in the domains $\xi \le 0$ and $\xi \ge 0$ and match the solutions at $\xi = 0$. For further mathematical convenience, we take the weight distribution to be an exponential function

$$w(x) = \frac{1}{2d}\mathrm{e}^{-|x|/d}, \tag{5.6}$$

where d determines the range of the synaptic interactions. We fix the spatial scale by setting $d = 1$; a typical value of d is 1 mm. We first consider the case of right-moving

waves ($c > 0$). On the domain $\xi \geq 0$, the particular solution is $U_>(\xi) = \kappa e^{-\xi}$, with κ related to the speed c according to the self-consistency condition

$$\kappa = \frac{c+\varepsilon}{2(c^2 + c[1+\varepsilon] + \varepsilon[1+\beta])}, \qquad c \geq 0. \tag{5.7}$$

In the domain $\xi \leq 0$ the solution consists of complementary and particular parts:

$$U_<(\xi) = \mathcal{A}_+ e^{\mu_+ \xi} + \mathcal{A}_- e^{\mu_- \xi} + \mathcal{A} e^{\xi} + U_+, \tag{5.8}$$

where

$$\mu_\pm = \frac{1}{2c}\left[1 + \varepsilon \pm \sqrt{(1+\varepsilon)^2 - 4\varepsilon(1+\beta)}\right]. \tag{5.9}$$

The coefficient $\mathcal{A}$ is obtained by direct substitution into the differential equation for U, whereas the coefficients $\mathcal{A}_\pm$ are determined by matching solutions at the boundary $\xi = 0$, that is, $U_<(0) = \kappa$ and $U_<'(0) = -\kappa$. Thus we find

$$\mathcal{A} = \frac{c-\varepsilon}{2(c^2 - c[1+\varepsilon] + \varepsilon[1+\beta])}, \tag{5.10}$$

$$\mathcal{A}_+ = \frac{\mu_- U_+ + (\mu_- - 1)\mathcal{A} - (1+\mu_-)\kappa}{\mu_+ - \mu_-}, \tag{5.11}$$

$$\mathcal{A}_- = \frac{-\mu_+ U_+ + (1-\mu_+)\mathcal{A} + (1+\mu_+)\kappa}{\mu_+ - \mu_-}. \tag{5.12}$$

In the limit $\beta \to 0$ we recover the standard result for an excitatory network without feedback [5]:

$$U(\xi) = \begin{cases} \dfrac{1}{2(c+1)} e^{-\xi} & \text{for } \xi > 0, \\ 1 + (\kappa - 1)e^{\xi/c} + \dfrac{1}{2(c-1)}\left[e^{\xi} - e^{\xi/c}\right] & \text{for } \xi < 0 \end{cases} \tag{5.13}$$

with

$$\kappa = \frac{1}{2(c+1)}, \qquad c \geq 0. \tag{5.14}$$

A similar analysis can be carried out for left-moving waves. Now the speed c is determined by the particular solution in the domain $\xi \leq 0$, which takes the form $U_<(\xi) = -\hat{\kappa} e^{\xi} + U_+$ with $\hat{\kappa} = (1+\beta)^{-1} - \kappa$. This leads to the self-consistency condition

$$\hat{\kappa} = -\frac{c-\varepsilon}{2(c^2 - c[1+\varepsilon] + \varepsilon[1+\beta])}, \qquad c \leq 0. \tag{5.15}$$

The existence of traveling front solutions can now be established by finding positive real solutions of (5.7) and negative real solutions of (5.15). For concreteness, we will assume that the threshold κ is fixed and determine the solution branches as a

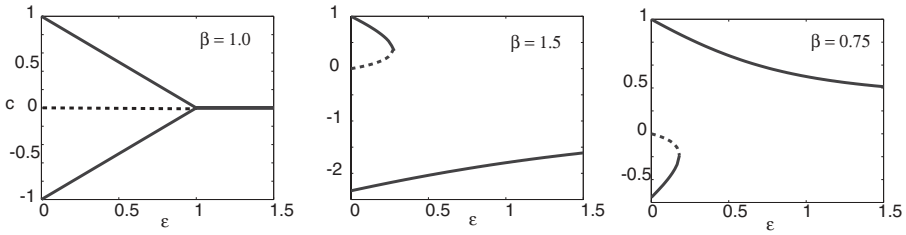


FIG. 5.1. *Plot of wavefront speed c as a function of ε for various values of β and a fixed threshold $\kappa = 0.25$: (i) $2\kappa(1+\beta) = 1$, (ii) $2\kappa(1+\beta) > 1$, (iii) $2\kappa(1+\beta) < 1$. Stable (unstable) branches are shown as solid (dashed) curves.*

function of the feedback parameters ε, β with $1/\kappa - 1 > \beta > 0$. The roots of (5.7) and (5.15) can be written explicitly as

$$c = \frac{1}{2}\left[-\left(1+\varepsilon-\frac{1}{2\kappa}\right) \pm \sqrt{\left(1+\varepsilon-\frac{1}{2\kappa}\right)^2 - 4\varepsilon\left(1+\beta-\frac{1}{2\kappa}\right)}\right] \tag{5.16}$$

and

$$c = \frac{1}{2}\left[\left(1+\varepsilon-\frac{1}{2\hat{\kappa}}\right) \pm \sqrt{\left(1+\varepsilon-\frac{1}{2\hat{\kappa}}\right)^2 - 4\varepsilon\left(1+\beta-\frac{1}{2\hat{\kappa}}\right)}\right]. \tag{5.17}$$

Using the fact that $\operatorname{sign}\left(1+\beta-\frac{1}{2\kappa}\right) = -\operatorname{sign}\left(1+\beta-\frac{1}{2\hat{\kappa}}\right)$, we find that there are three bifurcation scenarios, as shown in Figure 5.1:

(i) If $2\kappa(1+\beta) = 1$, then there exists a stationary front for all ε. At a critical value of ε the stationary front undergoes a pitchfork bifurcation, leading to the formation of a left- and a right-moving wave. This is the high gain limit of the front bifurcation analyzed in section 3 for smooth f.

(ii) If $2\kappa(1+\beta) > 1$, then there is a single left-moving wave for all ε. There also exists a pair of right-moving waves that annihilate in a saddle-node bifurcation at a critical value of ε that approaches zero as $\beta \to 0$.

(iii) If $2\kappa(1+\beta) < 1$, then there is a single right-moving wave for all ε. There also exists a pair of left-moving waves that annihilate in a saddle-node bifurcation at a critical value of ε that approaches zero as $\beta \to 0$.

5.2. Stability analysis of stationary fronts (inhomogeneous case). Stationary front solutions of (1.1) with $f(u) = H(u-\kappa)$ in the case of an inhomogeneous input $I(x)$ satisfy the equation

$$(1+\beta)U(x) = \int_{-\infty}^{x_0} w(x-x')dx' + I(x). \tag{5.18}$$

Suppose that $I(x)$ is a monotonically decreasing function of x. Since the system is no longer translation invariant, the position of the front is pinned to a particular location x_0, where $U(x_0) = \kappa$. Monotonicity of $I(x)$ ensures that $U(x) > \kappa$ for $x < x_0$ and $U(x) < \kappa$ for $x > x_0$. The center x_0 satisfies

$$(1+\beta)\kappa = \frac{1}{2} + I(x_0) \tag{5.19}$$

under the normalization $\int_0^\infty w(y)dy = 1/2$. Equation (5.19) implies that in contrast to the homogeneous case, there exists a stationary front over a range of threshold values (for fixed β); changing the threshold κ simply shifts the position of the center x_0. In the particular case of the exponential weight distribution (5.6), we have

$$(1+\beta)U(x) = \begin{cases} \dfrac{\mathrm{e}^{x_0-x}}{2} + I(x), & x > x_0, \\ 1 - \dfrac{\mathrm{e}^{x-x_0}}{2} + I(x), & x < x_0. \end{cases} \tag{5.20}$$

If the stationary front is stable, then it will prevent wave propagation. Stability is determined by writing $u(x,t) = U(x) + p(x,t)$ and $v(x,t) = V(x) + q(x,t)$ with $V(x) = U(x)$ and expanding (1.1) to first-order in $(p,\ q)$:

$$\frac{\partial p(x,t)}{\partial t} = -p(x,t) + \int_{-\infty}^{\infty} w(x-x')H'(U(x'))p(x',t)dx' - \beta q(x,t),$$

$$\frac{1}{\varepsilon}\frac{\partial q(x,t)}{\partial t} = -q(x,t) + p(x,t). \tag{5.21}$$

We assume that $p, q \in L^2(\mathbf{R})$. The spectrum of the associated linear operator is found by taking $p(x,t) = \mathrm{e}^{\lambda t}p(x)$ and $q(x,t) = \mathrm{e}^{\lambda t}q(x)$. Using the identity

$$\frac{dH(U(x))}{dU} = \frac{\delta(x-x_0)}{|U'(x_0)|} \tag{5.22}$$

we obtain the equation

$$(\lambda+1)p(x) = \frac{w(x-x_0)}{|U'(x_0)|}p(x_0) - \frac{\varepsilon\beta p(x)}{\lambda+\varepsilon}. \tag{5.23}$$

Equation (5.23) has two classes of solution. The first consists of any function $p(x)$ such that $p(x_0) = 0$, for which $\lambda = \lambda_\pm^{(0)}$, where

$$\lambda_\pm^{(0)} = \frac{-(1+\varepsilon) \pm \sqrt{(1+\varepsilon)^2 - 4\varepsilon(1+\beta)}}{2}. \tag{5.24}$$

Note that $\lambda_\pm^{(0)}$ belong to the essential spectrum since they have infinite multiplicity. The second class of solution is of the form $p(x) = Aw(x-x_0)$, $A \neq 0$, for which λ is given by the roots of the equation

$$\lambda + 1 + \frac{\varepsilon\beta}{\lambda+\varepsilon} = \frac{1}{2|U'(x_0)|}. \tag{5.25}$$

Since

$$U'(x_0) = \frac{1}{1+\beta}\left[I'(x_0) - \frac{1}{2}\right], \tag{5.26}$$

it follows that $\lambda = \lambda_\pm$, where

$$\lambda_\pm = \frac{-\Lambda \pm \sqrt{\Lambda^2 - 4(1-\Gamma)\varepsilon(1+\beta)}}{2} \tag{5.27}$$

with

$$\Lambda = 1 + \varepsilon - (1+\beta)\Gamma \tag{5.28}$$

and

$$\Gamma = \frac{1}{1+2D}, \quad D = |I'(x_0)|. \tag{5.29}$$

We have used the fact that $I'(x_0) \leq 0$. The eigenvalues $\lambda_\pm$ determine the discrete spectrum.

5.3. Hopf bifurcation to a breathing front. Equation (5.27) implies that the stationary front is locally stable, provided that $\Lambda > 0$ or, equivalently, the gradient of the inhomogeneous input at x_0 satisfies

$$D > D_c \equiv \frac{1}{2}\frac{\beta - \varepsilon}{1+\varepsilon}. \tag{5.30}$$

Since $D \geq 0$, it follows that the front is stable when $\beta < \varepsilon$, that is, when the feedback is sufficiently weak or fast. On the other hand, if $\beta > \varepsilon$, then there is a Hopf bifurcation at the critical gradient $D = D_c$. The corresponding critical Hopf frequency is

$$\omega_H = \sqrt{\frac{2D_c\varepsilon(1+\beta)}{2D_c+1}} = \sqrt{\varepsilon(\beta-\varepsilon)}. \tag{5.31}$$

Note that the frequency depends only on the size and rate of the negative feedback but is independent of the details of the synaptic weight distribution and the size of the input. This should be contrasted with the corresponding Hopf frequency in the case of a smooth nonlinearity f and a weak step-inhomogeneity; see (4.14). The latter depends on the input amplitude and the form of the stationary solution $\overline{U}$, which itself depends on the weight distribution w.

In order to investigate the nature of solutions around the Hopf bifurcation point, we consider the particular example of a smooth ramp inhomogeneity

$$I(x) = -\frac{s}{2}\tanh(\gamma x), \tag{5.32}$$

where s is the size of the step and γ determines its steepness. A stationary front will exist provided that

$$s > \bar{s} \equiv |1 - 2\kappa(1+\beta)|. \tag{5.33}$$

The gradient $D = s\gamma\,\mathrm{sech}^2(\gamma x_0)/2$ depends on x_0, which is itself dependent on β and κ through (5.19). Using the identity $\mathrm{sech}^2 x = 1 - \tanh^2 x$, it follows that

$$D = \frac{\gamma}{2s}\left(s^2 - \bar{s}^2\right). \tag{5.34}$$

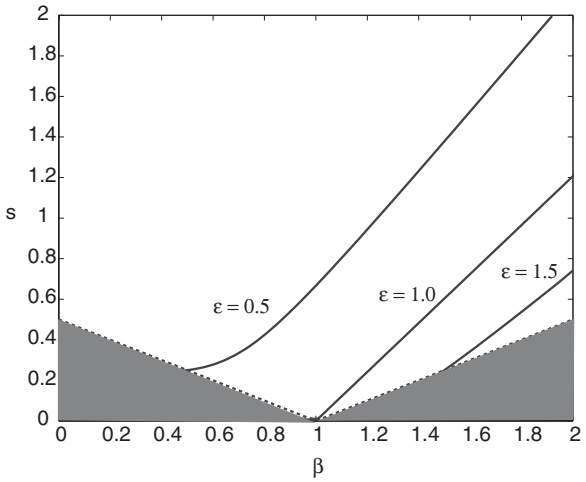


FIG. 5.2. *Stability phase diagram for a stationary front in the case of a step input* $I(x) = -s\tanh(\gamma x)/2$, *where* γ *is the steepness of the step and* s *its height. Hopf bifurcation lines (solid curves) in* $s-\beta$ *parameter space are shown for various values of* ε. *In each case the stationary front is stable above the line and unstable below it. The shaded area denotes the region of parameter space where a stationary front solution does not exist. The threshold* $\kappa = 0.25$ *and* $\gamma = 0.5$.

Combining (5.30) and (5.34), we obtain an expression for the critical value of s that determines the Hopf bifurcation points:

$$s_c = \frac{1}{2\gamma}\left[\frac{\beta-\varepsilon}{1+\varepsilon} + \sqrt{\left(\frac{\beta-\varepsilon}{1+\varepsilon}\right)^2 + 4\bar{s}^2\gamma^2}\right]. \tag{5.35}$$

The critical height s_c is plotted as a function of β for various values of ε and fixed κ, γ in Figure 5.2. Note that in the homogeneous case ($s = 0$) a stationary solution exists only at the particular value of β given by $\beta = 1/(2\kappa) - 1$. This solution is stable for $\varepsilon > \beta$ and unstable for $\varepsilon < \beta$, which is consistent with the pitchfork bifurcation shown in Figure 5.1. Close to the front bifurcation $\varepsilon = \beta$, the Hopf bifurcation occurs in the presence of a weak input inhomogeneity, which is the case considered in section 2. Now, however, it is possible to determine the bifurcation curve without any restrictions on the size of the input.

Numerically solving the full system of equations (1.1) for a step input $I(x)$, exponential weights $w(x)$, and threshold nonlinearity $f(u) = H(u-\kappa)$ shows that the Hopf bifurcation is supercritical, in which there is a transition to a small amplitude breather whose frequency of oscillation is approximately equal to the Hopf frequency ω_H. As the input amplitude s is reduced beyond the Hopf bifurcation point, the amplitude of the oscillation increases until the breather itself becomes unstable and there is a secondary bifurcation to a traveling front. This is illustrated in Figure 5.3, which shows a space-time plot of the developing breather as the input amplitude is slowly reduced. Note that analogous results have been obtained for pulses in the presence of stationary Gaussian inputs, where a reduction in the input amplitude induces a Hopf bifurcation to a pulse-like breather [3, 6]. Interestingly, the localized breather can itself undergo a secondary instability leading to the periodic emission of traveling waves. In one dimension such waves consist of pairs of counterpropagating pulses, whereas in two dimensions the waves are circular target patterns [6].

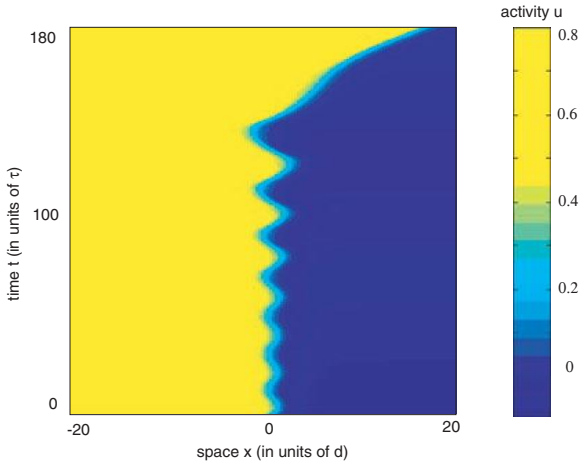


FIG. 5.3. *Breather-like solution arising from a Hopf instability of a stationary front due to a slow reduction in the size s of the step input inhomogeneity* (5.32). *Here* $\varepsilon = 0.5, \gamma = 0.5, \beta = 1, \kappa = 0.25$. *The input amplitude* $s = 2$ *at* $t = 0$ *and* $s = 0$ *at* $t = 180$. *The amplitude of the oscillation steadily grows until it destabilizes at* $s \approx 0.05$, *leading to the generation of a traveling front.*

5.4. Locking to a moving input. We conclude our analysis of the exactly solvable model by considering the effects of a moving input stimulus. This is interesting from a number of viewpoints. First, introducing a persistent stationary input into an in vitro cortical slice can damage the tissue, whereas a moving input (at least if it is localized) will not. Second, in vivo inputs into the intact cortex are typically nonstationary, as exemplified by inputs to the visual cortex induced by moving visual stimuli. We consider the particular problem of whether or not a traveling front can lock to a step-like input $I(x) = I_0\chi(x - vt)$ traveling with constant speed v, where

$$\chi(x) = \begin{cases} -1, & x > 0, \\ 0, & x = 0, \\ +1, & x < 0. \end{cases}$$

Such a front moves at the same speed as the input but may be shifted in space relative to the input.

We proceed by introducing the traveling wave coordinate $\xi = x - vt$ and deriving existence conditions for a front solution $U(\xi)$ satisfying $U(\xi) \to 0$ as $\xi \to \infty$, $U(\xi) \to (1+\beta)^{-1}$ as $\xi \to -\infty$, and $U(\xi_0) = \kappa$. Substituting into (1.1) gives

$$-vU'(\xi) = -U(\xi) + \int_{-\infty}^{\xi_0} w(\xi - \eta)d\eta - \beta V(\xi) + I_0\chi(\xi), \tag{5.36}$$

$$-vV'(\xi) = \epsilon(-V(\xi) + U(\xi)). \tag{5.37}$$

Setting $W(\xi) = \int_{\xi}^{\infty} w(\eta)d\eta$, we can rewrite this pair of equations in the matrix form

$$\mathbf{LS} \equiv \begin{pmatrix} vU' - U - \beta V \\ vV' + \epsilon U - \epsilon V \end{pmatrix} = -\begin{pmatrix} N_{\mathrm{E}} \\ 0 \end{pmatrix}, \tag{5.38}$$

where

$$\mathbf{S} = (U, V)^T, \qquad N_{\mathrm{E}}(\xi) = W(\xi - \xi_0) + I_0\chi(\xi). \tag{5.39}$$

We use variation of parameters to solve this linear equation. The homogeneous problem $\mathbf{L}S = 0$ has the two linearly independent solutions,

$$\mathbf{S}_+(\xi) = \begin{pmatrix} \beta \\ m_+ - 1 \end{pmatrix} \exp(\mu_+ \xi), \tag{5.40}$$

$$\mathbf{S}_-(\xi) = \begin{pmatrix} \beta \\ m_- - 1 \end{pmatrix} \exp(\mu_- \xi), \tag{5.41}$$

where

$$\mu_\pm = \frac{m_\pm}{v}, \qquad m_\pm = \frac{1}{2}\Big(1 + \epsilon \pm \sqrt{(1-\epsilon)^2 - 4\epsilon\beta}\Big).$$

By variation of parameters we define

$$\mathbf{S}(\xi) = [\mathbf{S}_+(\xi)|\mathbf{S}_-(\xi)] \begin{pmatrix} a(\xi) \\ b(\xi) \end{pmatrix},$$

where $[\mathbf{A}|\mathbf{B}]$ denotes the matrix whose first column is defined by the vector $\mathbf{A}$ and whose second column is defined by the vector $\mathbf{B}$. Then

$$\begin{aligned} \mathbf{LS} &= v\frac{\partial}{\partial \xi}\left([\mathbf{S}_+(\xi)|\mathbf{S}_-(\xi)] \begin{pmatrix} a(\xi) \\ b(\xi) \end{pmatrix}\right) - \begin{pmatrix} 1 & \beta \\ -\epsilon & \epsilon \end{pmatrix}\left([\mathbf{S}_+(\xi)|\mathbf{S}_-(\xi)] \begin{pmatrix} a(\xi) \\ b(\xi) \end{pmatrix}\right) \\ &= v[\mathbf{S}_+(\xi)|\mathbf{S}_-(\xi)]\frac{\partial}{\partial \xi} \begin{pmatrix} a(\xi) \\ b(\xi) \end{pmatrix}, \end{aligned} \tag{5.42}$$

since $\mathbf{LS}_\pm = 0$. Thus (5.38) reduces to

$$[\mathbf{S}_+(\xi)|\mathbf{S}_-(\xi)]\frac{\partial}{\partial \xi} \begin{pmatrix} a(\xi) \\ b(\xi) \end{pmatrix} = -\frac{1}{v} \begin{pmatrix} N_{\rm E} \\ 0 \end{pmatrix}. \tag{5.43}$$

The matrix $[\mathbf{S}_+(\xi)|\mathbf{S}_-(\xi)]$ is invertible. Introducing the vector-valued functions

$$\mathbf{Z}_+(\xi) = \begin{pmatrix} 1 - m_- \\ \beta \end{pmatrix} \exp(-\mu_+ \xi), \tag{5.44}$$

$$\mathbf{Z}_-(\xi) = -\begin{pmatrix} 1 - m_+ \\ \beta \end{pmatrix} \exp(-\mu_- \xi), \tag{5.45}$$

we have

$$[\mathbf{S}_+|\mathbf{S}_-][\mathbf{Z}_+|\mathbf{Z}_-]^{\rm T} = [\mathbf{Z}_+|\mathbf{Z}_-]^{\rm T}[\mathbf{S}_+|\mathbf{S}_-] = \beta(m_+ - m_-)\mathrm{I},$$

where I denotes the identity matrix. Multiplying (5.43) by $[\mathbf{Z}_+|\mathbf{Z}_-]^{\rm T}$ finally yields the first-order equation

$$\frac{\partial}{\partial \xi} \begin{pmatrix} a(\xi) \\ b(\xi) \end{pmatrix} = -\frac{1}{v\beta(m_+ - m_-)}[\mathbf{Z}_+(\xi)|\mathbf{Z}_-(\xi)]^{\rm T} \begin{pmatrix} N_{\rm E}(\xi) \\ 0 \end{pmatrix}. \tag{5.46}$$

In order to solve (5.46) we need to specify the sign of v. First, suppose that $v > 0$, which corresponds to a right-moving front. Integrating over the interval $[\xi, \infty)$ gives

$$\begin{pmatrix} a(\xi) \\ b(\xi) \end{pmatrix} = \begin{pmatrix} a_\infty \\ b_\infty \end{pmatrix} + \frac{1}{v\beta(m_+ - m_-)} \int_\xi^\infty [\mathbf{Z}_+(\eta)|\mathbf{Z}_-(\eta)]^{\rm T} \begin{pmatrix} N_{\rm E}(\eta) \\ 0 \end{pmatrix} d\eta,$$

where a_∞, b_∞ denote the values of a, b at ∞. Since we seek a bounded solution $\mathbf{S}(\xi)$, we must require that $a_\infty = b_\infty = 0$. Hence the solution is

$$\begin{pmatrix} a(\xi) \\ b(\xi) \end{pmatrix} = \frac{1}{v\beta(m_+ - m_-)} \int_\xi^\infty [\mathbf{Z}_+(\eta)|\mathbf{Z}_-(\eta)]^{\mathrm{T}} \begin{pmatrix} N_{\mathrm{E}}(\eta) \\ 0 \end{pmatrix} d\eta,$$

so that

$$\mathbf{S}(\xi) = \frac{1}{v\beta(m_+ - m_-)} [\mathbf{S}_+(\xi)|\mathbf{S}_-(\xi)] \int_\xi^\infty [\mathbf{Z}_+(\eta)|\mathbf{Z}_-(\eta)]^{\mathrm{T}} \begin{pmatrix} N_{\mathrm{E}}(\eta) \\ 0 \end{pmatrix} d\eta. \tag{5.47}$$

Further simplification occurs by introducing the functions

$$M_\pm(\xi) = \frac{1}{v} \left(\frac{1}{m_+ - m_-} \right) \int_\xi^\infty e^{\mu_\pm(\xi - \eta)} N_{\mathrm{E}}(\eta) d\eta.$$

We can then express the solution for $(U(\xi), V(\xi))$ as follows:

$$U(\xi) = (1 - m_-)M_+(\xi) - (1 - m_+)M_-(\xi), \tag{5.48}$$

$$V(\xi) = \beta^{-1}(m_+ - 1)(1 - m_-) \left[M_+(\xi) - M_-(\xi) \right]. \tag{5.49}$$

To ensure that such a front exists we require that $U(\xi_0) = \kappa$, i.e.,

$$\kappa = (1 - m_-)M_+(\xi_0) - (1 - m_+)M_-(\xi_0). \tag{5.50}$$

Taking $w(x) = \mathrm{e}^{-|x|}/2$ so that

$$W(\xi) = \begin{cases} 1 - \dfrac{1}{2} e^{\xi}, & \xi < 0, \\ \\ \dfrac{1}{2} e^{-\xi}, & \xi \geq 0, \end{cases}$$

we can calculate $M_\pm(\xi_0)$ explicitly as

$$M_\pm(\xi_0) = \frac{1}{(m_+ - m_-)} \left(\frac{1}{2(v + m_\pm)} - \frac{1}{m_\pm} F(\xi_0) \right),$$

where

$$F(\xi_0) = \begin{cases} I_0(2e^{\mu_\pm \xi_0} - 1), & \xi_0 < 0, \\ \\ I_0, & \xi_0 \geq 0. \end{cases}$$

The case of a left-moving front for which $v < 0$ follows along similar lines by integrating (5.46) over $(-\infty, \xi_0]$:

$$U(\xi) = (m_- - 1)\breve{M}_+(\xi) - (m_+ - 1)\breve{M}_-(\xi), \tag{5.51}$$

$$V(\xi) = \beta^{-1}(m_+ - 1)(1 - m_-) \left[\breve{M}_+(\xi) - \breve{M}_-(\xi) \right], \tag{5.52}$$

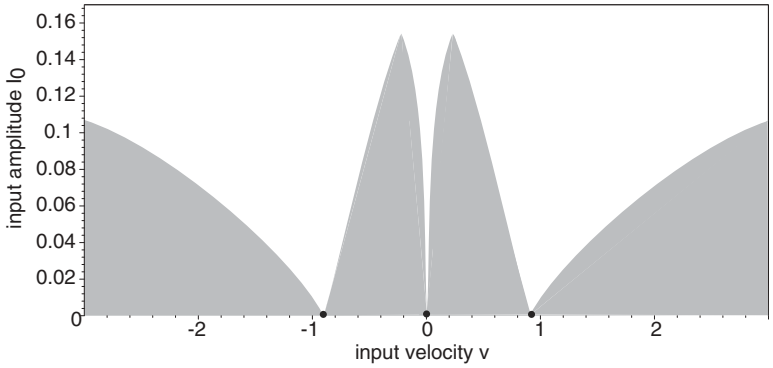


FIG. 5.4. *Locking of a traveling front to a moving step input with velocity v and amplitude I_0. Other parameter values are $\beta = 1, \varepsilon = 0.1, \kappa = 0.25$. Unshaded regions show where locking can occur in the (v, I_0)-plane. When $I_0 = 0$ there are three front solutions corresponding to a stationary front ($v = 0$) and two counterpropagating fronts, which is consistent with the front bifurcation shown in Figure* 5.1. *Each of these solutions forms the vertex of a distinct locking region whose width increases monotonically with I_0 so that ultimately the locking regions merge.*

where

$$\breve{M}_{\pm}(\xi_0) = \frac{1}{(m_+ - m_-)} \left(\frac{1}{2} \frac{m_{\pm} - 2v}{m_{\pm}(v - m_{\pm})} - \frac{1}{m_{\pm}} G(\xi_0) \right)$$

and

$$G(\xi_0) = \begin{cases} -I_0, & \xi_0 < 0, \\ I_0(1 - 2e^{\mu_{\pm}\xi_0}), & \xi_0 \geq 0. \end{cases}$$

This leads to the following threshold condition for $v < 0$:

$$\kappa = (m_- - 1)\breve{M}_+(\xi_0) - (m_+ - 1)\breve{M}_-(\xi_0). \tag{5.53}$$

We can now numerically solve (5.50) and (5.53) to determine the range of input velocities v and input amplitudes I_0 for which locking occurs. For the sake of illustration, we assume the threshold condition $2\kappa(1 + \beta) = 1$ and take $\varepsilon < \beta$. This ensures that, in the absence of any input, there exists an unstable stationary front and a pair of stable counterpropagating waves (see Figure 5.1). The continuation of these stationary and traveling fronts as I_0 increases from zero is shown in Figure 5.4. Since $2\kappa(1+\beta) = 1$, equations (5.50) and (5.53) are equivalent under the interchange $v \to -v$ and $\xi_0 \to -\xi_0$. This implies that the locking regions are symmetric with respect to v. For nonzero v the traveling front is shifted relative to the input such that $\xi_0 < 0$ when $v > 0$ and $\xi_0 > 0$ when $v < 0$. In other words, the wave is dragged by the input.

Figure 5.4 determines where locking can occur but not whether the resulting traveling wave is stable or unstable. Indeed, the stability analysis of traveling fronts is considerably more involved than that of stationary fronts. Nevertheless, we expect that for sufficiently small I_0 the locking regions around the counterpropagating fronts are stable, whereas the central region containing the stationary front is unstable. On the other hand, since $\beta > \varepsilon$, we know that the stationary front is stable for large inputs I_0 and undergoes a Hopf bifurcation as I_0 is reduced. This suggests that the Hopf bifurcation point at $v = 0$ lies on a Hopf curve within the locking region so that

a traveling front locked to a moving input can also be destabilized as the strength of the input is reduced (or as the input velocity changes relative to the intrinsic velocity of waves in the homogeneous network). Recently, Zhang [19] analyzed the asymptotic stability of traveling wave solutions of (1.1) in the case of homogeneous inputs by deriving the associated Evans function and evaluating it in the singular limit $\varepsilon \ll 1$. In future work we will extend this analysis to the case of inhomogeneous inputs and finite ε, thus determining the stability of the locking regions shown in Figure 5.4. We will also construct corresponding locking regions for traveling pulses in the presence of moving Gaussian inputs, and numerically explore the types of oscillatory solutions bifurcating from these waves.

Acknowledgment. We would like to thank Yue-Xian Li (University of British Columbia) for many helpful discussions regarding his work on wavefront instabilities in reaction–diffusion equations.

REFERENCES

[1] S. Amari, *Dynamics of pattern formation in lateral inhibition type neural fields*, Biol. Cybern., 27 (1977), pp. 77–87.

[2] M. Bode, *Front-bifurcations in reaction-diffusion systems with inhomogeneous parameter distributions*, Phys. D, 106 (1997), pp. 270–286.

[3] P. C. Bressloff, S. E. Folias, A. Prat, and Y.-X. Li, *Oscillatory waves in inhomogeneous neural media*, Phys. Rev. Lett., 91 (2003), article 178101.

[4] R. D. Chervin, P. A. Pierce, and B. W. Connors, *Periodicity and directionality in the propagation of epileptiform discharges across neocortex*, J. Neurophysiol., 60 (1988), pp. 1695–1713.

[5] G. B. Ermentrout and J. B. McLeod, *Existence and uniqueness of travelling waves for a neural network*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 461–478.

[6] S. E. Folias and P. C. Bressloff, *Breathing pulses in an excitatory neural network*, SIAM J. Appl. Dynam. Systems, to appear.

[7] D. Golomb and Y. Amitai, *Propagating neuronal discharges in neocortical slices: Computational and experimental study*, J. Neurophysiol., 78 (1997), pp. 1199–1211.

[8] A. Hagberg and E. Meron, *Pattern formation in non-gradient reaction-diffusion systems: The effects of front bifurcations*, Nonlinearity, 7 (1994), pp. 805–835.

[9] A. Hagberg, E. Meron, I. Rubinstein, and B. Zaltzman, *Controlling domain patterns far from equilibrium*, Phys. Rev. Lett., 76 (1996), pp. 427–430.

[10] M. A. P. Idiart and L. F. Abbott, *Propagation of excitation in neural network models*, Network, 4 (1993), pp. 285–294.

[11] Y.-X. Li, *Tango waves in a bidomain model of fertilization calcium waves*, Phys. D, 186 (2003), pp. 27–49.

[12] D. J. Pinto and G. B. Ermentrout, *Spatially structured activity in synaptically coupled neuronal networks*: I. *Traveling fronts and pulses*, SIAM J. Appl. Math., 62 (2001), pp. 206–225.

[13] A. Prat and Y.-X. Li, *Stability of front solutions in inhomogeneous media*, Phys. D, 186 (2003), pp. 50–68.

[14] J. Rinzel and D. Terman, *Propagation phenomena in a bistable reaction-diffusion system*, SIAM J. Appl. Math., 42 (1982), pp. 1111–1137.

[15] J. E. Rubin, *Stability, bifurcations and edge oscillations in standing pulse solutions to an inhomogeneous reaction-diffusion system*, Proc. Roy. Soc. Edinburgh Sect. A, 129 (1999), pp. 1033–1079.

[16] P. Schutz, M. Bode, and H.-G. Purwins, *Bifurcations of front dynamics in a reaction-diffusion system with spatial inhomogeneities*, Phys. D, 82 (1995), pp. 382–397.

[17] H. R. Wilson and J. D. Cowan, *A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue*, Kybernetik, 13 (1973), pp. 55–80.

[18] J.-Y. Wu, L. Guan, and Y. Tsau, *Propagating activation during oscillations and evoked responses in neocortical slices*, J. Neurosci., 19 (1999), pp. 5005–5015.

[19] L. Zhang, *On stability of traveling wave solutions in synaptically coupled neuronal networks*, Differential Integral Equations, 16 (2003), pp. 513–536.

SIAM J. APPL. MATH.
Vol. 65, No. 1, pp. 152–174

SWARMING PATTERNS IN A TWO-DIMENSIONAL KINEMATIC MODEL FOR BIOLOGICAL GROUPS*

CHAD M. TOPAZ† AND ANDREA L. BERTOZZI†

Abstract. We construct a continuum model for the motion of biological organisms experiencing social interactions and study its pattern-forming behavior. The model takes the form of a conservation law in two spatial dimensions. The social interactions are modeled in the velocity term, which is nonlocal in the population density and includes a parameter that controls the interaction length scale. The dynamics of the resulting partial integrodifferential equation may be uniquely decomposed into incompressible motion and potential motion. For the purely incompressible case, the model resembles one for fluid dynamical vortex patches. There exist solutions which have constant population density and compact support for all time. Numerical simulations produce rotating structures which have circular cores and spiral arms and are reminiscent of naturally observed phenomena such as ant mills. The sign of the social interaction term determines the direction of the rotation, and the interaction length scale affects the degree of spiral formation. For the purely potential case, the model resembles a nonlocal (forwards or backwards) porous media equation. The sign of the social interaction term controls whether the population aggregates or disperses, and the interaction length scale controls the balance between transport and smoothing of the density profile. For the aggregative case, the population clumps into regions of high and low density. The characteristic length scale of the density pattern is predicted and confirmed by numerical simulations.

Key words. aggregation, integrodifferential equation, pattern, swarming, vortex

AMS subject classifications. 92, 35, 76

DOI. 10.1137/S0036139903437424

1. Introduction.

Examples of collective motion abound in nature. Swarming, schooling, flocking, and herding have been observed among zooplankton, locusts, fish, birds, wolves, and other organisms; see [2, 21, 24] for discussions of some of these groups. These biological aggregations all consist of individuals moving in a coordinated manner, and yet they still represent a rich spectrum of phenomena [23]. For instance, length scales of different swarms may vary widely, ranging up to 100 km^2 in cross-sectional area for African migratory locust swarms [11]. Groups may also have different dimensionalities, from the three-dimensional rolling structure formed by the locusts [11] to the flat, two-dimensional structure of vortex-like ant mills [26]. Aggregations also vary in their degree of organization. Some groups, such as fish schools, are made up of individuals with highly correlated velocities, while other groups, such as mosquito swarms, accomplish their movement in a more disorganized manner [21]. Nonetheless, a common, remarkable aspect of these aggregations is that coordinated movement is achieved even though interactions between individuals via sight, smell, hearing, or other senses are typically limited to much shorter distances than the size of the group. Swarming, then, constitutes an example of how a global structure can arise from more localized rules.

While studies of emergent global structure and pattern formation have a long history in the realm of physics [7], the corresponding endeavors in biology are more

*Received by the editors November 7, 2003; accepted for publication (in revised form) March 24, 2004; published electronically September 24, 2004.
http://www.siam.org/journals/siap/65-1/43742.html
†Department of Mathematics, UCLA, Los Angeles, CA 90095 (topaz@ucla.edu, bertozzi@math.ucla.edu). The research of the first author was supported by NSF VIGRE grants DMS-9983320 and DMS-9983726. The research of the second author was supported by ARO grant DAAD-19-02-1-0055, ONR grants N000140310073 and N000140410054, and NSF grant DMS-0244498.

nascent. It is only over the past few decades that scientists have begun to address swarming from a mathematical perspective. The mathematical investigations are of interest not only because swarming is a widespread phenomenon, and not only because it provides a rich example of biological pattern formation, but because ideas from swarming have applications to other fields. For instance, swarming has recently been used by engineers, computer scientists, and others as a paradigm for developing efficient algorithms to accomplish routing, cooperative transport, and other tasks [5].

Previous theoretical investigations of swarms have been carried out in a wide variety of mathematical settings. We briefly mention some of these here; an overview of modeling issues is also available in sources such as [10, 19, 21]. One fruitful approach to modeling swarms has been to treat each individual as a discrete particle. These "individual-based" models have been employed in quite a few biological and mathematical studies, including [1, 8, 12, 13, 15, 18, 27, 29, 30]. These works begin with simple rules of motion for each individual, involving some combination of self-propulsion, random movement, and interaction with neighboring organisms. The models typically take the form of coupled nonlinear difference or differential equations, which may be stochastic or deterministic, depending on the particular ingredients of each model. Numerical simulations have indeed revealed collective behavior. However, a principal disadvantage of such models is that, for realistic numbers of individuals, analytical results for the collective motion are difficult or impossible to obtain. It is worth mentioning that some progress has been made in obtaining analytical results for *stationary* groups. In [18], a discrete model is formulated, and a Lyapunov functional is used to successfully predict an equilibrium state of equally spaced organisms. However, to our knowledge, analytical (nonstatistical) descriptions of nonequilibrium states in discrete swarming models are scarce.

Other investigations of swarming have been carried out in a continuum setting, in which relevant quantities are described as scalar or vector fields. This approach was perhaps first popularized in [20]; reviews are provided in [14, 19]. Continuum models, such as those in [11, 15, 17, 29], may be constructed a priori or by coarse-graining a particle model. In general, continuum models provide a convenient setting in which to study large populations, since one may apply machinery from the analysis of partial differential equations. In the context of swarming, the focus has generally been on models in which the population density satisfies a convection-diffusion equation of the form

$$\rho_t + \nabla \cdot (\vec{v}\rho) = \nabla \cdot (D \nabla \rho). \tag{1.1}$$

Here, $\rho(\vec{x}, t)$ is the population density, $\vec{v}(\vec{x}, t)$ is the velocity field, and $\vec{x}$ is the (one-, two-, or three-dimensional) spatial coordinate. This equation states that the density is conserved while individuals travel with average velocity $\vec{v}$. The motion may involve diffusion, whose strength is measured by $D \equiv D(\vec{x}, \vec{v}, \rho)$.

Swarming models may also be classified as either dynamic or kinematic, depending on how velocities are determined. In dynamic models, velocities are determined by Newton's second law. For continuum models, a dynamic rule would couple (1.1) to a momentum equation for the velocity field, such as

$$\vec{v}_t + \vec{v} \cdot \nabla \vec{v} = \vec{f}(\vec{v}) - \vec{k}(\rho, \vec{v}) + \nu \nabla^2 \vec{v} + \vec{F}_{ext}. \tag{1.2}$$

Here, the left-hand side is the material (or convective) derivative; i.e., the time derivative in a reference frame moving with the velocity field $\vec{v}$. The right-hand side is simply a sum of forces. The force $\vec{f}$ represents the self-propulsion of individuals, and $\vec{k}$ is a

nonlocal force due to interactions with other members of the population, to be discussed momentarily. The remaining terms on the right-hand side of (1.2) represent a "viscosity," with strength proportional to ν, and an external (environmental) force $\vec{F}_{ext}$. An example of a dynamic model for swarming may be found in [15]. In that work, $\nu = 0, \vec{F}_{ext} = \vec{0}$, and $\vec{f}(\vec{v}) = \alpha\vec{v}/|\vec{v}| - \beta\vec{v}$, so that in the absence of social interactions, individuals experience a self-propulsion of strength α in their direction of motion and a frictional drag of strength β.

In contrast, in kinematic models such as that in [17], inertial effects are not important. The motion of bodies is determined without consideration of the forces acting upon them. For continuum models, then, the velocity does not satisfy a momentum equation, but rather is simply a functional of the population density, i.e.,

$$\vec{v} = \vec{V}(\rho). \tag{1.3}$$

The functional $\vec{V}$ may include effects like those captured in (1.2), such as self-propulsion, social interactions, and environmental influence.

The essence of animal aggregations is the presence of social interactions. (Here we distinguish from animal "congregations," which may arise when organisms gather at a common attractant, such as a food or light source.) For the velocity equation (1.2), these interactions are represented by the term $\vec{k}(\rho, \vec{v})$, and for (1.3), they are contained in the functional $\vec{V}(\rho)$. The social interaction terms might describe effects such as attraction or repulsion between individuals sufficiently close to each other (a brief review of attraction/repulsion between organisms is provided in [21]) or the tendency of individuals to orient themselves similarly to their neighbors. Within the context of continuum models, then, the social term takes the form of an integral operator (most often of convolution form), and the governing equations are actually partial integrodifferential equations, as in [15, 17].

One challenge associated with continuum models has been the difficulty of obtaining swarm solutions with biologically realistic characteristics such as sharp boundaries, relatively constant internal population densities, and long lifetimes [23]. For swarms in one spatial dimension, some progress was made in [17], which also contains extensive background and an associated literature review on this issue. We believe that a related issue is the dimensionality of the model. Most continuum swarm models, such as those in [11, 17, 31], have been investigated in only one spatial dimension. We expect swarming dynamics in higher-dimensional models to be qualitatively different, since those cases allow for organisms to vary their orientations continuously, as in the "mill" or "vortex" states that have been observed in fish schools, ant colonies, and other groups [3, 22, 23, 25, 26].

In this paper, we consider a two-dimensional continuum kinematic model for the motion of biological organisms experiencing nonlocal social interactions, characterize the dynamics which may occur, and explore the pattern-forming behavior. Our model is general and abstract, since one of our goals in this paper is to highlight a difference between the swarming problems for one and two dimensions. The possibility of rotational motion in two spatial dimensions allows for cohesively moving swarms with sharp boundaries and infinite lifetimes, even in the absence of a local drift velocity, which is a key ingredient for one-dimensional models. Another goal is to demonstrate a natural way of classifying swarming dynamics in two spatial dimensions, namely by using the Hodge decomposition theorem. Our final goal is to determine explicitly how properties of the social interaction terms (for instance, their associated length scales and signs) affect the large-scale dynamics of the population. Throughout, we

highlight connections between our continuum swarming model and fluid-dynamical phenomena such as vortex patches and flow through porous media.

This paper is organized as follows. In section 2 we mention some results for constant density traveling band solutions of a class of one-dimensional swarm models. These solutions, which are the appropriate mathematical descriptions of one-dimensional swarms, rely on the presence of local drift in the model. These results are presented for contrast with the two-dimensional case, the study of which constitutes the bulk of this paper.

In section 3 we formulate an abstract kinematic model for the motion of biological organisms experiencing nonlocal social interactions. We also discuss how the Hodge decomposition theorem provides a useful way of understanding the two-dimensional motion of the group, namely by decomposing it into incompressible motion and potential motion.

In section 4 we focus on the case of incompressible motion. The resulting problem bears many similarities to fluid-dynamical vortex patches. Since we wish to study the motion of a biologically realistic swarm, we assume the initial condition to be a finite group with constant internal population density and sharp edges (compact support). We show that such a swarm retains these characteristics for all time. Numerical simulations demonstrate that the dynamics of the swarm are rotational, and that the asymptotic states are vortex-like structures with circular cores and a potentially complex arrangement of spiral arms. The sign of the social interaction term determines the direction of rotation of the swarm, and the characteristic length scale of the interactions determines the degree of spiral formation. The spiral states we observe are qualitatively similar to the mill states observed in [3, 22, 23, 25, 26].

In section 5 we consider the complementary case of potential motion. In this case, the resulting problem resembles a (forwards or backwards) porous media equation. The sign of the social interaction term determines whether the interaction represents nonlocal repulsion or attraction. These effects lead, respectively, to dispersion or aggregation of the population. For the dispersive case, shorter interaction length scales result in smoothing of the population density profile, while larger interaction length scales lead to motion which is more convective. For the case of aggregation, a simple linear stability analysis enables us to identify a maximally unstable wavelength and thus to make a prediction about the characteristic length scale of the clumped population distribution that will form. We demonstrate these results by means of numerical simulations.

Finally, we conclude in section 6 with a brief summary and a discussion of directions for future investigation.

2. Elementary results for one-dimensional kinematic swarms. A detailed investigation of a swarming model in one spatial dimension has been carried out in [17]. In that work, the population density ρ is assumed to obey (1.1). The kinematic velocity rule is

$$v = a_e\rho + (A_a - A_r\rho)(K * \rho), \quad a_e, A_a, A_r \in \mathbb{R}. \tag{2.1}$$

Here, the first term represents a local density-dependent drift. The remaining terms are nonlocal components describing attraction and repulsion, with the asterisk operator having the meaning of convolution. Note that the repulsive effects are higher order in the population density than are the attractive ones. The interaction kernel

K is odd, piecewise constant, and has compact support. It is given by

$$K(x) = \begin{cases} \frac{1}{2d} & \text{if } |x| \le r, \\ 0 & \text{otherwise,} \end{cases} \tag{2.2}$$

where d is an interaction length scale parameter that may be freely chosen.

Analysis and numerical simulations in [17] reveal that the model supports swarm solutions with biologically realistic characteristics, namely, a nearly constant internal population density and sharp edges. For density-independent diffusion, the cohesive swarm has an exponentially long lifetime before the population is lost through "tails" in the density profile. For the case of small density-dependent diffusion, the model has true "traveling band" solutions which have compact support. In either case, the cohesive motion of the swarm is achieved by an effective cancellation of the social interactions. The internal density of the swarm is precisely that at which the attractive and repulsive effects cancel each other, so that the only remaining component of the velocity is local drift. We mention this for contrast with the results to be presented later in this paper, which demonstrate a nonlocal, i.e., cooperative, means by which a constant density swarm may move cohesively.

With purely nonlocally determined velocities, traveling band solutions are difficult to obtain even in the absence of diffusion. We now mention some simple existence and uniqueness results for one-dimensional swarms with no diffusion. The population density ρ satisfies the convection equation

$$\rho_t + \partial_x(v\rho) = 0. \tag{2.3}$$

We will show how, in one dimension, realistic velocity rules that are purely nonlocal cannot lead to a constant-speed translation of the population, and thus cohesive swarms cannot be maintained. Again, these results are presented for contrast with the two-dimensional results given later in this paper.

Since we are interested in making statements about constant density swarms with sharp boundaries, we make the constant-density traveling band (CDTB) ansatz

$$\rho(x,t) = \rho_0 W_L(x - ct), \tag{2.4}$$

$$W_L(x - ct)v(x,t) = cW_L(x - ct). \tag{2.5}$$

Here, ρ_0 is the constant population density, c is the speed of the traveling band, and $W_L(x)$ is the window function defined without loss of generality as

$$W_L(x) = \begin{cases} 1 & \text{if } x \in \Omega_L,\ \Omega_L = [0, L], \\ 0 & \text{otherwise.} \end{cases} \tag{2.6}$$

The ansatz (2.4)–(2.5) automatically satisfies the governing equation (2.3) in Ω_L, the support of ρ. Note that we have not placed any restrictions on the velocity field outside the support of ρ since the velocity in unpopulated areas is irrelevant to the propagation of the swarm.

We must also consider an equation defining the velocity field. For contrast, we consider two velocity rules, each of which may be written as a (degenerate) version of the generalized kinematic velocity rule

$$v(x,t) = F(\rho) + G_1(\rho)\left[K_1 * H_1(\rho)\right] + G_2(\rho)\left[K_2 * H_2(\rho)\right]. \tag{2.7}$$

This rule is a generalization of (2.1) from [17]. F is a functional which captures the local density-dependence of the velocity. It represents the drift velocity of organisms, irrespective of social forces. K_1 and K_2 are interaction kernels, and thus the $G[K*H]$ terms represent nonlocal effects which arise from repulsion and attraction between organisms. G represents the strength of the interaction, and H represents the dependence of the convolution on the population density. We will further assume that K_1 and K_2 are integrable. In contrast to [17], we will also assume that these interaction kernels are decreasing in their arguments, so that the influence of the population on a given organism's velocity weakens with distance. We allow the functionals F, G, H to be nonlinear for generality. We assume that F, G_1, G_2, H_1, H_2 are smooth, and that $H_1(0) = H_2(0) = 0$, so that velocities from social interactions are induced only by nonzero population.

The swarm density ρ_0 and the constant band speed c must satisfy a consistency condition via the velocity equation (2.7) in order for the ansatz (2.4)–(2.5) to be a solution to (2.3).

We first consider the case $F = 0$, $G_2 = 0$, so that (2.7) becomes

$$v = G_1(\rho)\left[K_1 * H_1(\rho)\right], \tag{2.8}$$

where the argument of ρ is understood to be $z = x - ct$. Equivalently, we may obtain a rule of this form by choosing $F = 0,\ G_1 = G_2,\ H_1 = H_2$, so that

$$v = G_1(\rho)\left[(K_1 + K_2) * H_1(\rho)\right], \tag{2.9}$$

and both attractive and repulsive effects are represented. Note that for these velocity rules, in the absence of interactions, there is no underlying (i.e., local) drift velocity.

Combining (2.4)–(2.5) and (2.8), we obtain the consistency condition

$$W_L c = W_L G_1(\rho_0 W_L)\left[K_1 * H(\rho_0 W_L)\right], \tag{2.10}$$

which may be rewritten as

$$c = G_1^0 H_1^0 \int_{z-L}^{z} K_1(\zeta)\, d\zeta \quad \text{for } z \in \Omega_L, \tag{2.11}$$

where

$$G_1^0 \equiv G_1(\rho_0) \tag{2.12}$$

and

$$H_1^0 \equiv H_1(\rho_0). \tag{2.13}$$

By differentiating (2.11) with respect to z and applying the first fundamental theorem of calculus, we see that

$$K_1(z) = K_1(z - L) \quad \text{for } z \in \Omega_L. \tag{2.14}$$

Thus, K_1 satisfying (2.14) must be L-periodic on $[-L, L]$ in order to admit a CDTB solution. (The structure of K_1 outside of $[-L, L]$ is not relevant.) We will call the set of such kernels Υ. The logical implication goes in the reverse direction as well, as can be seen from writing down a Fourier series for $K_1 \in \Upsilon$, so that (2.11) is satisfied if and only if $K_1 \in \Upsilon$. In this case, (2.11) becomes

$$c = G_1^0 H_1^0 K_1^a, \tag{2.15}$$

where

$$K_1^a \equiv \int_0^L K(\zeta)\, d\zeta. \tag{2.16}$$

There may be families of traveling band solutions parameterized by c, which bifurcate depending on the structure of the nonlinear functions G_1 and H_1.

It is important to realize that the choice $K_1 \in \Upsilon$ is not biologically meaningful and contradicts our earlier assumption regarding the spatial decay of interaction kernels. As discussed above, biologically realistic kernels are expected to satisfy $dK_1/d|z| \leq 0$, so that, for a given individual, the effect of other individuals does not increase with distance. However, $K_1 \in \Upsilon$ cannot satisfy $dK_1/d|z| < 0$, so at best the kernel would be a constant, but even this choice is not expected to be a good biological model, except perhaps for very small L.

In contrast to the results just mentioned, we may now consider the case $G_2 = 0$, so that (2.7) becomes

$$v = F(\rho) + G_1(\rho)\left[K_1 * H_1(\rho)\right]. \tag{2.17}$$

Equivalently, we may choose $K_1 = K_2$, $H_1 = H_2$. (The velocity rule (2.1) in [17] takes this form.) Note that now there is a local drift, a self-induced contribution to the velocity, which is captured by F. Combining (2.4)–(2.5) and (2.17), we obtain the consistency condition

$$W_L c = W_L\left\{F(\rho_0 W_L) + G_1(\rho_0 W_L)\left[K_1 * H(\rho_0 W_L)\right]\right\}, \tag{2.18}$$

which may be rewritten as

$$c = F_1^0 + G_1^0 H_1^0 \int_{z-L}^{z} K_1(\zeta)\, d\zeta \quad \text{for } z \in \Omega_L. \tag{2.19}$$

Here,

$$F_1^0 \equiv F_1(\rho_0), \tag{2.20}$$

and H_1^0 and G_1^0 are given by (2.13) and (2.12).

There are two cases to consider. If $H_1^0 G_1^0 \neq 0$, then (2.19) becomes

$$\frac{c - F_1^0}{G_1^0 H_1^0} = \int_{z-L}^{z} K_1(\zeta)\, d\zeta \quad \text{for } z \in \Omega_L. \tag{2.21}$$

This is similar to the previous case. The condition (2.21) may be met only for $K_1 \in \Upsilon$, in which case the existence and uniqueness of solutions depend on the structure of $(c - F_1^0)/(G_1^0 H_1^0)$. For $K_1 \notin \Upsilon$, CDTB solutions are not possible.

On the other hand, if $H_1^0 G_1^0 = 0$, then CDTBs are possible for *any* choice of kernel K_1. In this case, the number of possible CDTB solutions depends on the number of roots of $H_1^0 G_1^0$ for positive ρ, of which there are expected to be a finite number. Looking at the problem from the forward (rather than inverse) perspective, for biologically realistic choices of K_1, the velocity rule (2.17) will lead to a finite number of CDTB solutions. The densities correspond to the solutions $\rho_0^* > 0$ of $H_1^0 G_1^0 = 0$, and the wave speeds are given by $F_1(\rho_0^*)$. Thus, the combination of local and nonlocal velocity terms selects particular densities and band speeds, rather than allowing entire families of solutions, as in the purely nonlocal case. The allowed CDTB densities are those at which the nonlocal interactions disappear. Further, since we imagine the total population to be fixed in number, this velocity rule actually dictates preferred swarm sizes L. These conclusions are similar to those reached in [17] for the particular choice of F_1, G_1, H_1, K_1 given by (2.1).

3. A kinematic two-dimensional swarm model. For the remainder of this paper, we study the dynamics of a two-dimensional swarming model. In constructing our model, we make the following assumptions:

1. The population is conserved; birth, death, immigration, and emigration of organisms are negligible on the time scale of the swarming dynamics.
2. The motion of organisms is due solely to social interactions, and thus velocities depend nonlocally on the population density.
3. Interactions between organisms are pairwise.
4. The social interactions are a linear functional of the population density.
5. The social interactions depend only on the distance between organisms, and become weaker with increasing distance.

Implicit in the second assumption is the supposition that random movement (e.g., due to fluctuations in the organisms' medium, or noise in their ability to move) is negligible. The third, fourth, and fifth assumptions are made for tractability of the model. The third assumption is made so that interaction effects on a given organism will be summable, and this will lead to a convolution in our model, similar to that of the model in [17].

In the spirit of the work in [17], we construct an abstract model, and thus we do not incorporate many biological specifics. Our model might be interpreted as a one for "flat" (two-dimensional) groups in the absence of disturbances such as predators or food sources. Even with the simple assumptions we have made, the dynamics are complex. As discussed in section 6, relaxing some of our assumptions to obtain a more biologically realistic model will be interesting for future work.

Under the assumptions described above, the model takes the form

$$\rho_t + \nabla \cdot (\vec{v}\rho) = 0, \tag{3.1}$$

$$\vec{v} = \int_{\mathbb{R}^2} \vec{K}(|\vec{x} - \vec{y}|)\rho(\vec{y})\, d\vec{y} \equiv \vec{K} * \rho. \tag{3.2}$$

Here, $\vec{x} = (x, y)$ is the two-dimensional spatial coordinate. Note that (3.1) is simply (1.1) with $D = 0$, and (3.2) is a two-dimensional analogue of a degenerate case of the velocity rule (2.7). $\vec{K}$ is our two-dimensional social interaction kernel, which is spatially decaying and isotropic.

Since our model includes no drift term, velocities decay in the far field, and we may apply the Hodge decomposition theorem (see, for instance, [16]). This theorem states that a vector field in the plane may be uniquely decomposed into a divergence-free component and a gradient component. That is to say, the velocity may be written as

$$\vec{v} = \psi + \nabla\Phi, \qquad \nabla \cdot \psi = 0. \tag{3.3}$$

For smooth vector fields decaying at infinity, the divergence-free part has a scalar stream function Ψ satisfying $\nabla^\perp\Psi = 0$. Thus, we can write

$$\vec{v} = \nabla^\perp\Psi + \nabla\Phi. \tag{3.4}$$

Using an analogy to fluid flow, we may think of Ψ as a stream function for the incompressible part of the flow and Φ as a pressure due to interactions. For functions with integrable gradients, convolution commutes with derivatives; i.e., $(\nabla P) * \rho = \nabla(P * \rho)$. Thus, for the model (3.1)–(3.2), we can directly apply the Hodge decomposition to the interaction kernel $\vec{K}$:

$$\vec{K} = \nabla^\perp N + \nabla P, \tag{3.5}$$

where P models the interaction pressure (motion towards and away from concentrations of density) and N models additional motion, which, as we will see, allows for rotation and a cohesive swarm.

To better understand the model (3.1)–(3.2), we separate the dynamics into the two cases which we have just discussed, namely incompressible motion and potential motion. In the following two sections we study each case in turn, and demonstrate how the macroscopic motion of the population is affected by the interaction kernel $\vec{K}$.

4. Incompressible motion. In this case,

$$\vec{K} = \nabla^{\perp} N \tag{4.1}$$

so that $\nabla \cdot \vec{v} = 0$. Incompressible motion in two spatial dimensions is the analogue of translational drift in one dimension, in that they both allow for cohesive movement of a swarm. However, while drift terms such as $F(\rho)$ in (2.7) cannot lead to pattern formation, incompressible velocities do lead to pattern formation (in particular, vortex patterns, as we demonstrate below) for the two-dimensional case.

The governing equations (3.1)–(3.2) may be written compactly as

$$\rho_t + \nabla \cdot [\rho(\nabla^{\perp} N * \rho)] = 0. \tag{4.2}$$

We take the scalar interaction function N to be a Gaussian of width d; i.e.,

$$G_d(|\vec{x}|) \equiv \frac{1}{d^2} e^{\frac{-|\vec{x}|^2}{d^2}}. \tag{4.3}$$

One might include an additional constant prefactor, but this would represent a velocity scale and may be removed by rescaling the time variable in the equations. The length scale d could also be removed by rescaling, in which case the only parameter remaining in the problem would be the initial condition. We choose to retain the length scale parameter d since it has a clear biological interpretation and since it is more convenient to vary than the length scale of the initial condition.

A Gaussian interaction was also considered for a linear stability analysis in [17]. Other works have used power functions or decaying exponentials [15, 18]. Our interaction function has a somewhat different meaning than the ones used in those previous works since it will be applied in two dimensions and thus has a rotationally symmetric structure. We choose Gaussian interaction functions since they are biologically realistic (in terms of being spatially decaying) and because they have convenient mathematical properties such as bounded norms and infinite differentiability. Many of our qualitative results will hold true for other classes of smooth spatially decaying interaction functions with normalized integral.

We begin by making some general statements about the effect of varying the interaction length scale d. For very small values of d, the interaction function N resembles a δ-function of strength π. For the limiting case $d \to 0$, (4.2) may be written as

$$\rho_t + \pi \nabla \cdot [\rho \nabla^{\perp} \rho] = 0 \tag{4.4}$$

since $\nabla^{\perp}$ commutes with convolution and the δ-function acts as the identity under convolution. A little algebra reveals that (4.4) is actually $\rho_t = 0$, and thus the swarm will be stationary. This makes intuitive sense. In the case that motion is perpendicular to population gradients in a completely local sense, the population

density profile cannot change, by construction. Of course, in a Lagrangian frame (tracking the coordinates of individual organisms) motion is possible, as long as it is perpendicular to the gradient.

On the other hand, for very large values of d, N is nearly a constant, namely zero. In the formal limit $d \to \infty$, $\nabla^{\perp} N = 0$, and once again (4.2) becomes $\rho_t = 0$. This result also makes intuitive sense. In this case, organisms can sense population gradients infinitely far away, but these gradients have no influence on velocity since social interactions are infinitely weak, and thus the organisms are stationary.

For simplicity, and for analogy with the results mentioned in section 2 and in [17], we now focus on constant density solutions of compact support. That is to say, we assume that the initial condition is a *swarm patch* with finite area and constant population density ρ_0. By making such a choice, we are not modeling the initial formation of a constant-density swarm. Rather, this model should be interpreted as a macroscopic description of a swarm in which attractive and repulsive forces have already come into balance. We will study the subsequent movement of such a swarm.

We use Green's formula to rewrite (3.1)–(3.2) as an integral over the boundary:

$$\vec{v}(\vec{x}) = \rho_0 \int_{\partial\Omega} N(|\vec{x} - \vec{y}|)\vec{t}(\vec{y}) ds(\vec{y}), \tag{4.5}$$

where Ω is the support of ρ and the boundary $\partial\Omega$ is parameterized in a clockwise orientation. Here s is the arc length, and $\vec{t}$ is the unit tangent vector to the boundary. Following the strategy from fluid dynamics, we adopt a Lagrangian framework and track points on the boundary of the swarm patch. That is to say, we write down a Lagrangian formulation of (4.5) which will be useful for numerical simulations. Taking α to parameterize the boundary of the swarm, we have the equation for $\vec{z}(\alpha, t)$, the patch boundary:

$$\frac{d\vec{z}(\alpha, t)}{dt} = \int_0^{2\pi} N(\vec{z}(\alpha, t) - \vec{z}(\alpha', t))\vec{z}_{\alpha}(\alpha', t) d\alpha', \tag{4.6}$$

where the subscript α indicates a derivative along the boundary.

This equation describes a self-deforming curve (which encloses a constant area). From a computational standpoint, this formulation is convenient because the dimension of the problem has been reduced by one. More importantly, we see that since the boundary is a self-deforming curve, *the swarm patch retains constant internal density and compact support for all time.* Equation (4.6) is similar to the contour-dynamics formulation of the two-dimensional Euler equations [33], which describe how a fluid region of constant vorticity, or a *vortex patch*, evolves in time. The difference is that for the swarm patch case, the interaction function N is expected to be spatially decaying in order to be biologically meaningful (cf. our modeling assumptions at the start of section 3, and our choice in (4.3)), while for the vortex patch problem, $N = \frac{1}{2\pi} \log |\vec{x}|$.

Before presenting numerical results for our model, we remark on the smoothness of the patch boundary, which historically, for the fluid vortex patches, has been a subject of interest. It is now known that, for the fluid vortex patches, solutions of (4.6) with smooth initial data stay smooth for all time [4, 6]. This turns out to be the case for our present swarm patch problem as well. See the appendix for a sketch of the proof.

Equation (4.6) may be solved numerically to find the evolution of the swarm patch boundary. We now briefly describe our simple numerical algorithm. An initial swarm patch shape is selected, and the boundary is discretized into n nodes. Depending on

our initial condition, we take the initial number of Lagrangian nodes to be between $n = 40$ and $n = 60$. The shape of the patch is evolved by using the discretized version of (4.6). The position of each node may be updated by computing its velocity and then using a time-stepping rule. We perform the spatial integral in (4.6) using Simpson's rule, which is an $\mathcal{O}(n^2)$ operation. As a start-up procedure, we take three time steps using a fourth-order Runge–Kutta method. However, since the Runge–Kutta method involves many evaluations of the right-hand side of (4.6), it is computationally expensive. Thus, we use a fourth-order multistep Adams–Bashforth rule for the remainder of the time steps. We take a time step of size $\Delta t = 0.02$. Checks are performed with smaller time steps and varying initial discretizations of the swarm patch boundary to verify that our solutions are sufficiently well converged.

Despite the fact that the boundary stays smooth, numerical simulations reveal that it develops complex structure (as we show below), which is also a feature of vortex patches. As the swarm patch evolves, it may be necessary to rediscretize the boundary in order to have an accurate solution, i.e., to have a fine enough mesh to capture new spatial complexity. Sophisticated "contour surgery" techniques have been developed to address this issue for vortex patches; see, for example, [9]. We use a simpler technique, which we apply at every time step. Nodes which are adjacent with respect to the Lagrangian parameter α are checked for spacing. If the Euclidean distance becomes too large, a node is inserted between them using linear interpolation. Similarly, if nodes become too close together, they are replaced with a node whose position is the spatial average of the original ones. While this latter step discards detail below a certain length scale, we perform it nonetheless so that the total number of nodes does not grow so quickly as to make the computation prohibitively slow. Finally, we periodically perform a check to verify that the swarm-patch boundary is not self-intersecting. (Self-intersection of the boundary would break the uniqueness of particle paths which the problem must obey.) If the boundary is found to self-intersect, the simulation is aborted and must be repeated with a finer threshold of spatial detail.

We note that, by symmetry arguments, a rotating disk is an exact solution to (4.2) (though the rotation is not solid-body rotation). This is true for fluid vortex patches as well; see [16] for a discussion. We assume that N is the Gaussian interaction function given by (4.3) and calculate the resulting velocity of points on the swarm boundary, a circle of radius R. After some algebra, we find the velocity $|\vec{v}(R)|$, from which we may compute the period of rotation

$$T(D) = \frac{2\pi R}{|\vec{v}(R)|} = \frac{d^2}{\rho_0} e^{\frac{2R^2}{d^2}} \left\{ I_1\left(\frac{2R^2}{d^2}\right) \right\}^{-1}, \tag{4.7}$$

where I_1 is the modified Bessel function of the first kind of order one. Figure 4.1 shows the period of rotation of the boundary, $T(d)$, for a swarm patch of radius $R = 1$ and population density $\rho_0 = 1$. The exact expression (4.7) is plotted as a line, while data from a numerical simulation, obtained by tracking one of the Lagrangian nodes, is plotted as dots. These results not only serve as a check on our algorithm, but also demonstrate our previous conclusions about the behavior of (4.2) in the limits $d \to 0$ and $d \to \infty$ for the particular case of a circular patch.

This example also illustrates the effect of the interaction function N on the direction of rotation. Since the boundary of the patch has a direction associated with its parameterization, the rotation is clockwise or counterclockwise according to whether the interaction function N is chosen to have, respectively, a positive or negative sign. This limitation results from our choice of kinematic velocity rule. In the case of a

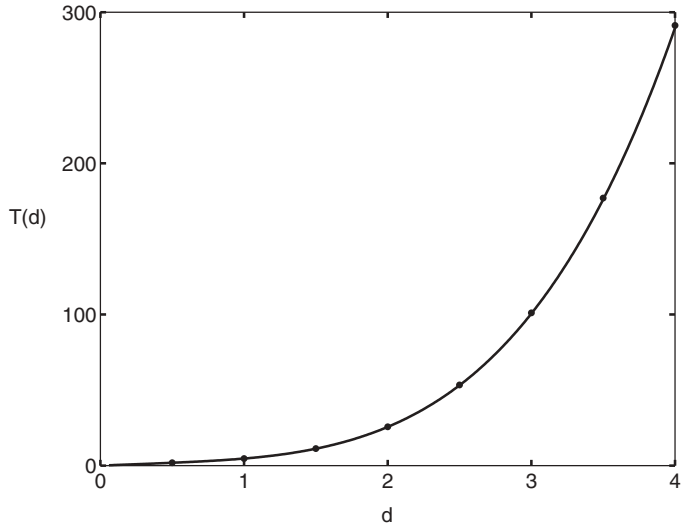


FIG. 4.1. *Period of rotation T for the boundary of a circular swarm patch under the model* (4.2) *with the interaction function* (4.3). *Here, T is shown as a function of d, the interaction length scale in* (4.3). *The line corresponds to the exact expression $T = 2\pi R/|\vec{v}(R)|$ given by* (4.7). *The dots correspond to a numerical simulation of the contour dynamics equation* (4.6). *For this example, the radius of the patch is $R = 1$, and the constant population density is $\rho_0 = 1$.*

dynamic velocity rule, inertial effects would give any initial swarm patch a natural direction of rotation. That is to say, for a dynamic rule, the swarm will have the freedom to nucleate a rotational state, for instance, as seen in simulations of the model in [15]. For simplicity, we present simulations in which N has a positive sign, resulting in clockwise rotation.

We now turn to a discussion of the behavior of the model for other (noncircular) initial conditions and for intermediate values of d when some nontrivial evolution occurs. We find that for the present case of incompressible velocity the dynamics are characterized by an overall rotational motion. The solutions at sufficiently long times are vortex-like, as we now illustrate with several examples.

Figure 4.2 shows the evolution of a swarm patch using the interaction function (4.3) with $d = 1$. The initial boundary of the population is the polar curve $r(\theta) = 1 + (1/10)(\cos 4\theta)$. The square-like initial swarm patch experiences clockwise rotation. At time $t = 1$, the beginnings of spiral arms are visible at the corners of the patch, where the initial curvature was greatest. By time $t = 3$, the spirals have grown longer and the core of the patch is becoming circular. This trend continues through the end of the simulation at $t = 10$, at which point the spiral arms have grown even longer and the core is nearly a perfect circle.

The evolution is qualitatively similar even for swarm patches whose initial shape is far from circular. For instance, Figure 4.3 shows the evolution of an elongated swarm patch, again with the interaction function (4.3) and $d = 1$. The boundary of the initial shape is an ellipse with a major semiaxis of 1 and a minor semiaxis of 0.1. As with the previous example, the swarm patch rotates in a clockwise direction. Spiral arms develop at the points of greatest curvature, and there is a movement of the population towards a developing circular core, which is noticeable at time $t = 8$ and well defined by $t = 10$.

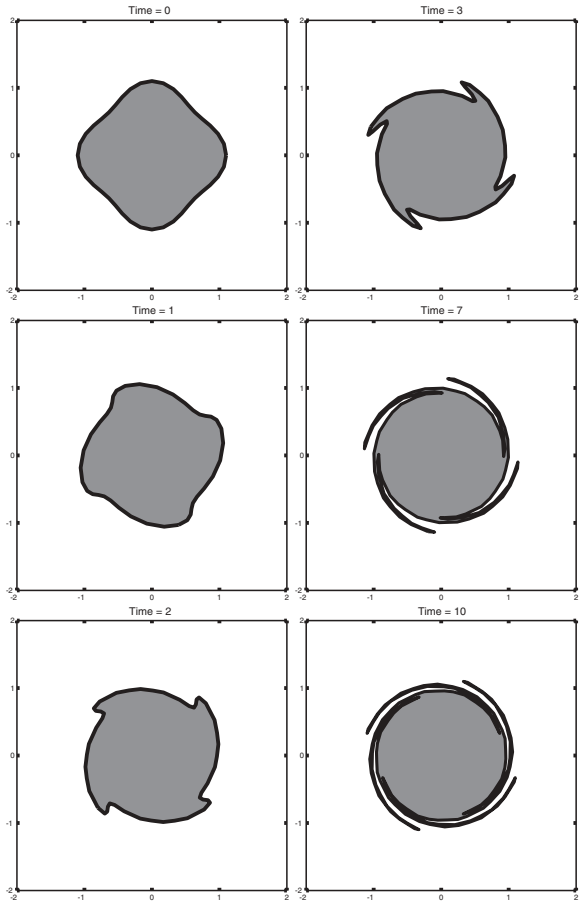


FIG. 4.2. *Evolution of a swarm patch under model* (4.2). *The boundary of the initial shape is a polar curve with radius given by* $r(\theta) = 1 + (1/10)(\cos 4\theta)$. *The constant population density is* $\rho_0 = 1$. *The interaction function* N *is the Gaussian given by* (4.3) *with interaction length scale* $d = 1$. *The constant density swarm patch rotates clockwise and develops spiral arms.*

Finally, we comment that a similar evolution occurs even for irregularly shaped swarm patches. Figure 4.4 shows the evolution of such a patch with the same interaction function as in the previous two examples. The initial shape is generated by the polar function $r = f(\theta)$, where f is a superposition of cosine components with randomly chosen amplitudes and randomly chosen low-integer frequencies. As with the previous examples, the patch rotates clockwise, developing a circular core and an irregular arrangement of spiral arms.

We have shown in this section that, in the case of incompressible motion, our simple nonlocal kinematic model has constant density solutions of compact support. It was seen directly from (4.2) that the evolution of any initial swarm patch slows for very large or very small values of the interaction length scale d. For intermediate values of d, numerical simulations demonstrated that the evolution of these swarm patches is rotational, with the direction of motion set a priori by the sign on the interaction function N. There is a movement of population towards the rotational center of the swarm, where a circular core develops. Spiral arms form at regions of the boundary

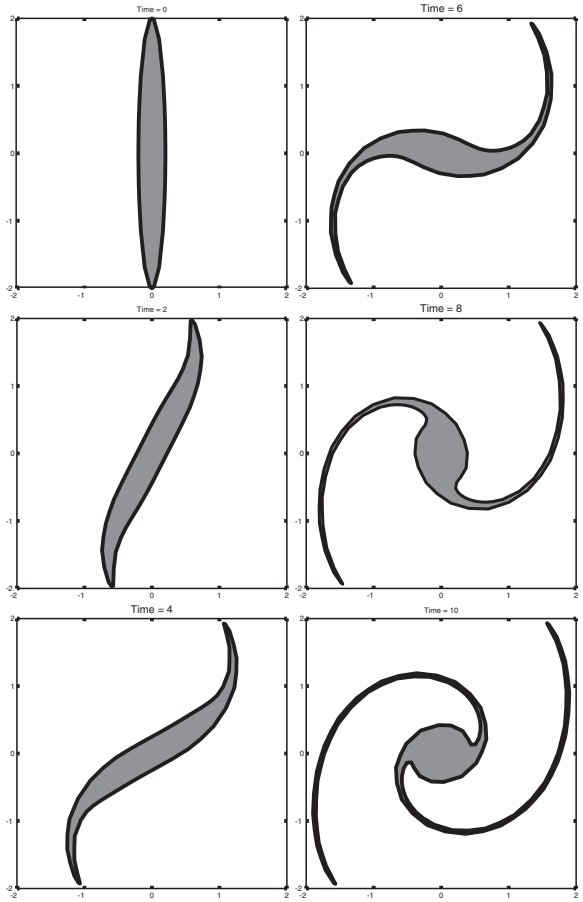


FIG. 4.3. *Evolution of an elongated swarm patch under model* (4.2). *The interaction function* N *is the Gaussian given by* (4.3) *with interaction length scale* $d = 1$. *The constant population density is* $\rho_0 = 1$. *The initial shape is an ellipse with a major semiaxis of* 1 *and a minor semiaxis of* 0.1. *The patch eventually evolves into a circular core with two spiral arms.*

where the curvature is very high. We saw that all of our numerical simulations resulted in asymptotic vortex states reminiscent of those observed in bacteria [3], fish [23], ants [26], cellular slime molds [25], and zooplankton [22]. Similar vortex-like states have also been observed in discrete dynamic swarming models such as those in [15, 8]. Of course, these discrete dynamic models are quite different from the model we study here. As previously discussed, the direction of swarm rotation is necessarily determined by the sign of the kernel in our kinematic model, whereas in dynamic models, the swarm has the capability to nucleate a natural direction of rotation. The work in [17] alludes to this difference between kinematic and dynamic models in the context of one-dimensional motion; the difference is more dramatically highlighted by considering the more general two-dimensional case.

We close this section by discussing the biological significance of incompressible velocities. A common characteristic of biological swarms is their ability to move and evolve in shape while maintaining a constant density. Within the class of kinematic models, the general incompressible type of model we consider here is the only one

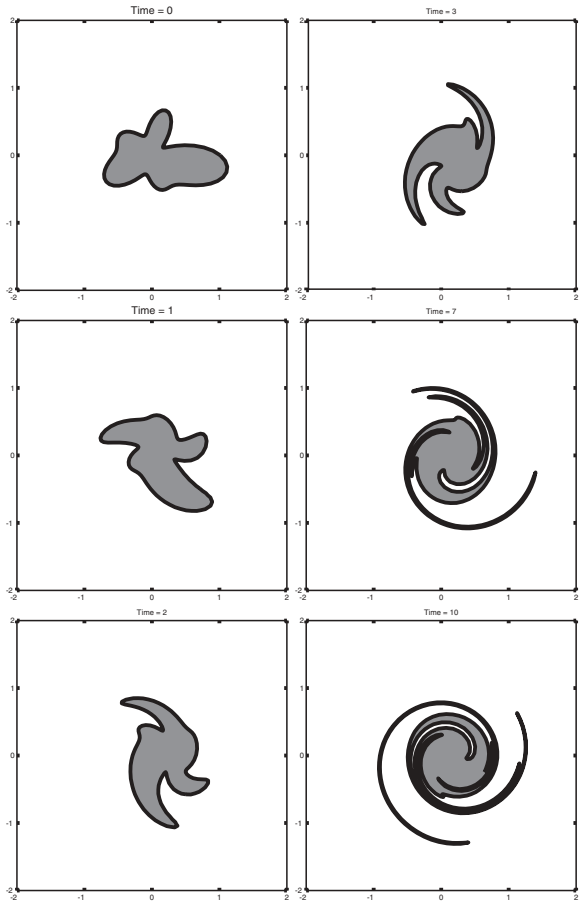


FIG. 4.4. *Evolution of an irregular swarm patch under model* (4.2). *The interaction function* N *is the Gaussian given by* (4.3) *with interaction length scale* $d = 1$. *The constant population density is* $\rho_0 = 1$. *The initial shape is generated by the polar function* $r = f(\theta)$, *where* f *is a superposition of cosine components with randomly chosen amplitudes and randomly chosen low-integer frequencies. The patch eventually evolves into a circular core with an irregular arrangement of spiral arms. We begin the simulation with* 84 *Lagrangian nodes on the swarm-patch boundary and end with* 485 *nodes.*

which captures precisely this effect. Stated differently: for a constant density swarm in which attractive and repulsive effects are in balance, organisms wishing to maintain constant population density *must* move with incompressible velocities. Potential velocities, discussed in the next section, will lead to variations in population density. Thus, one may think of the incompressible velocity terms as those which model the aggregate cooperative dynamics of organisms striving to maintain equal spacing.

5. Potential motion. Some models based on potential forces are reviewed in [21, 19]. For our model, potential motion means that

$$\vec{K} = \nabla P, \tag{5.1}$$

so that the model (3.1)–(3.2) may be written compactly as

$$\rho_t + \nabla \cdot [\rho(\nabla P * \rho)] = 0. \tag{5.2}$$

We take the scalar interaction function $P = \mp G_d$, where G_d is the Gaussian distribution of width d given by (4.3). In section 4, we made the same choice for the scalar interaction function N. In that case, the sign of N simply determined the direction of rotation of the swarm. For the present case, we will see that the sign of P has a much more dramatic effect on the evolution of the population. Specifically, it will determine whether organisms disperse or aggregate, as we discuss in the following two subsections.

5.1. Dispersion. Here, we take $P = -G_d$. Note that (5.2) has an analogy to Darcy's law for flow in porous media. In fact, in the limit $d \to 0$, $-G_d$ becomes a δ-function of strength π, and the governing equation (5.2) is the porous media equation. This is a well studied partial differential equation, and it possesses an exact self-similar solution, called Barenblatt's solution. For an initial population of size Q placed at the origin, Barenblatt's solution is

$$\rho(r,t) = \begin{cases} \frac{1}{2\pi}\sqrt{\frac{Q}{t}} - \frac{r^2}{8\pi t}, & r \le 2\,(Qt)^{1/4}, \\ 0, & r > 2\,(Qt)^{1/4}. \end{cases} \tag{5.3}$$

A discussion of the porous media equation as it relates to biological dispersal, along with a more general statement of Barenblatt's solution, may be found in [19]. In the opposite limit $d \to \infty$, i.e., when social interactions are extremely nonlocal, a bit of algebra again reveals that the equation becomes simply the steady state $\rho_t = 0$. The intuitive statement of this limiting case is similar to that in the previous section: organisms can sense population gradients infinitely far away, but these gradients have no influence on velocity because the strength of the social interactions is infinitely weak.

For intermediate values of the interaction length scale d, the population density profile experiences diffusion and convection. We may understand this better by writing the governing equation (5.2) in an alternate form. After some algebra, we obtain

$$\rho_t = \nabla\rho \cdot (\nabla G_d * \rho) + \rho(\nabla^2 G_d * \rho). \tag{5.4}$$

The first term on the right-hand side of (5.4) is convective and, due to the single derivative on G, scales like $1/d^4$. In contrast, the second term is diffusive and scales like $1/d^6$. Thus, for a given population density profile, as d is increased, we expect that convection will be more dominant than diffusion.

We demonstrate the role played by the interaction length scale d by means of numerical simulations. For these examples, we focus on a radially symmetric model, so that the density $\rho(r,t)$ is a function of the radial coordinate r, and the velocity v is as given above. Note that if ρ is radially symmetric and P is also radially symmetric, then the velocity field for this gradient flow points in the radial direction and is itself radially symmetric. We solve the governing equation on the unit disk with boundary condition $\rho = 0$ on the circumference. We use MacCormack's method, which is second-order accurate in space and time; see, for instance, [28]. We use $n = 64$ grid points with time steps of $\Delta t = 1 \times 10^{-5} - 5 \times 10^{-4}$. Checks are performed with finer meshes in space and time to verify that solutions are sufficiently well converged.

Two example simulations are shown in Figure 5.1. We choose as a random initial condition the function

$$\rho(r,0) = f(r)\left\{\frac{1}{2} + \frac{1}{2}\tanh(5 - 15r)\right\}. \tag{5.5}$$

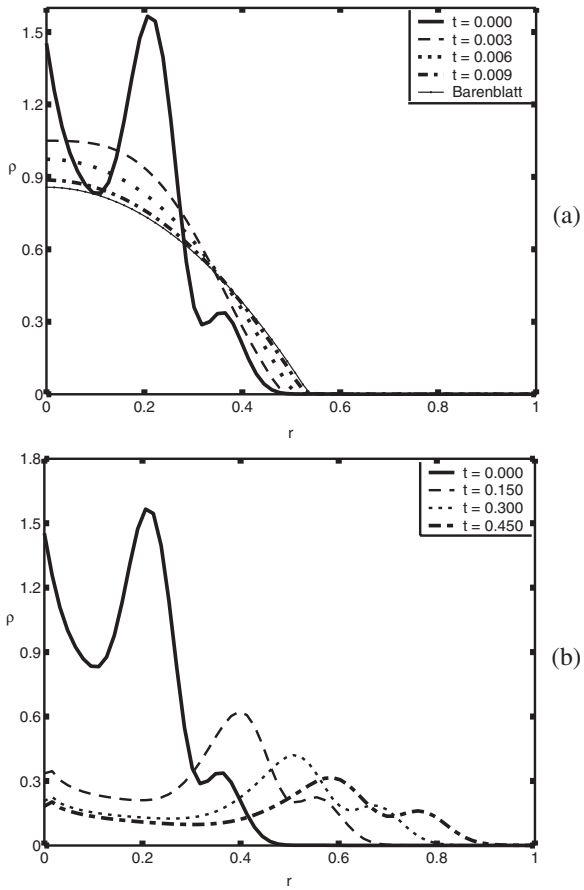


FIG. 5.1. *Time evolution of an identical random initial condition under* (5.2) *for two different social interaction length scales in a radially symmetric geometry with the potential interaction function* $P = -G_d$ *given by* (4.3). *The graphs show the population density* ρ *versus the radial coordinate* r. (a) *Interaction length scale* $d = 0.01$. *Interactions are very localized, and the dynamics are similar to those of the porous media equation. The curve labeled "Barenblatt" is a snapshot of the self-similar solution to the porous media equation given by* (5.3) *and should be compared to the numerical solution at time* $t = 0.009$. *(See text for details).* (b) *Interaction length scale* $d = 0.5$. *For this case of more nonlocal interactions, the population is convected away from the origin, and the smoothing of the population density profile occurs more slowly.*

Here, $f(r)$ is created by superposing Fourier modes with low integer wave numbers and random coefficients. Multiplication by the bracketed combination is carried out so that the initial condition decays smoothly towards zero. For the first example, this state is evolved with $d = 0.01$, so that social interactions are only very slightly nonlocal. In this case, (5.2) is nearly the porous media equation, and the "bumpy" initial condition quickly smooths out and approaches the parabolic profile given by Barenblatt's solution, which we verify in the following manner. We fit the numerical solution at time $t = 0.006$ to (5.3). Then the numerical solution is evolved numerically, and the fit to Barenblatt's solution is evolved analytically. The two are compared again at time $t = 0.009$. Both curves are contained in Figure 5.1(a). The curves nearly overlay each other, and the maximum error between the two is 3%.

For contrast, we have taken the same random initial condition and integrated it

with the more nonlocal interaction length scale $d = 0.5$. In this case, shown in Figure 5.1(b), Fourier modes are damped much more slowly, and the bumpy initial conditions retain their shape much longer. The motion of the swarm is much more convective, and the population is transported away from the origin.

5.2. Aggregation. In this case, we take $P = G_d$. Whereas P was strictly negative in the previous subsection, it is now strictly positive, and this change has dramatic consequences for the dynamics. Now the governing equation states that velocities are up, rather than down, population gradients (nonlocally), so that the population will tend to form groups.

We may understand this grouping by means of a linear stability analysis. To do so, we consider small perturbations $\hat{\rho}$ to a constant density steady state ρ_0. Linearizing (5.2), we obtain

$$\hat{\rho}_t = -\rho_0 G_d * \nabla^2 \hat{\rho}, \tag{5.6}$$

and thus we see that the perturbation obeys a nonlocal backwards heat equation. Taking a Fourier ansatz for the perturbation, i.e., $\hat{\rho} = \hat{\rho}_0 e^{i(\vec{k}\cdot\vec{x}+\sigma t)}$, we find that the linear growth rate is given by

$$\sigma(k) = \pi \rho_0 k^2 e^{\frac{-k^2}{(4d^2)}}, \tag{5.7}$$

where $k = |\vec{k}|$. By computing the critical points of (5.7), we see that the most unstable modes are those with wave number $k_u = 2/d$. The growth of this most unstable mode provides a mechanism for the clumping of organisms. We expect that extremely localized interactions will lead to the formation of a larger number of small groups, i.e., a density distribution pattern with a small characteristic length scale. On the other hand, more nonlocal interactions will result in a smaller number of large groups, i.e., a density distribution pattern with a larger characteristic length scale.

We confirm this prediction by means of numerical simulations. Using a pseudospectral Fourier method with 128 modes on each axis, we integrate (5.2) on a $2\pi \times 2\pi$ box with periodic boundary conditions. We choose the initial density distribution to be $\rho = 1$ plus a small random perturbation constructed by superposing low wave number ($k < 15$) Fourier modes with random coefficients. For time-stepping, we initialize with a forward Euler step and then use a second-order Adams–Bashforth method. Depending on the value of the interaction length scale d, we take time steps of $\Delta t = 4 \times 10^{-5} - 1 \times 10^{-3}$. Checks are performed with different numbers of modes and different time steps to verify convergence.

Our results are shown in Figure 5.2. Figure 5.2(a) shows the initial condition used for the two simulations. Dark patches correspond to regions of higher density. Figure 5.2(b) shows the center of the power spectrum of the initial perturbation, which is noisy. Figure 5.2(c) shows the evolution of the state in Figure 5.2(a) at time $t = 0.132$ with interaction length scale $d = 0.4$. Notice the patches of high population density. By the linear stability arguments given above, the characteristic wave number of the grouping pattern is predicted to be $k_u = 2/d = 5$. Figure 5.2(d) shows a blow-up of the center part of the power spectrum of the evolution of the perturbation. As predicted, the strongest peaks are centered around the circle $k = 5$. Figures 5.2(e) and 5.2(f) are analogous to 5.2(c) and 5.2(d), but at time $t = 2.74$ and with the more nonlocal interaction length scale $d = 1$. In this case, fewer groups form, and they are larger. The most unstable wave number from linear analysis is $k_u = 2$, and indeed

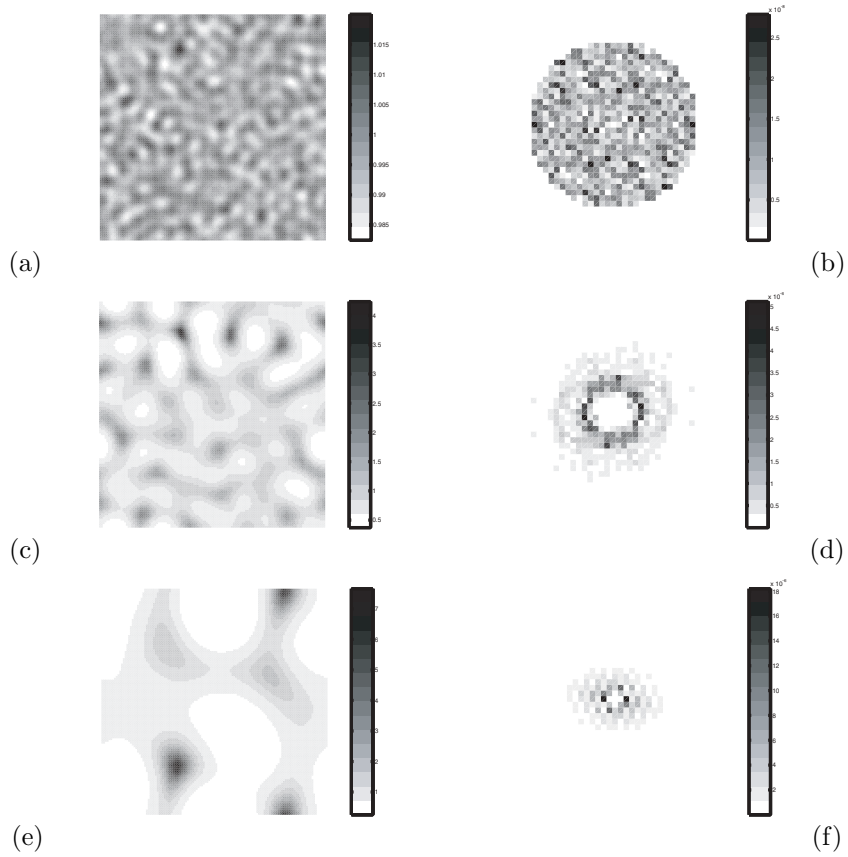


FIG. 5.2. *Results from integrating* (5.2) *on a* $2\pi \times 2\pi$ *periodic box with interaction function* $P = G_d$ *given by* (4.3). (a) *Initial population density, given by* $\rho = 1$ *plus a small random perturbation.* (b) *Center of the power spectrum of the initial perturbation.* (c) *Population density at* $t = 0.132$ *with interaction length scale* $d = 0.4$. *Note the formation of small, high-density groups.* (d) *Center of the power spectrum of the perturbation at* $t = 0.132$. *The strongest peaks are at wave number* $k = 5$, *which is the most unstable mode as predicted by linear analysis. Figures* (e) *and* (f) *are analogous to* (c) *and* (d), *but data is taken at* $t = 2.74$, *and the longer interaction length scale* $d = 1$ *is used. In this more nonlocal case, the most unstable wave number is* $k = 2$, *and the population clumps into fewer, larger groups.*

this is the wave number corresponding to the strongest peaks in the power spectrum. Finally, we comment that these simulations are not continued for longer times because they experience exponential blow-up, due to the lack of any effects to counterbalance the attractive forces in the model. See the appendix for a mathematical discussion of the blow-up. Modifications to the model to prevent blow-up will be a key aspect of future work, as mentioned in the next section.

6. Conclusions. Our work on biological groups in two dimensions is in the spirit of the one-dimensional study in [17]. The overarching goal in this paper is to make specific statements about how social interactions between organisms affect the large-scale motion of a biological group. In summary, we formulated and studied a simple kinematic continuum model which includes nonlocal spatially decaying social interactions between individuals. We decomposed the dynamics of our model into

incompressible motion and potential motion, or alternatively, motion perpendicular to population gradients and motion along gradients.

Whereas a constant drift and cancellation of nonlocal effects are necessary for maintaining a cohesive swarm in one dimension, this is not the case in two dimensions. For the special case of an incompressible kernel, where organisms move perpendicular to population gradients (in a nonlocal sense), the equations have constant density solutions of compact support. Through numerical simulations beginning with a variety of initial conditions, we showed that the dynamics for incompressible interactions are rotational, and that swarm patches eventually develop vortex-like structure. This rotational motion is a cooperative mechanism by which a swarm may maintain motion and cohesion once potential (attractive and repulsive) effects have come into balance. The sign of the social interaction term determines the direction of rotation, and the social interaction length scale determines the degree of macroscopic group movement. The observed asymptotic vortex states are intriguingly similar to actual mill vortices seen in biological systems [3, 22, 23, 25, 26].

In contrast, potential kernels model repulsion or attraction between organisms. For the repulsive case, the interaction length scale determines a balance between diffusive motion and convective motion. Very localized interactions lead to greater smoothing, while more nonlocal interactions result in slower smoothing but more outward convection motion. For the attractive case, the length scale determines a most-unstable mode, the growth of which results in the clumping of the population into regions of high and low density.

This work leaves open many possibilities for future research using nonlocal continuum models and ideas from fluid dynamics. One route would be to conduct fully two-dimensional simulations of biological groups under the simultaneous influence of incompressible and potential interactions. Another would be to relax some of the simplifying modeling assumptions made in section 3. The focus would be on more complicated velocity rules containing both nonlocal and local components, each of which may be nonlinear. Additional future work might consider dynamic, rather than kinematic, velocity rules. These rules would take into account inertial forces that might capture "phase changes" in animal group behaviors such as the transition from milling to translational motion.

While a model ultimately should base interaction rules on specific biological socialization functions of the organisms, field and laboratory data leading to such models is, unfortunately, very limited. We hope that the general discussion of this paper will help to focus further research in this direction. Furthermore, our results do suggest some additional measurements for future experimental work. One such measurement is motivated by our observation that (at least within the simple class of kinematic models we have considered here) we see quite different kinds of patterns depending on whether we consider incompressible or compressible motion. Thus, measuring local density variations for different swarm morphologies might elucidate the relationship between typical swarming patterns (like the vortex states) and the mechanisms that create them. Additionally, careful measurements of the nucleation of such patterns will also help in understanding the dynamics of their creation and their robustness to disturbances and other influences.

Appendix. In this appendix we sketch proofs for two results mentioned in the body of this paper, namely regularity of a swarm-patch boundary for the model examined in section 4 and an exponential upper bound for the blow-up of the model examined in section 5.2.

A.1. Regularity of the boundary of swarm patches. In this subsection we discuss the swarm-patch model of section 4 and the regularity of the swarm-patch boundary. The boundary parametrized by $\vec{z}(\alpha,t)$ is a solution of the integrodifferential equation

$$\frac{d\vec{z}}{dt} = \vec{v}(\vec{z},t), \quad \vec{v}(\vec{x},t) = \vec{K} * \rho(\vec{x},t), \quad \rho(\vec{x},t) = \chi_\Omega, \tag{A.1}$$

where $\Omega(t)$ is the interior of the swarm patch.

If we consider the problem as an integrodifferential equation for ρ,

$$\frac{d\rho}{dt} + \vec{v}\cdot\nabla\rho = 0, \qquad \vec{v} = \vec{K} * \rho, \tag{A.2}$$

where $\vec{K} = \nabla^\perp N$ for a smooth radial function N decaying at infinity, then the swarm patch is an example of a weak solution of this problem with initial condition $\rho_0 \in L^1 \cap L^\infty(\mathbb{R}^2)$. Existence and uniqueness of this problem can be proved following the classical theory of Yudovich for vortex patches [32] which is also detailed in [16]. Such a discussion is beyond the scope of this paper. However, we present some straightforward estimates that can be used to prove that the boundary of the patch remains smooth if initially smooth, as in the case of vortex patches for which the kernel $\vec{K}$ is more singular (and the proof is correspondingly more difficult).

If the swarm density persists as the characteristic function of a domain $\Omega(t)$, then it is uniformly bounded in $L^1 \cap L^\infty$ for all time. Since $\vec{v} = \vec{K} * \rho$ with $\vec{K} \in C^\infty$, then we have an a priori bound for all derivatives of $\vec{v}$, $D^k\vec{v} < \infty$ for all multi-indices k. Smoothness of the patch boundary then follows from the fact that the map $\vec{z}$ satisfies the ODE (A.1) with initial condition

$$\vec{z}|_{t_0} = \vec{z}_0(\alpha) \tag{A.3}$$

for smooth $\vec{z}_0$. Since $\vec{v}$ is C^∞ by standard regularity theory for solutions of ODEs, we see that $\vec{z}$ itself is smooth. Note that the mapping $\vec{z}$ cannot develop a critical point at a later time t because the Lagrangian derivative of $\vec{z}$, $\vec{z}_\alpha$ satisfies the ODE

$$\frac{d\vec{z}_\alpha}{dt} = \nabla\vec{v}\,|_{\vec{z}(\alpha,t)}\ \vec{z}_\alpha. \tag{A.4}$$

Since $\nabla\vec{v}$ is bounded for all time, $\vec{z}_\alpha$ remains bounded away from zero and infinity if it is so bounded at time zero, by Grönwall's lemma.

A.2. Boundedness of the swarm density for a general velocity rule. In section 5.2 we presented numerical computations that showed $\rho(\vec{x},t)$ can exhibit blow-up when the convolution kernel $\vec{K} = \nabla P = \nabla G_d$ is positive. In the limit as $d \to 0$ this formally corresponds to a backward time porous media equation.

Here we derive an a priori bound that shows that, for a smooth kernel of any sign, the maximum of ρ is bounded by an exponential in time. Thus the blow-up seen numerically must be an infinite time blow-up, not finite time. The bound we derive depends on the L^∞ norm of the convolution kernel. Thus as $d \to 0$, the bound itself becomes unbounded, as it should because we are approaching the ill-posed limit in the positive kernel case.

In Eulerian coordinates, ρ satisfies a reaction convection equation

$$\rho_t + \vec{v}\cdot\nabla\rho = -\rho\nabla\cdot\vec{v}. \tag{A.5}$$

This problem can be transformed to Lagrangian coordinates using the method of characteristics. Let $\vec{X}(\alpha, t)$ denote the solution of the ODE

$$\frac{d\vec{X}}{dt} = \vec{v}(\vec{X}, t), \quad \vec{X}|_{t=0} = \alpha. \tag{A.6}$$

Then in the Lagrangian coordinate α, ρ satisfies

$$\frac{d\rho}{dt} = -(\nabla \cdot \vec{v})|_{\vec{X}(\alpha,t)}\rho. \tag{A.7}$$

Thus we have a differential inequality for the L^∞ norm of ρ,

$$\frac{d}{dt}\|\rho\|_{L^\infty} \le C\|\nabla \cdot \vec{v}\|_{L^\infty}\|\rho\|_{L^\infty}. \tag{A.8}$$

Since ρ is a density, it has an a priori L^1 bound,

$$\int \rho(\vec{x}, t)d\vec{x} = \int \rho(\vec{x}, 0)d\vec{x} = \|\rho\|_{L^1}. \tag{A.9}$$

Since $\vec{v} = \vec{K} * \rho$ for smooth $\vec{K}$, we have

$$\|\nabla \cdot \vec{v}\|_{L^\infty} \le \|\nabla \cdot \vec{K}\|_{L^\infty}\|\rho\|_{L^1}. \tag{A.10}$$

Combining this with (A.8) and the a priori bound on the L^1 norm of ρ, we have

$$\frac{d}{dt}\|\rho\|_{L^\infty} \le C\|\rho\|_{L^1}\|\rho\|_{L^\infty}, \quad C = \|\nabla \cdot \vec{K}\|_{L^\infty}. \tag{A.11}$$

Grönwall's lemma then gives

$$\|\rho\|_{L^\infty} \le \exp\,(C\|\rho_0\|_{L^1}t)\|\rho_0\|_{L^\infty}. \tag{A.12}$$

Acknowledgments. We are grateful to Daniel Marthaler for providing input on numerical simulations, and to Anke Ordemann for numerous helpful comments.

REFERENCES

[1] M. Aldana and C. Huepe, *Phase transitions in self-driven many-particle systems and related non-equilibrium models: A network approach*, J. Statist. Phys., 112 (2003), pp. 135–153.

[2] W. Alt and G. Hoffman, eds., *Biological Motion: Proceedings of a Workshop Held in Königswinter, Germany, March* 16–19, 1989, Lecture Notes in Biomath. 89, Springer-Verlag, Berlin, 1990.

[3] E. Ben-Jacob, I. Cohen, A. Czirók, T. Vicsek, and D.L. Gutnick, *Chemomodulation of cellular movement, collective formation of vortices by swarming bacteria, and colonial development*, Phys. A, 238 (1997), pp. 181–197.

[4] A.L. Bertozzi and P. Constantin, *Global regularity for vortex patches*, Comm. Math. Phys., 152 (1993), pp. 19–28.

[5] E. Bonabeu, M. Dorigo, and G. Theraulaz, *Swarm Intelligence: From Natural to Artificial Systems*, Santa Fe Institute Studies in the Sciences of Complexity, Oxford University Press, New York, 1999.

[6] J.Y. Chemin, *Persistency of geometric structures in bidimensional incompressible fluids*, Ann. Sci. Eć ole Norm. Sup., 26 (1994), pp. 517–542.

[7] M.C. Cross and P.C. Hohenberg, *Pattern formation outside of equilibrium*, Rev. Mod. Phys., 65 (1993), pp. 851–1112.

[8] I.D. Couzin, J. Krause, R. James, G.D. Ruxton, and N.R. Franks, *Collective memory and spatial sorting in animal groups*, J. Theoret. Biol., 218 (2002), pp. 1–11.
[9] D. G. Dritschel, *Contour dynamics and contour surgery: Numerical algorithms for extended, high-resolution modeling of vortex dynamics in two-dimensional, inviscid, incompressible flows*, Comp. Phys. Rep., 10 (1989), pp. 77–146.
[10] L. Edelstein-Keshet, *Mathematical models of swarming and social aggregation*, in Proceedings of the 2001 International Symposium on Nonlinear Theory and Its Applications, Miyagi, Japan, 2001, pp. 1–7.
[11] L. Edelstein-Keshet, J. Watmough, and D. Grunbaum, *Do travelling band solutions describe cohesive swarms? An investigation for migratory locusts*, J. Math. Biol., 36 (1998), pp. 515–549.
[12] U. Erdmann and W. Ebeling, *Collective motion of Brownian particles with hydrodynamic interactions*, Fluct. Noise Lett., 3 (2003), pp. L145–L154.
[13] U. Erdmann, W. Ebeling, and V.S. Anishchenko, *Excitation of rotational modes in two-dimensional systems of driven Brownian particles*, Phys. Rev. E, 65 (2002), paper 061106.
[14] G. Flierl, D. Grünbaum, S. Levin, and D. Olson, *From individuals to aggregations: The interplay between behavior and physics*, J. Theoret. Biol., 196 (1999), pp. 397–454.
[15] H. Levine, W.J. Rappel, and I. Cohen, *Self-organization in systems of self-propelled particles*, Phys. Rev. E, 63 (2001), paper 017101.
[16] A.J. Majda and A.L. Bertozzi, *Vorticity and Incompressible Flow*, Texts Appl. Math., Cambridge University Press, Cambridge, UK, 2002.
[17] A. Mogilner and L. Edelstein-Keshet, *A non-local model for a swarm*, J. Math. Biol., 38 (1999), pp. 534–570.
[18] A. Mogilner, L. Edelstein-Keshet, L. Bent, and A. Spiros, *Mutual interactions, potentials, and individual distance in a social aggregation*, J. Math. Biol., 47 (2003), pp. 353–389.
[19] J.D. Murray, *Mathematical Biology* I: *An Introduction*, 3rd ed., Interdiscip. Appl. Math. 17, Springer, New York, 2002.
[20] A. Okubo, *Diffusion and Ecological Problems*, Springer, New York, 1980.
[21] A. Okubo, D. Grunbaum, and L. Edelstein-Keshet, *The dynamics of animal grouping*, in Diffusion and Ecological Problems, 2nd ed., A. Okubo and S. Levin, eds., Interdiscip. Appl. Math. 14., Springer, New York, 1999, pp. 197–237.
[22] A. Ordemann, F. Moss, and G. Balazsi, *Motions of Daphnia in a light field: Random walks with a zooplankton*, Nova Acta Leopoldina, 88 (2003), pp. 87–103.
[23] J.K. Parrish and L. Edelstein-Keshet, *Complexity, pattern, and evolutionary trade-offs in animal aggregation*, Science, 284 (1999), pp. 99–101.
[24] J.K. Parrish and W.M. Hamner, eds., *Animal Groups in Three Dimensions*, Cambridge University Press, Cambridge, UK, 1997.
[25] W.J. Rappel, A. Nicol, A. Sarkissian, and H. Levine, *Self-organized vortex state in two-dimensional Dictyostelium dynamics*, Phys. Rev. Lett., 83 (1999), pp. 1247–1250.
[26] T.C. Schneirla, *Army Ants: A Study in Social Organization*, W.H. Freeman, San Francisco, 1971.
[27] F. Schweitzer, W. Ebeling, and B. Tilch, *Statistical mechanics of canonical-dissipative systems and applications to swarm dynamics*, Phys. Rev. E, 64 (2001), paper 021110.
[28] J.C. Strikwerda, *Finite Difference Schemes and Partial Differential Equations*, Chapman & Hall, New York, 1989.
[29] J. Toner and Y. Tu, *Flocks, herds, and schools: A quantitative theory of flocking*, Phys. Rev. E, 58 (1998), pp. 4828–4858.
[30] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet, *Novel type of phase transition in a system of self-driven particles*, Phys. Rev. Lett., 75 (1995), pp. 1226–1229.
[31] T. Vicsek, A. Czirók, I.J. Farkas, and D. Helbing, *Application of statistical mechanics to collective motion in biology*, Phys. A, 274 (1999), pp. 182–189.
[32] V. Yudovich, *Non-stationary flows of an ideal incompressible fluid*, Zh. Vychisl. Mat. Mat. Fiz., 3 (1963), pp. 1032–1066.
[33] N.J. Zabusky, M.H. Hughes, and K.V. Roberts, *Contour dynamics for the Euler equations in two dimensions*, J. Comput. Phys., 30 (1979), pp. 96–106.

SIAM J. APPL. MATH.
Vol. 65, No. 1, pp. 175–193

STABILITY OF STATIONARY SOLUTIONS OF THE MULTIFREQUENCY RADIATION DIFFUSION EQUATIONS*

OLE H. HALD† AND ALEKSEI I. SHESTAKOV‡

Abstract. A nondimensional model of the multifrequency radiation diffusion equation is derived. A single material ideal gas equation-of-state is assumed. Opacities are proportional to the inverse of the cube of the frequency. Inclusion of stimulated emission implies a Wien spectrum for the radiation source function. It is shown that the solutions are uniformly bounded in time and that stationary solutions are stable. The spatially independent solutions are asymptotically stable, while the spatially dependent solutions of the linearized equations approach zero.

Key words. stationary solutions, multifrequency radiation diffusion

AMS subject classifications. 85A25, 35B35, 35B41, 35B50, 35K57

DOI. 10.1137/S0036139902407303

1. Introduction. This paper derives asymptotic stability properties for a system of equations modeling the frequency dependent radiation diffusion equation coupled to the matter energy balance equation. In order to obtain a tractable system, we simplify the equations and first derive a system in nondimensional form. Our derived system preserves the salient features of the original dimensional set of equations—in particular, frequency dependent opacities and a nonlinear relationship between the matter temperature and the radiation emission term. We assume matter to consist of a single material characterized by an ideal gas equation-of-state, i.e., a specific energy proportional to temperature. In the analysis, we impose zero flux boundary conditions on the radiation field. However, since we are interested in studying how the two fields (radiation and temperature) equilibrate, and since the radiation diffusion equation is itself derived assuming near isotropy, i.e., short mean free paths, our boundary condition is not overly restrictive.

We begin with the multifrequency radiation diffusion equations (CGS units)

$$\partial_t u_\nu = \nabla D_\nu \nabla u_\nu + c\,\kappa_\nu\,(\,B_\nu - u_\nu\,)\,, \tag{1}$$

$$\rho \partial_t e = -\,c\int_0^\infty d\nu\,\kappa_\nu\,(\,B_\nu - u_\nu\,)\,. \tag{2}$$

Equations (1)–(2) simulate diffusive transport and emission-absorption. The equations do not contain scattering or effects due to fluid motion (advection and spectral shifts.) The variable u_ν (erg sec cm^{-3}) represents the spectral energy density of the

*Received by the editors May 8, 2002; accepted for publication (in revised form) February 26, 2004; published electronically September 24, 2004.

http://www.siam.org/journals/siap/65-1/40730.html

†Mathematics Department, University of California, Berkeley, CA 94720 (hald@math.berkeley.edu). The work of this author was supported in part by the Office of Science, Office of Advanced Scientific Computing Research, Mathematical, Information, and Computational Sciences Division, Applied Mathematical Sciences Subprogram, of the U.S. Department of Energy under contract DE-AC03-76SF00098.

‡Lawrence Livermore National Laboratory, Livermore CA 94550 (shestakov@llnl.gov). The work of this author was performed under the auspices of the U.S. Department of Energy by the University of California Lawrence Livermore National Laboratory under contract W-7405-Eng-48.

radiation field, ν ($\sec^{-1}$) is the frequency coordinate, ρ is the mass density (assumed fixed), and e is the matter specific energy density. The latter's temporal derivative is expressed in terms of the specific heat $c_{\rm v}$ and the matter temperature T,

$$\partial_t e = \frac{\partial e}{\partial T}\,\partial_t T \doteq c_{\rm v}\partial_t T. \tag{3}$$

On the right side of (1), the diffusion coefficient $D_\nu = c/3\kappa_\nu$, where c is the speed of light and κ_ν (cm^{-1}) is the inverse mean free path.[1] (In most applications, D_ν is modified by including the total scattering cross-section (Mihalas and Weibel-Mihalas [9]) and a flux limiter (Lund and Wilson [7].) Section 6 discusses how such generalizations impact our model. The coefficient κ_ν is a complicated function of frequency. In this paper, we consider one valid for free-free transitions [12],

$$\kappa_\nu = (2\pi/3m_e^3k)^{1/2}\,(4Z_e^2e^6N_+N_e/3hc)\,T^{-1/2}\,\nu^{-3}\,, \tag{4}$$

where m_e is the electron mass, k is the Boltzmann constant, Z_e is the ionic charge, e is the fundamental charge, N_+ and N_e are the number densities (# cm^{-3}) of the ions and free electrons, and h is the Planck constant. We focus our attention on a fully ionized hydrogen plasma of density ρ. Thus, $Z_e = 1$, and $N_+ = N_e = \rho\mathcal{A}$, where $\mathcal{A}$ is the Avogadro constant. This implies that

$$\kappa_\nu = \kappa_0\ \nu^{-3}\ \ ,\quad \kappa_0 = 1.34\cdot 10^{56}\,\rho^2/\sqrt{T}. \tag{5}$$

In order to obtain a tractable set of equations, we ignore the weak temperature dependence of κ_0 and, since ρ is assumed to be fixed, absorb $\rho^2/\sqrt{T}$ into κ_0.[2] Thus, taking characteristic values for ρ and T, we let $\kappa_0\rho^2/\sqrt{T} \to \kappa_0$. Since κ_ν has units of cm^{-1}, κ_0 now has units of $\mathrm{cm}^{-1}\ \sec^{-3}$.

Last, we define the radiation emission term B_ν. To simplify the algebra, instead of the usual Planck function we assume a Wien distribution and write

$$B_\nu = (8\pi h/c^3)\,\nu^3\,\exp(-h\nu/kT). \tag{6}$$

The difference between (6) and the Planck function is that the latter replaces the exponential term with $[\exp(h\nu/kT)-1]^{-1}$. For our purposes, this is of little consequence. Both distributions, Planck and Wien, have the same exponential decay at high frequencies. At low frequencies, (6) and the Planck function are proportional to ν^3 and ν^2, respectively. Equation (6) and the Planck function peak at $h\nu/kT = 3$ and 2.82, respectively. Also, the frequency integrated emission term

$$\int_0^\infty d\nu\, B_\nu = 7.00\cdot 10^{-15}\,T^4\,,$$

which is approximately 8% less than the Planck result, $7.56\cdot 10^{-15}\,T^4$. The relative error of the two distributions, (Planck − Wien)/Planck $= \exp(-h\nu/kT)$. Thus, the error is greatest at $\nu = 0$, but at the peak of the Planck function, the error is less than 6%.

We now derive the nondimensional system of equations and begin by choosing characteristic values for the density ρ, specific heat $c_{\rm v}$, and temperature T_0. As

[1] Some authors express the mean free path as $(\rho\kappa_\nu)^{-1}$, where κ_ν (cm^2/g) is the opacity.

[2] Section 6 discusses the effect of keeping the $1/\sqrt{T}$ dependence of κ_ν.

mentioned above, incorporating these values into (5) yields $\kappa_\nu = \kappa_0 \nu^{-3}$. Denoting the normalization values with zero subscripts and the normalized variables with primes, we define $\nu' = \nu/\nu_0$, $T' = T/T_0$, $u' = u/u_0$, $t' = t/t_0$, and if x is the spatial coordinate, $x' = x/x_0$. This implies that $\kappa_\nu = \kappa_0(\nu_0\nu')^{-3}$, $\partial_t = (1/t_0)\,\partial_{t'}$, $\nabla = (1/x_0)\nabla'$, etc. The constants ν_0, t_0, etc. may be expressed in terms of the original three, ρ, c_v, and T_0. The choices

$$
\begin{aligned}
\nu_0 &= kT_0/h\ , \\
t_0 &= \nu_0^3/(c\,\kappa_0)\ , \\
x_0 &= \nu_0^3/(\sqrt{3}\,\kappa_0)\ , \\
u_0 &= 8\pi\, h\,\nu_0^3/c^3\ , \\
R &= \rho\, c_\mathrm{v}\, c^3/(8\pi\, k\,\nu_0^3)
\end{aligned}
\tag{7}
$$

yield the desired system,

$$
\partial_t u = \nabla \cdot \nu^3\, \nabla u + \left(\nu^3 e^{-\nu/T} - u\right)/\nu^3\,, \tag{8}
$$

$$
R\partial_t T = -T + \int_0^\infty (u/\nu^3)\, d\nu\,, \tag{9}
$$

where $u = u(x,t,\nu)$ and $T = T(x,t)$. In (8)–(9) and henceforth, we drop the primes from the nondimensional variables, t, ν, u, etc. Note that the first term on the right side of (9) stems from integrating over frequency, i.e.,

$$
\int_0^\infty d\nu\, (\nu^3 e^{-\nu/T})/\nu^3 = \int d\nu\, e^{-\nu/T} = T\,.
$$

It is instructive to choose characteristic values for T_0, ρ, and c_v and evaluate the normalization constants in (7). For example, the choices $T_0 = 10^5$ °K, $\rho = 10^{-6}$ g cm^{-3}, and $c_\mathrm{v} = 10^8$ erg/(g °K) imply $\nu_0 = 2.08 \cdot 10^{15}$ Hz (≈ 10 eV), $t_0 = 7.12 \cdot 10^{-7}$ sec, $x_0 = 1.24 \cdot 10^4$ cm, $u_0 = 5.59 \cdot 10^{-11}$ erg sec/cm^3, and $R = 85.8$. Note that the above value for ν_0 corresponds to a wavelength $\lambda_0 = 1.44 \cdot 10^{-5}$ cm, which lies just outside the so-called near UV range $2\text{–}4 \cdot 10^{-5}$ cm.

The nondimensional equations (8)–(9) preserve important properties of the original system (1)–(2). Conservation of total energy follows by ignoring boundary fluxes (e.g., by imposing homogeneous Neumann boundary conditions on u) and integrating the total energy density $RT + \int d\nu\, u$ over the spatial domain. The high frequency photons, i.e., u with large ν, in (8)–(9) are characterized by fast transport and slow absorption. In (8), $1/\nu^3$ plays the role of a coupling coefficient, the low frequencies coupling the fastest. Matter preferentially emits radiation into frequencies where $\nu^3\, e^{-\nu/T}$ is maximal.

Before launching into our theme, we summarize previously published related work. Andreev, Kozmanov, and Rachilov [1] consider the system coupling the matter energy equation to either the equation of radiative transfer (i.e., for the radiation intensity I) or to the radiation diffusion equation. They examine the grey, i.e., frequency-integrated, as well as the spectral cases. Andreev, Kozmanov, and Rachilov assume Dirichlet boundary conditions; that is, the incoming intensity is specified, or, in the diffusion limit, the radiation energy density. Isotropic scattering is included in the transfer equation. Assuming existence of solutions, the paper proves a maximum principle (MP), viz., that the solution is bounded by the initial data and the boundary values. The paper does not discuss whether the equilibrium solution is an attractor.

Mercier [8] considers the system coupling the spectral radiative transfer and matter energy equations. The system includes Thomson scattering, i.e., no energy exchange due to scattering. By applying the theory of accretive operators, Mercier shows that (1) a solution exists, (2) the solution is bounded by the initial and boundary data, and (3) as time progresses, the distance between the solution and the equilibrium point decreases, i.e., the latter is an attractor. Mercier makes the following assumptions regarding the behavior of the coefficients of the equations: The matter internal energy has a positive minimum and the specific heat has a finite upper bound. For the opacity κ, there is Lipschitz continuity with respect to T, a finite upper bound for all frequency, and κ decreases with T. And radiation emission κB increases with T. These assumptions are largely satisfied by our model, with the exceptions that our κ diverges with decreasing frequency (as ν^{-3}) and that κ is independent of T.

The existence of an MP has implications for the design of numerical methods for these equations. Ideally, the schemes should also satisfy an MP. In [6], Larsen and Mercier apply the theoretical work of [1] and [8] to analyze the Monte Carlo method of Fleck and Cummings [5]. Larsen and Mercier prove that for sufficiently small time steps, the algorithm does satisfy the MP but that it fails to do so for large time steps. The analysis is confirmed by their numerical simulations and by Fleck and Cummings's own results [5, Figure 4, p. 332].

There are two competing Monte Carlo schemes for the equations of radiative transfer. One is the previously mentioned one of Fleck and Cummings [5]; another is the method of Carter and Forest [2]. Both schemes rewrite the matter energy equation in terms of the integrated radiation energy density function $u_r(T) = aT^4$, where a is the radiation constant. The schemes require a value of u_r throughout the course of the time step, but they differ in how that term is approximated. The choice of approximation impacts the stability of the scheme, i.e., whether or not the numerical solution satisfies the MP. In [3], Densmore and Larsen apply the Larsen and Mercier analysis [6] to the Carter–Forest scheme. Densmore and Larsen's analysis shows that the Fleck–Cummings scheme appears to allow larger time steps than the Carter–Forest scheme before the numerical solution violates the MP. The analysis is confirmed by numerical results on a problem simulating radiative flow into initially cold matter.

In this paper, we extend the work just cited by analyzing the rate at which solutions equilibrate. Since our interests lie with the multifrequency radiation diffusion equations, we restrict our attention to the normalized system derived above.

We now summarize the rest of the paper. Section 2 confirms that solutions are positive and uniformly bounded by the initial conditions. Section 3 shows that the stationary solution is unique and if the initial conditions of a given problem lie within an envelope about the stationary solution, so does the solution for all time. In section 4, we consider a linearized form of (8)–(9) and prove that a small perturbation about the stationary solution decays to zero exponentially fast if the perturbation is spatially varying. Section 5 considers the nonlinear system (8)–(9) and shows that nonspatially varying perturbations also decay to zero. However, we are unable to give an estimate for the rate of decay and surmise that the decay may be slow. A summary, motivation for this work, and a description of ongoing related research appear in section 6. Section 6 also discusses the impact that more realistic assumptions, e.g., nonideal gases, have on our model.

2. A priori bounds. Let U be an open, bounded, and connected subset of $\mathbb{R}^3$ with smooth boundary. We consider the equations

$$\partial_t u(x,t,\nu) + \frac{1}{\nu^3} u(x,t,\nu) = \nu^3 \Delta u(x,t,\nu) + e^{-\nu/T(x,t)}, \tag{10}$$

$$R\,\partial_t T(x,t) + T(x,t) = \int_{\nu=0}^{\infty} \frac{u(x,t,\nu)}{\nu^3} d\nu \tag{11}$$

for $x \in U$, $t > 0$, and $0 < \nu < \infty$ with boundary conditions

$$\frac{\partial u}{\partial n}(x,t,\nu) = 0, \qquad x \in \partial U, \tag{12}$$

and initial conditions

$$u(x,0,\nu) = u_0(x,\nu), \qquad T(x,0) = T_0(x), \qquad x \in U.$$

Here $T(x,t)$ is the matter temperature at position x and time t, $u(x,t,\nu)$ denotes the radiation energy density at frequency ν, and $\Delta u = \sum_{i=1}^{3} \partial_{x_i}^2 u$. The constant R is positive.

LEMMA 1. *Let $u(x,t,\nu)$, $T(x,t)$ be smooth solutions of* (10)–(12). *If*

$$0 \le u(x,0,\nu) \le \nu^3 e^{-\nu/T_2} \qquad \textit{and} \qquad 0 < T(x,0) < T_2,$$

then, for $t > 0$,

$$0 < u(x,t,\nu) < \nu^3 e^{-\nu/T_2} \qquad \textit{and} \qquad 0 < T(x,t) < T_2.$$

Remark. Lemma 1 shows that the solutions of (10)–(12) are uniformly bounded in time. This is to be expected as the temperature and the radiation energy are nonnegative quantities and there is no energy flowing through the boundary. The result is also valid when U is a box with periodic boundary conditions.

Proof. The idea is to rewrite (10)–(11) as integral equations and use proof by contradiction. It follows from (10) that for each $\nu > 0$,

$$\partial_t (e^{t/\nu^3} u) = \nu^3 \Delta (e^{t/\nu^3} u) + e^{t/\nu^3} e^{-\nu/T}.$$

Using Duhamel's principle [4, p. 49], we obtain

$$\begin{aligned} u(x,t,\nu) = {} & e^{-t/\nu^3} \int_U G_\nu(x,y,t)\, u(y,0,\nu) dy \\ & + \int_{s=0}^{t} e^{-(t-s)/\nu^3} \int_U G_\nu(x,y,t-s) e^{-\nu/T(y,s)} dy\, ds\,. \end{aligned} \tag{13}$$

Here $G_\nu(x,y,t)$ is the Green's function for $\partial_t w = \nu^3 \Delta w$ with Neumann boundary conditions. From (11) we see that

$$\partial_t (e^{t/R} T) = \frac{e^{t/R}}{R} \int_{\nu=0}^{\infty} \frac{u}{\nu^3} d\nu.$$

Integrating with respect to t and using (13) yield

$$
\begin{aligned}
T(x,t) &= e^{-t/R}\,T(x,0) \\
&\quad + \int_{\tau=0}^{t} \frac{e^{-(t-\tau)/R}}{R} \int_{\nu=0}^{\infty} \frac{e^{-\tau/\nu^3}}{\nu^3} \int_U G_\nu(x,y,\tau) u(y,0,\nu)\,dy\,d\nu\,d\tau \\
&\quad + \int_{\tau=0}^{t} \frac{e^{-(t-\tau)/R}}{R} \int_{\nu=0}^{\infty} \int_{s=0}^{\tau} \frac{e^{-(\tau-s)/\nu^3}}{\nu^3} \int_U G_\nu(x,y,\tau-s) e^{-\nu/T(y,s)}\,dy ds d\nu d\tau. \qquad (14)
\end{aligned}
$$

To bound T we assume that either $\min_{\bar U} T(x,t) = 0$ or $\max_{\bar U} T(x,t) = T_2$ for some $t > 0$. Let t_2 be the smallest such t. Then $0 < T(x,t) < T_2$ for $0 \le t < t_2$. Since $T_0 > 0$, $G_\nu > 0$, $u_0 \ge 0$, $R > 0$, and $e^{-\nu/T} > 0$, for $t < t_2$ we conclude from (14) that $T(x,t_2) > 0$. Thus, the upper bound must hold at $t = t_2$. Let x_2 define the point where $T(x_2,t_2) = T_2$. Recall that $\int G_\nu dy \equiv 1$. Since $u(y,0,\nu) \le \nu^3 e^{-\nu/T_2}$ and $e^{-\nu/T(y,s)} < e^{-\nu/T_2}$, for $s < t_2$ it follows from (14), after integrating over s, that

$$
\begin{aligned}
T_2 = T(x_2,t_2) &\le e^{-t_2/R} T(x_2,0) \\
&\quad + \int_{\tau=0}^{t_2} \frac{e^{-(t_2-\tau)/R}}{R} \int_{\nu=0}^{\infty} e^{-\tau/\nu^3} e^{-\nu/T_2}\,d\nu d\tau \\
&\quad + \int_{\tau=0}^{t_2} \frac{e^{-(t_2-\tau)/R}}{R} \int_{\nu=0}^{\infty} (1 - e^{-\tau/\nu^3}) e^{-\nu/T_2}\,d\nu d\tau\,.
\end{aligned}
$$

Cancelling the terms involving $e^{-\tau/\nu^3}$ and evaluating the remaining integrals yield the contradiction

$$
T_2 \le e^{-t_2/R}\,[T(x_2,0) - T_2] + T_2 < T_2\,.
$$

Thus, the bounds for $T(x,t)$ hold for all time.

To bound $u(x,t,\nu)$, we let $t > 0$ be given and choose $\epsilon > 0$ such that

$$
\epsilon < \min_{y\in\bar U,\ 0\le s\le t} T(y,s) \le \max_{y\in\bar U,\ 0\le s\le t} T(y,s) < T_2 - \epsilon\,.
$$

Since $u(x,0,\nu) \ge 0$, $G_\nu > 0$, and $\int G_\nu dy \equiv 1$, it follows from (13) that $u(x,t,\nu) \ge (1 - e^{-t/\nu^3}) e^{-\nu/\epsilon} > 0$, which establishes the lower bound. For the upper bound, we use the inequalities $u(x,0,\nu) \le \nu^3 e^{-\nu/T_2}$ and $e^{-\nu/T(y,s)} < e^{-\nu/(T_2-\epsilon)}$. Doing the integrals in (13) gives

$$
\begin{aligned}
u(x,t,\nu) &\le e^{-t/\nu^3}\nu^3 e^{-\nu/T_2} + (1 - e^{-t/\nu^3})\nu^3 e^{-\nu/(T_2-\epsilon)} \\
&\le \nu^3 e^{-\nu/T_2} - (1 - e^{-t/\nu^3})\nu^3 e^{-\nu/T_2}\big(\,1 - e^{-\nu\epsilon/[T_2(T_2-\epsilon)]}\,\big) \\
&< \nu^3 e^{-\nu/T_2}\,.
\end{aligned}
$$

This completes the proof. □

3. Stationary solutions. In this section we consider stationary solutions of (10)–(12). We shall show that such solutions cannot depend on x and that they are stable. The problem of asymptotic stability is discussed in sections 4 and 5.

LEMMA 2. *Let $u(x,\nu)$, $T(x)$ be a stationary solution of* (10)–(12) *and satisfy the bounds in Lemma* 1. *Then u, T are independent of x, and there is a constant M such that*

$$
u(\nu) = \nu^3 e^{-\nu/M}, \qquad T = M.
$$

Proof. It is clear that $u = \nu^3 e^{-\nu/M}$, $T = M$ is a stationary solution of (10)–(12). We will show that there are no others. It follows from (10)–(12) that

$$-\nu^6 \Delta u(x,\nu) + u(x,\nu) = \nu^3 e^{-\nu/T(x)},$$

$$T(x) = \int_{\nu=0}^{\infty} \frac{u(x,\nu)}{\nu^3} d\nu, \tag{15}$$

where u satisfies the boundary condition (12). Since $-\nu^6 \Delta w + w = f$ with $\partial w/\partial n = 0$ has a unique solution and $w = 1$ if $f = 1$, we see that

$$u(x,\nu) = \int_U g_\nu(x,y) \nu^3 e^{-\nu/T(y)} dy, \tag{16}$$

$$1 = \int_U g_\nu(x,y) dy. \tag{17}$$

One can show that the Green's function satisfies $g_\nu(x,y) = g_\nu(y,x) > 0$ when $x \neq y$ and is singular when $x = y$. Let $M = \max_{\bar{U}} T(y)$. Then $M = T(x)$ for some x in $\bar{U}$ and by using (17), (15), and (16), we find

$$\begin{aligned} 0 &= M - T(x) \\ &= \int_{\nu=0}^{\infty} e^{-\nu/M} \int_U g_\nu(x,y) dy d\nu - \int_{\nu=0}^{\infty} \frac{1}{\nu^3} \int_U g_\nu(x,y) \nu^3 e^{-\nu/T(y)} dy d\nu \\ &= \int_{\nu=0}^{\infty} \int_U g_\nu(x,y) \left[e^{-\nu/M} - e^{-\nu/T(y)} \right] dy d\nu \,. \end{aligned}$$

Since $g_\nu > 0$ and $[e^{-\nu/M} - e^{-\nu/T(y)}] \geq 0$, we conclude that $T(y) = M$ for all y in U. Equations (16)–(17) then imply that $u(x,\nu) = \nu^3 e^{-\nu/M}$ for all x in U. This completes the proof. □

How do we determine the stationary solution corresponding to particular initial data? Let

$$K(t) = \int_U \left[R\,T(x,t) + \int_{\nu=0}^{\infty} u(x,t,\nu) d\nu \right] dx \tag{18}$$

define the total (matter and radiation) energy for the system. Differentiating with respect to t and using (10)–(12), we see that $K(t)$ is independent of time. If $T(x,t) \to M$ and $u(x,t,\nu) \to \nu^3 e^{-\nu/M}$ as $t \to \infty$, we can find M by solving

$$R\,M + 6M^4 = K(0) \,/\, \mathrm{vol}(U) \,. \tag{19}$$

If $K(0) > 0$, this equation has a unique solution. We will now show that the stationary solution is stable.

THEOREM 1. *Let $u(x,t,\nu)$, $T(x,t)$ be a smooth solution of* (10)–(12). *If*

$$\nu^3 e^{-\nu/T_1} \leq u(x,0,\nu) \leq \nu^3 e^{-\nu/T_2} \qquad \textit{and} \qquad T_1 < T(x,0) < T_2,$$

then for $t > 0$

$$\nu^3 e^{-\nu/T_1} < u(x,t,\nu) < \nu^3 e^{-\nu/T_2} \qquad \textit{and} \qquad T_1 < T(x,t) < T_2.$$

Remark. Equations (18)–(19) show that if u and T satisfy the assumptions in Theorem 1, then $R\,T_1+6T_1^4 < R\,M+6M^4 < R\,T_2+6T_2^4$. Therefore, the temperature M for the stationary solution lies in the interval $T_1 < M < T_2$.

Proof. We have already established the upper bounds in Lemma 1. Our proof for the lower bounds is similar. Assume $\min_{\bar U} T(x,t) = T_1$ for some $t > 0$, and let t_1 be the first such t. Then, $T_1 = T(x_1,t_1)$ for some $x_1 \in \bar U$, and $T(y,s) > T_1$ for $y \in U$ and $s < t$. Since $\int G_\nu dy \equiv 1$, $u(y,0,\nu) \geq \nu^3 e^{-\nu/T_1}$, and $e^{-\nu/T(y,s)} > e^{-\nu/T_1}$ for $s < t_1$, it follows from (14), after integrating over s, that

$$\begin{aligned} T_1 = T(x_1,t_1) &\geq e^{-t_1/R}T(x_1,0) \\ &\quad + \int_{\tau=0}^{t_1} \frac{e^{-(t_1-\tau)/R}}{R} \int_{\nu=0}^{\infty} e^{-\tau/\nu^3} e^{-\nu/T_1} d\nu d\tau \\ &\quad + \int_{\tau=0}^{t_1} \frac{e^{-(t_1-\tau)/R}}{R} \int_{\nu=0}^{\infty} (1-e^{-\tau/\nu^3}) e^{-\nu/T_1} d\nu d\tau. \end{aligned}$$

Cancelling the $e^{-\tau/\nu^3}$ terms and evaluating the remaining integrals, we get

$$T_1 \geq e^{-t_1/R}\,[T(x_1,0) - T_1] + T_1 > T_1\,.$$

The contradiction shows that the lower bound for $T(x,t)$ holds for all time.

To bound $u(x,t,\nu)$ from below we let $t > 0$ be given and choose $\epsilon > 0$ such that

$$T_1 + \epsilon < \min_{y\in\bar U,\ 0\leq s\leq t} T(y,s)\,.$$

Since $u(x,0,\nu) \geq \nu^3 e^{-\nu/T_1}$ and $e^{-\nu/T(y,s)} > e^{-\nu/(T_1+\epsilon)}$, it follows from (13) that

$$\begin{aligned} u(x,t,\nu) &\geq e^{-t/\nu^3}\nu^3 e^{-\nu/T_1} + (1-e^{-t/\nu^3})\nu^3 e^{-\nu/(T_1+\epsilon)} \\ &\geq \nu^3 e^{-\nu/T_1} + (1-e^{-t/\nu^3})\nu^3 e^{-\nu/(T_1+\epsilon)}(1-e^{-\nu\epsilon/[T_1(T_1+\epsilon)]}) \\ &> \nu^3 e^{-\nu/T_1}\,. \end{aligned}$$

This completes the proof. □

4. Linearized equations. In section 3 we have shown that (10)–(12) have a stable stationary solution. The important question is whether the stationary solution is asymptotically stable; i.e., will nearby solutions tend to the stationary solution as time increases? We approach this from two directions. In this section we linearize (10)–(12) around the stationary solution and prove that the linearized equation is asymptotically stable when the initial data are orthogonal to constant functions. In the next section we show that (10)–(12) are asymptotically stable if the initial data are expressed in terms of constant functions.

To get the linearized equations we assume that $u = \nu^3 e^{-\nu/M} + u'$ and $T = M + T'$, where u' and T' are small. Inserting (10)–(12), expanding $e^{-\nu/T}$ in a Taylor series, and discarding the higher order terms lead to the linearized equations

$$\partial_t u'(x,t,\nu) + \frac{1}{\nu^3}u'(x,t,\nu) = \nu^3\Delta u'(x,t,\nu) + \frac{\nu e^{-\nu/M}}{M^2}T'(x,t), \tag{20}$$

$$R\,\partial_t T'(x,t) + T'(x,t) = \int_{\nu=0}^{\infty} \frac{u'(x,t,\nu)}{\nu^3} d\nu, \tag{21}$$

where u' satisfies the original boundary condition (12). Since M satisfies (19), we conclude from (18) that for every $t \geq 0$,

$$\int_U \left[R\,T'(x,t) + \int_{\nu=0}^{\infty} u'(x,t,\nu)d\nu \right] dx = 0.$$

We solve (20)–(21) by separation of variables. Let

$$u'(x,t,\nu) = \sum_{k=0}^{\infty} a_k(t,\nu)e_k(x), \qquad T'(x,t) = \sum_{k=0}^{\infty} \alpha_k(t)e_k(x), \tag{22}$$

where $\Delta e_k + \lambda_k e_k = 0$ in U, $(\partial/\partial n)e_k = 0$ on ∂U, and $\int_U e_i e_j dx = \delta_{ij}$. It can be shown that $\lambda_0 = 0 < \lambda_1 \leq \lambda_2 \leq \cdots$ with $e_0 = 1/\sqrt{\mathrm{vol}(U)}$. Inserting (22) into (20)–(21), we get

$$\frac{d}{dt}a_k(t,\nu) + \left(\frac{1}{\nu^3} + \lambda_k \nu^3\right) a_k(t,\nu) = \frac{\nu e^{-\nu/M}}{M^2}\alpha_k(t), \tag{23}$$

$$R\frac{d}{dt}\alpha_k(t) + \alpha_k(t) = \int_{\nu=0}^{\infty} \frac{a_k(t,\nu)}{\nu^3}\, d\nu, \tag{24}$$

$$R\alpha_0(t) + \int_{\nu=0}^{\infty} a_0(t,\nu)d\nu = 0.$$

Note that the last equation, which follows from energy conservation, (18)–(19), and the orthogonality property of the e_k, does not constrain the coefficients α_k, a_k for $k \geq 1$. Following the proof of Lemma 1, we can show that solutions of (23)–(24) satisfy $|a_k(t,\nu)| \leq \nu^4 M^{-2} e^{-\nu/M} B$ and $|\alpha_k(t)| < B$ for $t > 0$ if the conditions hold at $t = 0$.

To analyze the solutions of (23)–(24) we need the following result.

LEMMA 3. *Let $M, R > 0$. For every $\lambda > 0$ there exist a $\sigma = \sigma(\lambda) > 0$ such that*

$$\int_{\nu=0}^{\infty} \frac{\nu e^{-\nu/M}}{M^2(1 + \lambda\nu^6 - \sigma\nu^3)}\, d\nu = 1 - R\sigma.$$

The function $\sigma(\lambda)$ is an increasing function of λ, and there is an $\eta > 0$ such that $1 + \lambda\nu^6 - \sigma(\lambda)\nu^3 > \eta^2$ for all $\lambda > 0$, $\nu > 0$.

Proof. Since $1 + \lambda\nu^6 - \sigma\nu^3 = 1 - \sigma^2/(4\lambda) + \lambda[\nu^3 - \sigma/(2\lambda)]^2$, we see that the integral—call it $f(\sigma)$—is positive and well defined for $\sigma < 2\sqrt{\lambda}$. It is a smooth increasing function of σ and becomes unbounded as $\sigma \to 2\sqrt{\lambda}$. Using the property that $1 - R\sigma$ decreases with σ and the inequality $0 < f(0) < 1$, we can find a $\sigma = \sigma(\lambda) > 0$ such that $f(\sigma) = 1 - R\sigma$. Note that $\sigma < \min(1/R, 2\sqrt{\lambda})$. Differentiating the identity in Lemma 3 with respect to λ, we obtain

$$\frac{d}{d\lambda}\sigma(\lambda) = \int_{\nu=0}^{\infty} \frac{\nu^7 e^{-\nu/M}\, d\nu}{M^2(1 + \lambda\nu^6 - \sigma\nu^3)^2} \Bigg/ \left(R + \int_{\nu=0}^{\infty} \frac{\nu^4 e^{-\nu/M}\, d\nu}{M^2(1 + \lambda\nu^6 - \sigma\nu^3)^2} \right)$$

and $(d^2/d\lambda^2)\sigma(\lambda) < 0$. This implies that $\sigma(\lambda)$ is an increasing function of λ and concave down. If λ is small and positive, we have

$$\sigma(\lambda) = 0 + \frac{5040M^6}{R + 24M^3}\lambda - \cdots.$$

Let $g(\lambda) = 1 - \sigma^2(\lambda)/(4\lambda)$ with $g(0) = 1$. Then, $g(\lambda) > 0$ for all $\lambda > 0$. Since $\sigma^2(\lambda)/(4\lambda) \to 0$ as $\lambda \to 0$ and $g(\lambda) \geq 3/4$ when $\lambda \geq 1/R^2$, we see that $g(\lambda)$ is continuous at $\lambda = 0$ and conclude that $g(\lambda) > \eta^2$ for all $\lambda \geq 0$. But $1+\lambda\nu^6-\sigma(\lambda)\nu^3 \geq g(\lambda)$. This completes the proof. □

THEOREM 2. *Let $a_k(t,\nu)$, $\alpha_k(t)$ be solutions of (23)–(24) satisfying*

$$|a_k(0,\nu)| \leq \frac{\nu^4 e^{-\nu/M}}{M^2} b_k\,, \qquad |\alpha_k(0)| \leq b_k\,,$$

for $k = 0, 1, \ldots$ and $\nu > 0$. Let $\sigma > 0$ satisfy the nonlinear equation in Lemma 3 with $\lambda = \lambda_1$, where λ_1 is the smallest nonzero eigenvalue of $\Delta e(x) + \lambda e(x) = 0$, $x \in U$, with $(\partial/\partial n)e(x) = 0$ for $x \in \partial U$. If $b_0 = 0$ and $\sum_{k=1}^{\infty} b_k^2 < \infty$, there is a $C > 0$ such that for $t \geq 0$

$$\|u'(x,t,\nu)\|_{L^2(U)} \leq \frac{\nu^4 e^{-\nu/M}}{M^2} C e^{-0.9\sigma t}\,, \qquad \|T'(x,t)\|_{L^2(U)} \leq C e^{-0.9\sigma t}\,.$$

Remarks. (1) The theorem asserts that if u', T' initially consist of only spatially varying components, i.e., $a_0(0,\nu) = \alpha(0) = 0$, the a_k, α_k modes for $k \geq 1$ decay to zero exponentially. Thus, for such perturbations about the stationary solution, the latter is an attractor. (2) If $a_0(0,\nu)$, $\alpha(0) \neq 0$, it is possible to show that $a_0(t,\nu)$ and $\alpha_0(t)$ tend to zero. However, the decay is very slow and requires a different proof. We discuss this in section 5.

Proof. To prove Theorem 2, we derive bounds for a_k, α_k as functions of time. One might think that the higher Fourier coefficients go to zero rapidly, but this is not the case. The relaxation to equilibrium depends on R and the size of the domain and can be slow for large R and large domains.

Let $k \geq 1$ be fixed and replace a_k, α_k, λ_k in (23)–(24) by a, α, λ. We will show that a particular linear combination of a and α goes to zero. Let $\sigma > 0$ be determined as in Lemma 3. If we subtract σa from both sides of (23), divide throughout by $1 + \lambda\nu^6 - \sigma\nu^3$, integrate for $\nu > 0$, and apply Lemma 3, we obtain

$$\begin{aligned}
&\frac{d}{dt}\int_{\nu=0}^{\infty} \frac{a(t,\nu)\,d\nu}{1+\lambda\nu^6-\sigma\nu^3} + \int_{\nu=0}^{\infty} \frac{a(t,\nu)\,d\nu}{\nu^3} \\
&= \int_{\nu=0}^{\infty} \frac{(\nu/M^2)\,e^{-\nu/M}\,d\nu}{1+\lambda\nu^6-\sigma\nu^3}\,\alpha(t) - \sigma\int_{\nu=0}^{\infty} \frac{a(t,\nu)\,d\nu}{1+\lambda\nu^6-\sigma\nu^3} \\
&= (1-R\sigma)\alpha(t) - \sigma\int_{\nu=0}^{\infty} \frac{a(t,\nu)\,d\nu}{1+\lambda\nu^6-\sigma\nu^3}\,.
\end{aligned}$$

Combining this result with (24) leads to

$$\frac{d}{dt}\left[R\alpha(t) + \int_{\nu=0}^{\infty} \frac{a(t,\nu)\,d\nu}{1+\lambda\nu^6-\sigma\nu^3}\right] = -\sigma\left[R\alpha(t) + \int_{\nu=0}^{\infty} \frac{a(t,\nu)\,d\nu}{1+\lambda\nu^6-\sigma\nu^3}\right].$$

We can solve this differential equation and get

$$R\alpha(t) + \int_{\nu=0}^{\infty} \frac{a(t,\nu)\,d\nu}{1+\lambda\nu^6-\sigma\nu^3} = C_1 e^{-\sigma t}, \tag{25}$$

where C_1 is the initial value of the left-hand side.

Our next goal is to derive a differential inequality for

$$A(t) = \int_{\nu=0}^{\infty} \frac{M^2 e^{\nu/M} a^2(t,\nu)\,d\nu}{\nu(1+\lambda\nu^6-\sigma\nu^3)}\,. \tag{26}$$

This time we multiply both sides of (23) by $2\nu^{-1}M^2e^{\nu/M}a/(1+\lambda\nu^6-\sigma\nu^3)$, integrate for $\nu>0$, and obtain

$$\begin{aligned}&\frac{d}{dt}A(t)+2\int_{\nu=0}^{\infty}\frac{M^2e^{\nu/M}a^2(t,\nu)}{\nu^4}\,d\nu\\ &=2\int_{\nu=0}^{\infty}\frac{a(t,\nu)}{1+\lambda\nu^6-\sigma\nu^3}\,d\nu\,\alpha(t)\;-2\sigma A(t)\,.\end{aligned}\tag{27}$$

Using (25) and completing the square, we write the penultimate term as

$$\begin{aligned}&2\int_{\nu=0}^{\infty}\frac{a(t,\nu)}{1+\lambda\nu^6-\sigma\nu^3}\,d\nu\,\alpha(t)\\ &=-\frac{2}{R}\left(\int_{\nu=0}^{\infty}\frac{a(t,\nu)}{1+\lambda\nu^6-\sigma\nu^3}\,d\nu\;-\frac{1}{2}C_1e^{-\sigma t}\right)^2+\frac{C_1^2}{2R}e^{-2\sigma t}\,.\end{aligned}\tag{28}$$

Inserting (26) and (28) into (27) yields

$$\begin{aligned}&\frac{d}{dt}A(t)+2\sigma A(t)+2\int_{\nu=0}^{\infty}\frac{M^2e^{\nu/M}a^2(t,\nu)}{\nu^4}\,d\nu\\ &+\frac{2}{R}\left(\int_{\nu=0}^{\infty}\frac{a(t,\nu)}{1+\lambda\nu^6-\sigma\nu^3}\,d\nu\;-\frac{1}{2}C_1e^{-\sigma t}\right)^2=\frac{C_1^2}{2R}e^{-2\sigma t}\,.\end{aligned}$$

Neglecting the integral terms gives a differential inequality for $A(t)$ with the solution

$$A(t)\le\left(A(0)+\frac{C_1^2}{2R}\,t\right)e^{-2\sigma t}\,.$$

In the arguments below it is inconvenient to carry the term proportional to t. However, as $te^{-\sigma t/5}\le 5/(\sigma e)$ for $t\ge 0$ we can use the weaker result

$$A(t)\le\left(A(0)+\frac{C_1^2}{R\sigma}\right)e^{-1.8\sigma t}\,.\tag{29}$$

Our next task is to estimate $\alpha(t)$ and $a(t,\nu)$. Using Cauchy–Schwarz in (25), we see that

$$\begin{aligned}|\alpha(t)|&=\frac{1}{R}\left|C_1e^{-\sigma t}-\int_{\nu=0}^{\infty}\frac{a(t,\nu)}{1+\lambda\nu^6-\sigma\nu^3}\,d\nu\right|\\ &\le\frac{1}{R}\left[|C_1|e^{-\sigma t}+\left(\int_{\nu=0}^{\infty}\frac{M^2e^{\nu/M}a^2(t,\nu)}{\nu(1+\lambda\nu^6-\sigma\nu^3)}\,d\nu\right)^{1/2}\right.\\ &\qquad\left.\times\left(\int_{\nu=0}^{\infty}\frac{\nu e^{-\nu/M}}{M^2(1+\lambda\nu^6-\sigma\nu^3)}\,d\nu\right)^{1/2}\right].\end{aligned}$$

In the product, the first integral is $A(t)$, while the second is $1-R\sigma$; see (26) and Lemma 3. Since $R\sigma<1$ and $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$, we find from (29) that

$$\begin{aligned}|\alpha(t)|&\le\frac{1}{R}\left[|C_1|e^{-\sigma t}+\left(A(0)+\frac{C_1^2}{R\sigma}\right)^{1/2}e^{-0.9\sigma t}(1-R\sigma)^{1/2}\right]\\ &\le\frac{1}{R}\left[\sqrt{A(0)}+\frac{2|C_1|}{\sqrt{R\sigma}}\right]e^{-0.9\sigma t}\\ &=C_2e^{-0.9\sigma t}\,.\end{aligned}\tag{30}$$

To estimate $a(t,\nu)$ we rewrite (23) as

$$a(t,\nu) = e^{-(\nu^{-3}+\lambda\nu^3)t}a(0,\nu) + \int_{s=0}^{t} \frac{\nu e^{-\nu/M}}{M^2} e^{-(\nu^{-3}+\lambda\nu^3)(t-s)}\alpha(s)ds\,.$$

Using (30) and integrating with respect to s, we obtain

$$|a(t,\nu)| \le e^{-(\nu^{-3}+\lambda\nu^3)t}|a(0,\nu)| + \frac{\nu e^{-\nu/M}}{M^2}\, C_2\, \frac{e^{-0.9\sigma t} - e^{-(\nu^{-3}+\lambda\nu^3)t}}{\nu^{-3}+\lambda\nu^3 - 0.9\sigma}\,.$$

Since $1+\lambda\nu^6-\sigma\nu^3 > \eta^2$ (see Lemma 3), it follows that $-(\nu^{-3}+\lambda\nu^3) < -0.9\sigma$ and

$$\begin{aligned} |a(t,\nu)| &\le \frac{\nu^4 e^{-\nu/M}}{M^2}\left(b_k + \frac{C_2}{\eta^2}\right) e^{-0.9\sigma t} \\ &= \frac{\nu^4 e^{-\nu/M}}{M^2}\, C_3\, e^{-0.9\sigma t}\,. \end{aligned} \tag{31}$$

Next, we derive bounds for $A(0)$, C_1, C_2, and C_3. After setting $t=0$ in (26) and recalling the assumption $|a(0,\nu)| \le \nu^4 M^{-2} e^{-\nu/M} b_k$, it follows from Lemma 3 that

$$\begin{aligned} |A(0)| &\le \int_{\nu=0}^{\infty} \frac{\nu^7\, e^{-\nu/M} b_k^2}{M^2(1+\lambda\nu^6-\sigma\nu^3)}\, d\nu \\ &\le \frac{7!\, M^6\, b_k^2}{\eta^2}\,. \end{aligned}$$

Similar arguments applied to (25) give

$$\begin{aligned} |C_1| &\le R|\alpha(0)| + \int_{\nu=0}^{\infty} \frac{|a(0,\nu)|}{1+\lambda\nu^6-\sigma\nu^3}\, d\nu \\ &\le \left(R + \frac{24\, M^3}{\eta^2}\right) b_k. \end{aligned}$$

Combining these bounds, we conclude from (30) and (31) that

$$\begin{aligned} C_2 &\le \frac{1}{R}\left[\frac{71M^3}{\eta} + \frac{2}{\sqrt{R\sigma}}\left(R + \frac{24\, M^3}{\eta^2}\right)\right] b_k \quad \text{and} \\ C_3 &\le \left(1 + \frac{1}{R\eta^2}\left[\frac{71M^3}{\eta} + \frac{2}{\sqrt{R\sigma}}\left(R + \frac{24\, M^3}{\eta^2}\right)\right]\right) b_k. \end{aligned} \tag{32}$$

So far we have assumed that k is fixed. However, note that M, R, and η are independent of k. Since $\lambda_1 \le \lambda_2 \le \cdots$ and $\sigma(\lambda)$ is an increasing function of λ, we see that $\sigma(\lambda_1) \le \sigma(\lambda_k)$ for $k \ge 1$. Thus, if we replace σ by $\sigma(\lambda_1)$ in the bounds for C_2, C_3, we find that $C_2,\, C_3 \le C_4\, b_k$, where C_4 is independent of k. Observe also that the slowest decay in (30), (31) occurs for $\sigma = \sigma(\lambda_1)$, where λ_1 is the eigenvalue of the lowest spatially varying eigenmode.

Finally, we must reassemble the Fourier series for u', T'. Since $\sum_{k=1}^{\infty} b_k^2 < \infty$, it follows from (22), (31), and (32) that

$$\begin{aligned} \|u'(x,t,\nu)\|_{L^2(U)} &= \left\| \sum_{k=1}^{\infty} a_k(t,\nu)\, e_k(x) \right\|_{L^2(U)} \\ &= \left(\sum_{k=1}^{\infty} a_k^2(t,\nu) \right)^{1/2} \\ &\le \frac{\nu^4 e^{-\nu/M}}{M^2}\, C\, e^{-0.9\sigma t}\,, \end{aligned}$$

where $C = C_4 \left(\sum b_k^2\right)^{1/2}$. We can use the same arguments for T'. This completes the proof. □

5. Asymptotic stability. In section 4, we linearized (10)–(12) around a stationary solution and showed that the solution of the linearized equations tends to zero exponentially. We therefore expect that for every ν the solutions of the original equations will be approximately constant as times increases. In this section we consider a class of functions which are independent of x and show that they tend to a stationary solution. The stationary solution is therefore asymptotically stable for this class of initial data.

THEOREM 3. *Let $u(t,\nu)$, $T(t)$ be independent of x and satisfy* (10)–(11) *and the bounds in Lemma* 1. *If*

$$R\,T(0) + \int_{\nu=0}^{\infty} u(0,\nu)\,d\nu \;=\; R\,M + 6M^4 \tag{33}$$

and

$$\left(\int_{\nu=0}^{\infty} \frac{M^2\,e^{\nu/M}}{\nu}\,|u(0,\nu) - \nu^3 e^{-\nu/M}|^2\,d\nu\right)^{1/2} < \frac{RM}{7}, \tag{34}$$

then

$$\begin{aligned} u(t,\nu) &\to \nu^3 e^{-\nu/M} \qquad && as\ t \to \infty,\ \nu > 0\ fixed, \\ T(t) &\to M && as\ t \to \infty\,. \end{aligned}$$

Proof. As in the proof of Theorem 2 we set $u' = u - \nu^3 e^{-\nu/M}$ and $T' = T - M$, but now we do not assume that u', T' are small. It follows from (10)–(11) and the constraint (33) that

$$\frac{\partial}{\partial t}u'(t,\nu) + \frac{1}{\nu^3}u'(t,\nu) = e^{-\nu/[M+T'(t)]} - e^{-\nu/M}, \tag{35}$$

$$R\frac{d}{dt}T'(t) + T'(t) = \int_{\nu=0}^{\infty} \frac{u'}{\nu^3}\,d\nu, \tag{36}$$

$$R\,T'(t) + \int_{\nu=0}^{\infty} u'(t,\nu)\,d\nu = 0. \tag{37}$$

Equation (37) follows from the energy conservation property. We proceed as in the previous proof. This time we derive a differential inequality for

$$A(t) = \int_{\nu=0}^{\infty} \frac{M^2 e^{\nu/M}}{\nu}|u'(t,\nu)|^2\,d\nu \tag{38}$$

and show that $A(t)$ decays. The function $A(t)$ defined in (38) is the analogue of the one defined in (26), except now we are considering the $\lambda = 0$ case, which according to Lemma 3 corresponds to $\sigma = 0$. Note that (34) and (38) imply that

$$0 < A(0) < (RM/7)^2\,. \tag{39}$$

Using Cauchy–Schwarz in (37) yields

$$\begin{aligned} R|T'(t)| &\le \int_{\nu=0}^{\infty} |u'(t,\nu)|\, d\nu \\ &\le \left(\int_{\nu=0}^{\infty} \frac{\nu e^{-\nu/M}}{M^2}\, d\nu\right)^{1/2} \left(\int_{\nu=0}^{\infty} \frac{M^2 e^{\nu/M}}{\nu} |u'(t,\nu)|^2\, d\nu\right)^{1/2} \\ &= 1 \cdot \sqrt{A(t)}\,. \end{aligned} \tag{40}$$

To analyze (35), we expand the right-hand side using Taylor's formula,

$$\begin{aligned} e^{-\nu/(M+T')} - e^{-\nu/M} &= \frac{\nu e^{-\nu/M}}{M^2}\, T' \\ &+ \frac{1}{2}\, e^{-\nu/(M+\theta T')} \left[\frac{\nu^2}{(M+\theta T')^4} - \frac{2\nu}{(M+\theta T')^3}\right] (T')^2\,, \end{aligned} \tag{41}$$

where $0 < \theta < 1$.

We now prove that $|T'(t)| < M/7$ for all t and do this by contradiction. Equations (40), (38), and (39) show that at least for small t,

$$|T'(t)| \;\le\; R^{-1}\sqrt{A(t)} < M/7.$$

Let $t = t_2$ be the first time for which the inequality fails; thus, $|T'(t_2)| = M/7$. Since $|T'(t)| \le M/7$ for $t \le t_2$, we see that the absolute value of the term multiplying $(T')^2$ in (41) is bounded by

$$N(\nu) = \frac{1}{2}\, e^{-7\nu/(8M)} \left[\frac{\nu^2}{(6M/7)^4} + \frac{2\nu}{(6M/7)^3}\right].$$

Hence, there exists a $\gamma(\nu)$ with $|\gamma| \le 1$ such that

$$e^{-\nu/(M+T')} - e^{-\nu/M} = \frac{\nu e^{-\nu/M}}{M^2}\, T' + \gamma\, N(\nu)\, (T')^2\,.$$

Inserting this expression in (35), multiplying both sides by $\nu^{-1} M^2 e^{\nu/M} u'(t,\nu)$, and integrating over $\nu > 0$, we find

$$\begin{aligned} &\frac{1}{2}\,\frac{d}{dt} \int_{\nu=0}^{\infty} \frac{M^2 e^{\nu/M}}{\nu} |u'|^2\, d\nu + \int_{\nu=0}^{\infty} \frac{M^2 e^{\nu/M}}{\nu^4} |u'|^2\, d\nu \\ &= \left(\int_{\nu=0}^{\infty} u' d\nu\right) T' + \left(\int_{\nu=0}^{\infty} \frac{M^2 e^{\nu/M}}{\nu}\, u'\, \gamma\, N(\nu)\, d\nu\right) (T')^2\,. \end{aligned} \tag{42}$$

Note that the first term is $(1/2)dA/dt$. Let E_1, E_2 denote the terms on the right-hand side of (42). It is clear from (37) that $E_1 = -R\,[\,T'(t)\,]^2$. To estimate E_2, we recall that $|\gamma| \le 1$, apply Cauchy–Schwarz, and get

$$\begin{aligned} |E_2| &\le \left(\int_{\nu=0}^{\infty} \frac{M^2 e^{\nu/M}}{\nu} |u'(t,\nu)|^2\, d\nu\right)^{1/2} \\ &\quad \times \left(\int_{\nu=0}^{\infty} \left[\frac{M\, e^{\nu/(2M)}}{\sqrt{\nu}}\, N(\nu)\right]^2 d\nu\right)^{1/2} (T')^2\,. \end{aligned}$$

The first integral is $A(t)$ (see (38)), while an elementary but tedious calculation shows that the second integral is less than $36/M^2$. Thus, $|E_2| \le (6/M)\sqrt{A(t)}$. Inserting the expressions for E_1, E_2 in (42), we obtain the desired differential inequality

$$\frac{1}{2}\frac{d}{dt}A(t) + \int_{\nu=0}^{\infty} \frac{M^2\, e^{\nu/M}}{\nu^4} |u'(t,\nu)|^2\, d\nu + \left(R - \frac{6}{M}\sqrt{A(t)}\right) |\, T'(t)\,|^2 \le 0. \tag{43}$$

We note that for the linearized problem discussed in the previous section, we would have arrived at the same inequality, but without the term $(6/M)\sqrt{A(t)}$ and with u', T' replaced by a_0, α_0.

Equations (39) and (43) show that $A(t)$ is a decreasing function of t for $t \le t_2$. Thus, $A(t)^2 \le A(0)^2 < RM/7$. However, (40) shows that $|T'(t)| \le R^{-1}\sqrt{A(t)} < M/7$ for $0 \le t \le t_2$, which contradicts the definition of t_2. Thus, the bound

$$|T'(t)| \le M/7 \tag{44}$$

holds for all time.

We now show that $T'(t)$ decays to zero. If in (43) we ignore the middle term, note that $R - 6\sqrt{A(t)}/M > R/7$, and integrate, we obtain

$$A(t) + 2\int_{\tau=0}^{t} \frac{R}{7}\, |T'(\tau)|^2\, d\tau \le A(0)\,.$$

The function $|T'(t)|^2$ is therefore integrable on $(0,\infty)$. However, this is insufficient for our purposes since we wish to show that $T'(t) \to 0$ as $t \to \infty$ (which implies $T \to M$).

To prove $T' \to 0$, we first show that $|(d/dt)[T'(t)]^2|$ is bounded. Equation (44) leads to

$$\left|\frac{d(T')^2}{dt}\right| \le 2\,|T'(t)|\, \left|\frac{dT'}{dt}\right| \le \frac{2M}{7}\, \left|\frac{dT'}{dt}\right|. \tag{45}$$

Using (36), (44), and the definition of $u'(x,\nu)$, we obtain

$$\left|\frac{dT'}{dt}\right| \le \frac{1}{R}\, \left[\frac{M}{7} + \left|\int \nu^{-3}\,(u - \nu^3 e^{-\nu/M})\, d\nu\right|\right].$$

If T_2 is the upper bound defined in Lemma 1, then $\max(\,M,\, T_2 - M) < T_2$ bounds the last integral. Substituting these results into (45) yields

$$|(d/dt)[T'(t)]^2| \le (2M/7)\,(1/R)\,(T_2 + M/7) \doteq c\,.$$

Thus, the function $|T'(t)|^2$ is integrable and has a uniformly bounded derivative. Clearly, $|T'(t)|^2$, for large t, cannot stray too far from zero. In fact, for any $\epsilon > 0$, $|T'(t)|^2$ can only have a finite number of spikes above ϵ no less than $2\epsilon/c$ apart, because every neighborhood with radius ϵ/c of a spike contributes at least ϵ^2/c toward the area under the curve. Thus $|T'(t)|^2 < \epsilon$ for t sufficiently large.

To show that $u' \to 0$ when ν is fixed and $t \to \infty$, we solve (35) and get

$$\begin{aligned} u'(t,\nu) &= e^{-t/\nu^3}\, u'(0,\nu) + \int_{\tau=0}^{t} e^{-(t-\tau)/\nu^3}\, \left(e^{-\nu/[M+T'(\tau)]} - e^{-\nu/M}\right) d\tau \\ &\doteq E_3 + E_4\,. \end{aligned}$$

Clearly $E_3 \to 0$ as $t \to \infty$. For E_4 we use l'Hôpital. Since $e^{t/\nu^3} \to \infty$ and $T'(t) \to 0$ as $t \to \infty$, it follows that

$$\lim_{t\to\infty} E_4(t) = \lim_{t\to\infty} \frac{\int_{\tau=0}^{t} e^{\tau/\nu^3} \left(e^{-\nu/[M+T'(\tau)]} - e^{-\nu/M} \right) d\tau}{e^{t/\nu^3}}$$

$$= \lim_{t\to\infty} \frac{e^{-\nu/[M+T'(t)]} - e^{-\nu/M}}{(1/\nu^3)} = 0\,.$$

Combining the results for E_3, E_4 completes the argument. Our proof is nonconstructive and gives no rate of convergence. However, numerical experiments indicate very slow convergence. This completes the proof. □

6. Concluding remarks. We have derived a nondimensional form of the system that couples the multifrequency radiation diffusion equations to the matter energy balance equation. In deriving the nonlinear system, we strived for equations that preserve the salient properties of the underlying physics yet are amenable to analysis. Our equations may be applied in multiple spatial dimensions and arbitrarily sized domains on which one may impose boundary conditions appropriate to parabolic systems.

Our nondimensional system conserves total energy and a unique stationary solution is eventually established. For homogeneous Neumann boundary conditions, the stationary solution is spatially constant. We have shown that the decay to equilibrium is constrained by an envelope whose extent depends on the initial conditions. Spatially varying perturbations about the stationary solution decay to zero exponentially. Spatially constant perturbations also decay. However, we are unable to estimate the speed of decay; it may be very slow.

Although our nondimensional system is a valid model for multifrequency radiation diffusion, it is instructive to recall how it differs from equations typically used in high energy density computer codes.

We assume the matter to be composed of a single material, of constant density, characterized by an ideal gas equation-of-state. For "real" materials, the internal energy is a nonlinear function of temperature [12, p. 177]. This implies a temperature dependence for the specific heat, $c_{\rm v} = c_{\rm v}(T)$. Recalling (7), we obtain the dependence $R = R(T)$ for our model.

If matter density were to vary with position, so would the opacities. This implies spatial variations for the diffusion coefficient ν^3 in (8), the coupling coefficient $1/\nu^3$ in (8)–(9), and the coefficient R in (9). For real materials, especially for those of high atomic number and/or at low temperatures, the opacity is a complicated nonmonotonic function of frequency [12, p. 246]. Most notably, it features occasional spikes corresponding to the "bound-bound" transitions of the electrons.

Our model ignores the $1/\sqrt{T}$ dependence of the opacity. As we show below, including $1/\sqrt{T}$ leads to spatial variations of the diffusion and coupling coefficients. This complicates the analysis significantly. We are then unable to invoke Duhamel's principle since a general Green's function $G_\nu(x, y, t)$ does not exist.

We chose a Wien $\nu^3 e^{-\nu/T}$ rather than a Planck $\nu^3/(e^{\nu/T} - 1)$ distribution for radiation emission. The difference between the two is most pronounced for small ν yielding a smaller radiation source. However, since our absorption opacity ignores induced emission [12], multiplying our κ_ν by a Wien distribution gives the correct total radiation source if we ignore the $1/\sqrt{T}$ dependence of the opacity. To see this,

recall that to include induced emission, one multiplies κ_ν by the factor $(1 - e^{-y})$, where $y = h\nu/kT$. Then, if instead of (6) we use the Planck function, the exponential is replaced by the factor $1/(e^y - 1)$. Hence, if B_ν denotes the Planck function and κ'_ν the absorption coefficient corrected for stimulated emission, the $(e^y - 1)$ factors cancel and we obtain the same total radiation source,

$$c \int_0^\infty d\nu\, \kappa'_\nu\, B_\nu = \text{const}\, T.$$

Choosing a Wien rather than a Planck function brings an additional benefit when the system is discretized in the frequency direction in order to derive the "multigroup" equations. For this endeavor, one defines a spectral mesh

$$0 = \nu_0 < \nu_1 < \cdots < \nu_N.$$

If B_ν is the Wien function, integrals of the form

$$\int_{\nu_{j-1}}^{\nu_j} d\nu\, \nu^n\, B_\nu$$

can be done analytically. If B_ν is the Planck function, the integrals can only be approximated.

Last, we did not include a flux limiter in the diffusion coefficient. This leads to unphysically large photon speeds for our model, especially for those with high frequencies. To incorporate a flux limiter into our model, we note that (physically) the distance x that radiation can travel in a time interval t is limited by $x \le c\,t$. Recalling the nondimensional variables x' and t', the inequality reduces to $x' \le \sqrt{3}\, t'$. Hence, if we introduce a Lund–Wilson-type flux limiter, our nondimensional flux becomes

$$-\nu^3\, \nabla u \to \frac{-1}{1/(\nu^3\, \sqrt{T}) + |\nabla u|/(\sqrt{3}\, u)}\, \nabla u \doteq -D_\nu\, \nabla u. \tag{46}$$

Note that we have kept the dependence of κ_ν on $1/\sqrt{T}$.

We conclude the discussion of how to modify our model to account for some of the above effects. If we include the $\sqrt{T}$ dependence of the opacity, allow for a nonconstant c_v, write the emission function in the general form $B_\nu = B_0\, \nu^3\, g(\nu/T)$, but do not include effects of stimulated emission, the nondimensional equations become

$$\partial_t u = \nabla \cdot D_\nu \nabla u + \left(\nu^3\, g(\nu/T) - u\right) / \sqrt{T}\, \nu^3,$$
$$R(T)\, \partial_t T = -\frac{1}{\sqrt{T}} \left(\int_0^\infty d\nu\, \frac{1}{\nu^3} (\nu^3\, g(\nu/T) - u) \right),$$

where D_ν may be defined as in (46), with or without the flux limiter and/or the $\sqrt{T}$ factor, and $g(y)$ is chosen appropriate to either the Wien or Planck distribution.

The analysis described in this paper was motivated by an unexpected result from simulations and [10] and [11], which solve the multigroup system derived from (8)–(9). The papers simulate the relaxation to steady state of an initial condition in which the two fields, u and T, are wildly out of equilibrium.

In the problem of interest, the spatial domain is $|x| < 1$. Initially, T is sharply peaked:

$$T|_{t=0} = \begin{cases} 6.4775 & \text{if } |x| < 0.04, \\ 0.0027 & \text{otherwise.} \end{cases}$$

Using a "specific heat" $R = 2.0$ yields an initial matter energy $\mathcal{E}_m(t)|_{t=0} = 1.0469$. The spectral radiation energy density u is initialized with a Wien profile characterized with a spatially constant "radiation temperature" $T_r = 0.0172$. Hence, the initial radiation energy $\mathcal{E}_r(0) = 1.0469 \cdot 10^{-6}$. The conditions imply that initially 99% of the total energy $\mathcal{E}_m + \mathcal{E}_r$ is contained in the central region $|x| < 0.04$. The problem is designed so that the stationary solution consists of spatially uniform fields with the matter energy equal to 1, i.e., with $T = 0.25$.

In the simulation, the hot central region pumps a prodigious amount of energy into the high frequencies. (This is a result consistent with the Wien profile which peaks at $\nu/T = 3$.) The high frequencies are characterized by fast transport. Thus, this energy quickly diffuses away from the hot spot. However, the high frequency energy is slow to absorb since the coupling coefficient is $1/\nu^3$. In effect, the high frequency energy is trapped for a time proportional to ν^3. Because of the trapping, the ratio $\mathcal{E}_r(t)/\mathcal{E}_m(t)$ exhibits a surprising behavior. It quickly rises to approximately 0.6 (more than 10 times the equilibrium value), then gradually decays, albeit very slowly. That very slow relaxation to equilibrium is the subject of section 5. During the slow-relaxation phase, the fields have little spatial variation.

Acknowledgments. Aleksei I. Shestakov is extremely grateful to Prof. Alexandre Chorin of the Mathematics Department of the University of California at Berkeley and the Lawrence Berkeley Laboratory for numerous discussions, suggestions, and the opportunity to pursue this work in the intellectually stimulating environment of the Mathematics Department of the Lawrence Berkeley Laboratory. We also thank the referee(s) for numerous suggestions. When we started the project, we were unaware of the analytical work of Andreev, Kozmanov, and Rachilov [1], Mercier [8], Larsen and Mercier [6], and Densmore and Larsen [3]. We hope that our contribution will complement those papers.

REFERENCES

[1] E. S. Andreev, M. Yu. Kozmanov, and E. B. Rachilov, *The maximum principle for a system of equations of energy and non-stationary radiation transfer*, Comput. Math. Math. Phys., 23 (1983), pp. 104–109.

[2] L. L. Carter and C. A. Forest, *Nonlinear Radiation Transport Simulation with an Implicit Monte Carlo Method*, LA-5038, Los Alamos National Laboratory, Los Alamos, NM, 1973.

[3] J. D. Densmore and E. W. Larsen, *Maximum principle analysis of Monte Carlo methods for grey radiative transfer*, in Proceedings of the ANS Topical Meeting: Nuclear, Mathematical, and Computational Sciences: A Century in Review—A Century Anew, Gatlinburg, TN, 2003, CD-ROM, American Nuclear Society, La Grange Park, IL, 2003.

[4] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, RI, 1998.

[5] J. A. Fleck, Jr., and J. D. Cummings, *An implicit Monte Carlo scheme for calculating time and frequency dependent nonlinear radiation transport*, J. Comput. Phys., 8 (1971), pp. 313–342.

[6] E. W. Larsen and B. Mercier, *Analysis of a Monte Carlo method for nonlinear radiative transfer*, J. Comput. Phys., 71 (1987), pp. 50–64.

[7] C. M. Lund and J. R. Wilson, *Some Numerical Methods for Time-Dependent Multifrequency Radiation Transport Calculations in One Dimension*, UCRL-84678, Lawrence Livermore National Laboratory, Livermore, CA, 1980.

[8] B. Mercier, *Application of accretive operators theory to the radiative transfer equations*, SIAM J. Math. Anal., 18 (1987), pp. 393–408.

[9] D. Mihalas and B. Weibel-Mihalas, *Foundations of Radiation Hydrodynamics,* Dover, Mineola, New York, 1999.

[10] A. I. Shestakov, *Solving the Multifrequency Radiation Diffusion Equations Using Branching Brownian Motion*, UCRL-JC-149411-Rev.1, Lawrence Livermore National Laboratory, Livermore, CA, 2003, J. Comput. Phys., to appear.

[11] A. I. Shestakov, *Solution of the Nonlinear Multigroup Radiation Diffusion Equation Using Multigrid and Pseudo Transient Continuation*, UCRL-JC-148873,Lawrence Livermore National Laboratory, Livermore, CA, 2002.
[12] Ya. B. Zel'dovich and Yu. P. Raizer, *Physics of Shock Waves and High-Temperature Hydrodynamic Phenomena, Vol.* I, Dover, Mineola, NY, 2001.

SIAM J. APPL. MATH.
Vol. 65, No. 1, pp. 194–208

ON THE FRÉCHET DERIVATIVE FOR OBSTACLE SCATTERING WITH AN IMPEDANCE BOUNDARY CONDITION*

HOUSSEM HADDAR† AND RAINER KRESS‡

Abstract. A technique introduced by Kress and Päivärinta to establish differentiability of the solution to obstacle scattering problems with respect to the boundary is extended to the case of the impedance boundary condition. For acoustic scattering an alternative proof of a differentiability result due to Hettlich is provided, and in electromagnetic scattering a new differentiability result is proven.

Key words. Helmholtz equation, Maxwell equations, acoustic scattering, electromagnetic scattering, obstacle scattering, impedance condition, far field pattern, inverse scattering, Fréchet derivative

AMS subject classifications. 35J05, 35Q60, 35R30, 45A05, 45P05

DOI. 10.1137/S0036139903435413

1. Introduction.

In time-harmonic inverse obstacle scattering, for the mathematical foundation and implementation of approximate solution methods by regularized iteration schemes via linearization, it is necessary to investigate the differentiability of the boundary to far field operator that maps the boundary of the obstacle onto the far field pattern of the scattered wave. In acoustic scattering, differentiability with respect to the boundary was considered by Roger [18], who first employed Newton-type iterations for inverse obstacle scattering problems. Rigorous foundations for the Fréchet differentiability including characterizations of the derivative both for the Dirichlet and Neumann boundary condition, i.e., for sound-soft and sound-hard obstacles, were given by Kirsch [7] and Hettlich [3] in the sense of a domain derivative via variational methods and by Potthast [15, 16] via boundary integral equation techniques. Alternative proofs were contributed by Hohage [6] and Schormann [19] via the implicit function theorem and by Kress and Päivärinta [11] via Green's theorems and a factorization of the difference of the far field for neighboring domains. Hettlich [3] also established differentiability for the impedance boundary condition.

In electromagnetic obstacle scattering from perfect conductors Fréchet differentiability was considered by Potthast [17] via boundary integral equations. Hettlich [5] treated the transmission problem for penetrable obstacles via variational methods. The technique due to Kress and Päivärinta was extended to the Maxwell equations for the perfect conductor case in [9]. The impedance boundary condition for electromagnetic obstacle scattering has not yet been considered in the literature.

It is the purpose of the present paper to extend the method introduced by Kress and Päivärinta to the case of the impedance boundary condition. In section 2 we will reestablish the results of Hettlich through an alternative proof, and in section 3 we will establish Fréchet differentiability with respect to the boundary for the electromagnetic impedance problem and provide a characterization of the derivative. In principle, one of the key ideas of the method is to extend the continuous dependence of the

*Received by the editors October 1, 2003; accepted for publication (in revised form) January 20, 2004; published electronically September 24, 2004.
http://www.siam.org/journals/siap/65-1/43541.html
†Inria-Rocquencourt, 78153, Le Chesnay Cedex, France (Houssem.Haddar@inria.fr).
‡Institut für Numerische und Angewandte Mathematik, Universität Göttingen, 37083 Göttingen, Germany (kress@math.uni-goettingen.de).

solution on the boundary that can be obtained via boundary integral equations to differentiability via the application of Green's integral and Green's representation theorems.

The paper will be concluded with a few remarks on the difficulties that arise in connection with investigating the nullspace of the linearization, i.e., the Fréchet derivative.

2. Impedance problem in acoustic scattering. Let $D \subset \mathbb{R}^3$ be a bounded domain with a connected boundary ∂D of class C^2 and outward unit normal ν. Consider the exterior impedance boundary value problem for acoustic waves: Given a continuous function f on ∂D and a constant $\lambda \in \mathbb{C}$ find a solution $v \in C^2(\mathbb{R}^3 \setminus \bar{D}) \cap C^1(\mathbb{R}^3 \setminus D)$ to the Helmholtz equation

$$\triangle v + k^2 v = 0 \quad \text{in } \mathbb{R}^3 \setminus \bar{D} \tag{2.1}$$

with wave number $k > 0$ that satisfies the impedance boundary condition

$$\frac{\partial v}{\partial \nu} + ik\lambda v = f \quad \text{on } \partial D \tag{2.2}$$

and the Sommerfeld radiation condition

$$\lim_{r\to\infty} r\left(\frac{\partial v}{\partial r} - ikv\right) = 0, \quad r = |x|, \tag{2.3}$$

uniformly for all directions. The impedance coefficient λ is assumed to satisfy the condition

$$\operatorname{Re}\lambda \geq 0 \tag{2.4}$$

that ensures uniqueness of a solution via Rellich's lemma. For existence of a solution we refer to [1, 2].

For an impedance obstacle D the scattering problem for time-harmonic waves is, given an incident field u^i as an entire solution of the Helmholtz equation, to find the total field $u = u^i + u^s$ as a solution to the Helmholtz equation in the exterior $\mathbb{R}^3 \setminus \bar{D}$ of D, such that u satisfies the impedance boundary condition

$$\frac{\partial u}{\partial \nu} + ik\lambda u = 0 \quad \text{on } \partial D$$

and u^s fulfills the Sommerfeld radiation condition. Clearly, this scattering problem is a special case of the above boundary value problem (2.1)–(2.3).

We introduce the fundamental solution to the Helmholtz equation in $\mathbb{R}^3$ by

$$\Phi(x,y) := \frac{1}{4\pi}\,\frac{e^{ik|x-y|}}{|x-y|}\,, \quad x \neq y.$$

For $x \in \mathbb{R}^3 \setminus \bar{D}$ let $w^s(x,\cdot)$ be the solution of (2.1)–(2.3) for the boundary values

$$f = -\frac{\partial \Phi(x,\cdot)}{\partial \nu} - ik\lambda\Phi(x,\cdot) \quad \text{on } \partial D$$

and let $w(x,\cdot) := \Phi(x,\cdot) + w^s(x,\cdot)$; that is, $w^s(x,\cdot)$ and $w(x,\cdot)$ are the scattered and the total field, respectively, for the scattering of a point source located at $x \in \mathbb{R}^3 \setminus \bar{D}$.

Note that the function w is the Green function for the boundary value problem (2.1)–(2.3) and that it enjoys the symmetry, i.e., the reciprocity

$$w(x,y) = w(y,x), \quad x, y \in \mathbb{R}^3 \setminus \bar{D},\ x \neq y,$$

which can be easily derived from Green's integral theorem. From Green's representation theorem and Green's integral formula it follows that (see also the proof of the following lemma) the unique solution of the impedance boundary value problem (2.1)–(2.3) can be represented in the form

$$v(x) = -\int_{\partial D} w(x,y) f(y)\, ds(y), \quad x \in \partial D. \tag{2.5}$$

We now consider a family of scatterers D_h with boundaries represented in the form

$$\partial D_h = \{x + h(x) : x \in \partial D\}, \tag{2.6}$$

where $h : \partial D \to \mathbb{R}^3$ is of class C^2. Provided h is sufficiently small in the C^2 norm on ∂D, then ∂D_h is well defined and the boundary of a C^2 domain D_h. By ν_h we denote its exterior unit normal and, in what follows, we will distinguish the above quantities related to the impedance scattering problem for the domain D_h through the subscript h.

LEMMA 2.1. *Assume that $\bar{D} \subset D_h$. Then*

$$u_h^s(x) - u^s(x) = \int_{\partial D_h} u_h \left\{ \frac{\partial w(x,\cdot)}{\partial \nu_h} + ik\lambda w(x,\cdot) \right\} ds \tag{2.7}$$

for $x \in \mathbb{R}^3 \setminus \bar{D}_h$.

Proof. Let $x \in \mathbb{R}^3 \setminus \bar{D}_h$. Combining Green's integral theorem for the incident field u^i and Green's representation theorem for the scattered field u^s we have that

$$u^s(x) = \int_{\partial D} \left\{ u\, \frac{\partial \Phi(x,\cdot)}{\partial \nu} - \Phi(x,\cdot)\, \frac{\partial u}{\partial \nu} \right\} ds. \tag{2.8}$$

From the impedance boundary condition for u and $w(x,\cdot)$ by straightforward calculations we obtain

$$u\, \frac{\partial \Phi(x,\cdot)}{\partial \nu} - \Phi(x,\cdot)\, \frac{\partial u}{\partial \nu} = -u\, \frac{\partial w^s(x,\cdot)}{\partial \nu} + w^s(x,\cdot)\, \frac{\partial u}{\partial \nu} \quad \text{on } \partial D.$$

Hence, using Green's integral theorem and the radiation condition, (2.8) implies that

$$u^s(x) = -\int_{\partial D} \left\{ u^i\, \frac{\partial w^s(x,\cdot)}{\partial \nu} - w^s(x,\cdot)\, \frac{\partial u^i}{\partial \nu} \right\} ds \tag{2.9}$$

and then again via Green's theorem

$$u^s(x) = -\int_{\partial D_h} \left\{ u^i\, \frac{\partial w^s(x,\cdot)}{\partial \nu_h} - w^s(x,\cdot)\, \frac{\partial u^i}{\partial \nu_h} \right\} ds. \tag{2.10}$$

From this, once again using Green's integral theorem and the radiation condition, we find that

$$u^s(x) = -\int_{\partial D_h} \left\{ u_h\, \frac{\partial w^s(x,\cdot)}{\partial \nu_h} - w^s(x,\cdot)\, \frac{\partial u_h}{\partial \nu_h} \right\} ds$$

and in view of the boundary condition for u_h we finally arrive at

$$u^s(x) = -\int_{\partial D_h} u_h \left\{ \frac{\partial w^s(x,\cdot)}{\partial \nu_h} + ik\lambda w^s(x,\cdot) \right\} ds.$$

On the other hand, from (2.8) applied to u_h and ∂D_h and the boundary condition for u_h we have that

$$u_h^s(x) = \int_{\partial D_h} u_h \left\{ \frac{\partial \Phi(x,\cdot)}{\partial \nu_h} + ik\lambda \Phi(x,\cdot) \right\} ds \tag{2.11}$$

and the statement (2.7) follows by combining the last two equations. □

Remark 2.2. We note that (2.7) remains valid for $x \notin \bar{D} \cup \bar{D}_h$ if we drop the assumption that $\bar{D} \subset D_h$ and assume that w^s can be extended as a solution to the Helmholtz equation in the exterior of D_h. By Theorem 5.7.1′ in [13, p. 169], this can be assured if ∂D is analytic and D_h does not differ too much from D. In this case, obviously, (2.10) follows from (2.9) via Green's theorem by introducing the corresponding integral over a sufficiently large sphere as an intermediate step. □

In a neighborhood of ∂D we introduce a coordinate system via

$$z = x + t\nu(x), \quad x \in \partial D,\ t \in [-T, T], \tag{2.12}$$

with some sufficiently small $T > 0$. Straightforward calculations show that the determinant G of the metric in the neighborhood

$$U_T := \{z = x + t\nu(x) : x \in \partial D,\ t \in [-T, T]\}$$

and the determinant g of the metric on ∂D are related via

$$G(z) = g(x)[1 - 2tH(x) + t^2 K(x)]^2, \tag{2.13}$$

where H denotes the mean curvature and K the Gaussian curvature of the surface ∂D (with respect to the exterior normal direction). In the neighborhood U_T we define an extension ν of the exterior unit normal ν to ∂D by setting

$$\nu(x + t\nu(x)) := \nu(x), \quad x \in \partial D,\ t \in [-T, T].$$

If we chose an orthonormal coordinate system at the point $x \in \partial D$, then we have that

$$\operatorname{div} \nu(x) = \frac{1}{\sqrt{G(z)}} \left. \frac{\partial \sqrt{G(z)}}{\partial t} \right|_{t=0}$$

and therefore (2.13) implies that

$$\operatorname{div} \nu = -2H \quad \text{on } \partial D. \tag{2.14}$$

We would like to point out that, in the literature, there is some ambiguity for the sign in the definition of the mean curvature. In this paper, in order to be consistent with [1, 3, 5] we choose the sign according to the classical definition in differential geometry. However, other authors prefer a definition where the sign in (2.14) is reversed (see, for example, Nédélec [14]).

We assume that h is small enough to ensure that ∂D_h is contained in U_T. For the further analysis we need to relate the normal ν_h to ∂D_h and the extended normal ν. To achieve this, we consider a fixed point $x_0 \in \partial D$ and a local parameterization

$$\partial D \cap V = \{\varphi(\xi) : \xi \in U\}$$

for a neighborhood V of x_0 and an open set $U \subset \mathbb{R}^2$. For convenience, we may assume that $\varphi(0) = x_0$ and that the tangential vectors satisfy

$$\left|\frac{\partial \varphi}{\partial \xi_1}(0)\right| = \left|\frac{\partial \varphi}{\partial \xi_2}(0)\right| = 1, \quad \frac{\partial \varphi}{\partial \xi_1}(0) \cdot \frac{\partial \varphi}{\partial \xi_2}(0) = 0,$$

and

$$\nu(x_0) = \frac{\partial \varphi}{\partial \xi_1}(0) \times \frac{\partial \varphi}{\partial \xi_2}(0).$$

The perturbed boundary ∂D_h is locally described by

$$\psi(\xi) = \varphi(\xi) + \tilde{h}(\xi), \quad \xi \in U,$$

where $\tilde{h} = h \circ \varphi$. We compute

$$\frac{\partial \psi}{\partial \xi_1}(0) \times \frac{\partial \psi}{\partial \xi_2}(0) = \nu(x_0) + \delta + O\left(\|h\|^2_{C^1(\partial D)}\right),$$

where we have set

$$\delta := \frac{\partial \tilde{h}}{\partial \xi_1}(0) \times \frac{\partial \varphi}{\partial \xi_2}(0) + \frac{\partial \varphi}{\partial \xi_1}(0) \times \frac{\partial \tilde{h}}{\partial \xi_2}(0).$$

From this it follows that

$$\nu_h(x_0 + h(x_0)) = \nu(x_0) + \delta - [\nu(x_0) \cdot \delta]\, \nu(x_0) + O\left(\|h\|^2_{C^1(\partial D)}\right).$$

Straightforward computations exploiting the orthonormality of the tangent vectors at x_0 yield that

$$\begin{aligned}
\delta - [\nu(x_0) \cdot \delta]\, \nu(x_0) &= \nu(x_0) \times [\delta \times \nu(x_0)] \\
&= -\sum_{j=1}^{2} \left[\nu(x_0) \cdot \frac{\partial \tilde{h}}{\partial \xi_j}(0)\right] \frac{\partial \varphi}{\partial \xi_j}(0) \\
&= \sum_{j=1}^{2} \left[h(x_0) \cdot \frac{\partial \nu}{\partial \xi_j}(0)\right] \frac{\partial \varphi}{\partial \xi_j}(0) - \mathrm{Grad}[\nu \cdot h](x_0),
\end{aligned}$$

where Grad denotes the surface gradient on ∂D. Combining this with the previous equation we find that

$$\begin{aligned}
\nu_h(x_0 + h(x_0)) - \nu(x_0) &= \sum_{j=1}^{2} \left[h(x_0) \cdot \frac{\partial \nu}{\partial \xi_j}(0)\right] \frac{\partial \varphi}{\partial \xi_j}(0) \\
&\quad - \mathrm{Grad}[\nu \cdot h](x_0) + O\left(\|h\|^2_{C^1(\partial D)}\right).
\end{aligned} \tag{2.15}$$

From Taylor's formula and $\nabla\nu \cdot \nu = 0$, using the coordinate system (2.12) we obtain

$$\begin{aligned}\nu(x_0 + h(x_0)) - \nu(x_0) &= [\nabla\nu](x_0)h(x_0) + O\left(\|h\|^2_{C^1(\partial D)}\right) \\ &= \sum_{j=1}^{2}\left[h(x_0)\cdot\frac{\partial\varphi}{\partial\xi_j}(0)\right]\frac{\partial\nu}{\partial\xi_j}(0) + O\left(\|h\|^2_{C^1(\partial D)}\right).\end{aligned}$$

Subtracting this from (2.15) yields

$$\begin{aligned}\nu_h(x_0 + h(x_0)) - \nu(x_0 + h(x_0)) &= \sum_{j=1}^{2}\left[\frac{\partial\nu}{\partial\xi_j}(0)\times\frac{\partial\varphi}{\partial\xi_j}(0)\right]\times h(x_0) \\ &\quad - \operatorname{Grad}[\nu\cdot h](x_0) + O\left(\|h\|^2_{C^1(\partial D)}\right).\end{aligned}$$

From this, in view of

$$\nu\cdot\frac{\partial\varphi}{\partial\xi_j} = 0, \quad j = 1, 2,$$

and

$$\frac{\partial\nu}{\partial\xi_j}\cdot\frac{\partial\varphi}{\partial\xi_\ell} = -\nu\cdot\frac{\partial^2\varphi}{\partial\xi_j\,\partial\xi_\ell} = \frac{\partial\nu}{\partial\xi_\ell}\cdot\frac{\partial\varphi}{\partial\xi_j}, \quad j, \ell = 1, 2,$$

we finally obtain the following technical lemma.

LEMMA 2.3. *The normal ν_h to ∂D_h and the extended normal ν are related by the estimate*

$$\nu_h(x + h(x)) - \nu(x + h(x)) = -\operatorname{Grad}[\nu\cdot h](x) + O\left(\|h\|^2_{C^1(\partial D)}\right) \tag{2.16}$$

uniformly for all $x \in \partial D$.

In particular, from (2.16) we can conclude that

$$\nu\cdot(\nu_h - \nu) = O\left(\|h\|^2_{C^1(\partial D)}\right). \tag{2.17}$$

LEMMA 2.4. *Let ∂D be analytic and K be a compact subset of $\mathbb{R}^3 \setminus \bar{D}$. Then*

$$u^s_h(x) - u^s(x) = -\int_{\partial D} w(x, y)\,(Bu)(y)\,ds(y) + o\left(\|h\|_{C^2(\partial D)}\right) \tag{2.18}$$

uniformly for x in K, where

$$Bu := [k^2(1 - \lambda^2) + 2ik\lambda H u](\nu\cdot h) + \operatorname{Div}[(\nu\cdot h)\operatorname{Grad} u]$$

and Div *and* Grad *denote the surface divergence and surface gradient, respectively.*

Proof. From (2.11), with the aid of the jump relations we obtain the boundary integral equation

$$u_h(x) = 2u^i(x) + 2\int_{\partial D_h}\left\{\frac{\partial\Phi(x,\cdot)}{\partial\nu_h} + ik\lambda\Phi(x,\cdot)\right\}u_h\,ds, \quad x \in \partial D_h. \tag{2.19}$$

Employing a perturbation argument based on pointwise convergence and collective compactness of the single- and double-layer boundary integral operators as $h \to 0$, from (2.19) it can be deduced that

$$|u_h(y + h(y)) - u(y)| \to 0, \quad \|h\|_{C^2(\partial D)} \to 0,$$

uniformly for all $y \in \partial D$. From this, in view of (2.7), using the impedance boundary condition for w together with Taylor's formula and the continuity of the second derivatives of w up to the boundary ∂D it follows that

$$u_h^s(x) - u^s(x) = \int_{\partial D_h} u \left\{ \frac{\partial w(x,\cdot)}{\partial \nu_h} + ik\lambda w(x,\cdot) \right\} ds + o\left(\|h\|_{C^2(\partial D)}\right) \tag{2.20}$$

uniformly for x on compact subsets of $\mathbb{R}^3 \setminus \bar{D}$. Abbreviating $w = w(x,\cdot)$, from (2.20) and the divergence theorem in view of (2.17) and the boundary condition for w we obtain

$$u_h^s(x) - u^s(x) = \int_{D_h^*} \operatorname{div} \{u \operatorname{grad} w + ik\lambda uw\nu\} \chi \, dy + o\left(\|h\|_{C^2(\partial D)}\right). \tag{2.21}$$

Here

$$D_h^* := \{y \in D_h : y \notin D\} \cup \{y \in D : y \notin D_h\}$$

and $\chi(y) = 1$ if $y \in D_h$ and $y \notin D$ and $\chi(y) = -1$ if $y \in D$ and $y \notin D_h$. Approximating the integral over D_h^* in (2.21) by an integral over ∂D, with the aid of (2.13) and Taylor's formula, we readily obtain that

$$u_h^s(x) - u^s(x) = \int_{\partial D} \operatorname{div} \{u \operatorname{grad} w + ik\lambda uw\nu\} (\nu \cdot h) \, ds + o\left(\|h\|_{C^2(\partial D)}\right).$$

Straightforward computations using the Helmholtz equation for w, the impedance boundary conditions for u and w, and (2.14) yield that

$$\operatorname{div} \{u \operatorname{grad} w + ik\lambda uw\nu\} = \operatorname{Grad} u \cdot \operatorname{Grad} w - k^2(1-\lambda^2)uw - 2ik\lambda H uw.$$

Inserting this into the previous equation and applying the Gauss surface divergence theorem we finally arrive at (2.18). □

The Sommerfeld radiation condition (2.3) implies an asymptotic behavior in the form of a spherical wave

$$v(x) = \frac{e^{ik|x|}}{|x|} \left\{ v_\infty(\hat{x}) + O\left(\frac{1}{|x|}\right) \right\}, \quad |x| \to \infty,$$

uniformly for all directions $\hat{x} = x/|x|$ where the function v_∞, defined on the unit sphere $\Omega := \{z \in \mathbb{R}^3 : |z| = 1\}$, is known as the far field pattern.

For a fixed incident field u^i, we define the operator

$$F : \partial D_h \to u_{\infty,h}$$

that maps the boundary ∂D_h onto the far field pattern $u_{\infty,h}$ of the scattered wave for scattering from the impedance obstacle D_h. For this operator we now are ready to prove the following result on Fréchet differentiability.

THEOREM 2.5. *Let ∂D be analytic. Then the boundary to far field operator $F : C(\partial D) \to L^2(\Omega)$ is Fréchet differentiable with the Fréchet derivative given through*

$$F'(\partial D)h = v_{\infty,h},$$

where $v_{\infty,h}$ is the far field pattern of the solution v_h to the Helmholtz equation in $\mathbb{R}^3 \setminus \bar{D}$ that satisfies the Sommerfeld radiation condition and the impedance boundary condition

$$\frac{\partial v_h}{\partial \nu} + ik\lambda v_h = [k^2(1-\lambda^2) + 2ik\lambda Hu](\nu \cdot h) + \operatorname{Div}[(\nu \cdot h)\operatorname{Grad} u] \quad \text{on } \partial D.$$

Proof. This is an immediate consequence of Lemma 2.4 and the representation formula (2.5). □

We expect that proceeding as in [11] Theorem 2.5 can be extended to the case of C^2 boundaries.

3. Impedance problem in electromagnetic scattering. Now we consider the exterior impedance boundary value problem for electromagnetic waves: Given a Hölder continuous tangential field c on ∂D and a constant $\lambda \in \mathbb{C}$ find a solution $E, H \in C^1(\mathbb{R}^3 \setminus \bar{D}) \cap C(\mathbb{R}^3 \setminus D)$ to the time-harmonic Maxwell equations

$$\operatorname{curl} E - ikH = 0, \quad \operatorname{curl} H + ikE = 0 \quad \text{in } \mathbb{R}^3 \setminus \bar{D} \tag{3.1}$$

that satisfies the impedance or Leontovich boundary condition

$$\nu \times H - \lambda(\nu \times E) \times \nu = c \quad \text{on } \partial D \tag{3.2}$$

and the Silver–Müller radiation condition

$$\lim_{r\to\infty} (H \times x - rE) = 0, \tag{3.3}$$

where $r = |x|$ and the limit holds uniformly for all directions $x/|x|$. The impedance coefficient λ is assumed to satisfy the condition

$$\operatorname{Re} \lambda \geq 0 \tag{3.4}$$

that ensures uniqueness and existence of a solution to (3.1)–(3.3) (see [1]). The limiting case $\lambda = 0$ corresponds to the case of a perfectly conducting obstacle with the roles of the fields E and H interchanged. We note that in this case the existence of a classical solution requires the existence and Hölder continuity of the surface divergence $\operatorname{Div} c$ of the given boundary data.

For an impedance obstacle D the scattering problem for time-harmonic waves is, given an incident field E^i, H^i as an entire solution of the Maxwell equations, to find the total field $E = E^i + E^s, H = H^i + H^s$ as a solution to the Maxwell equations in the exterior $\mathbb{R}^3 \setminus \bar{D}$ of D, such that E, H satisfies the impedance boundary condition

$$\nu \times H - \lambda(\nu \times E) \times \nu = 0 \quad \text{on } \partial D$$

and E^s, H^s satisfies the Silver–Müller radiation condition. Clearly, this scattering problem is a special case of the above boundary value problem (3.1)–(3.3).

On occasion, as an abbreviation we will use

$$E_T := (\nu \times E) \times \nu$$

for the tangential component of E on the boundary ∂D. For a matrix $M=(m_1,m_2,m_3)$ in $\mathbb{C}^3\times\mathbb{C}^3$ with columns m_1, m_2, m_3 we define matrices $\nu\times M$ and $M\times\nu$ by $\nu\times M := (\nu\times m_1, \nu\times m_2, \nu\times m_3)$ and $M\times\nu := -\nu\times M$.

Describing the electromagnetic field of an electric dipole with polarization $p\in\mathbb{R}^3$ located at a point $y\in\mathbb{R}^3$ we introduce matrices E_e^i, H_e^i through

$$E_e^i(x,y)p := \frac{i}{k}\,\mathrm{curl}_x\,\mathrm{curl}_x\,p\Phi(x,y), \quad H_e^i(x,y)p := \mathrm{curl}_x\,p\Phi(x,y)$$

in terms of the fundamental solution Φ to the Helmholtz equation. Note the symmetries

$$E_e^i(x,y) = [E_e^i(x,y)]^\top = E_e^i(y,x) \quad \text{and} \quad H_e^i(x,y) = [H_e^i(y,x)]^\top, \quad x\neq y.$$

For $z\in\mathbb{R}^3\setminus\bar{D}$ denote by $E_e^s(\cdot,z), H_e^s(\cdot,z)$ the matrices for which the pairs of corresponding columns are the solution of (3.1)–(3.3) for the boundary values given by the columns of

$$c = -\nu\times H_e^i(\cdot,z) + \lambda(\nu\times E_e^i(\cdot,z))\times\nu \quad \text{on } \partial D.$$

Then define $E_e := E_e^i + E_e^s, H_e := H_e^i + H_e^s$; that is, E_e^s, H_e^s and E_e, H_e describe the scattered and the total field, respectively, for the scattering of an electric dipole located at $x\in\mathbb{R}^3\setminus\bar{D}$. In view of the symmetry

$$E_e(x,y) = [E_e(y,x)]^\top, \quad x,y\in\mathbb{R}^3\setminus\bar{D},\ x\neq y,$$

that can be shown with the aid of the Gauss divergence theorem (see also section 7 in [8]), from the Stratton–Chu representation theorem it follows that (see also the proof of the following lemma) the unique solution of the impedance boundary value problem (3.1)–(3.3) can be represented in the form

$$E(x) = -\int_{\partial D} E_e(x,y)c(y)\,ds(y), \quad x\in\mathbb{R}^3\setminus\bar{D}. \tag{3.5}$$

As in the previous section we consider a family of domains D_h as defined in (2.6) and distinguish the above quantities related to the impedance scattering problem for the domain D_h through the subscript h.

LEMMA 3.1. *Assume that $\bar{D}\subset D_h$. Then*

$$E_h^s(x) - E^s(x) = -\int_{\partial D_h} \left\{[H_e(\cdot,x)]^\top + \lambda\,[\nu_h\times E_e(\cdot,x)]^\top\right\}[\nu_h\times E_h]\,ds \tag{3.6}$$

for $x\in\mathbb{R}^3\setminus\bar{D}_h$.

Proof. Let $x\in\mathbb{R}^3\setminus\bar{D}_h$. Combining the Stratton–Chu representation theorem for the incident field E^i, H^i and for the scattered field E^s, H^s we have that

$$E^s(x) = -\int_{\partial D} \left\{[E_e^i(\cdot,x)]^\top[\nu\times H] + [H_e^i(\cdot,x)]^\top[\nu\times E]\right\} ds \tag{3.7}$$

(see Corollary 2.2 in [8]). From the impedance boundary condition for E, H, and $E_e(\cdot,x), H_e(\cdot,x)$ by straightforward calculations we obtain

$$[E_e^i(\cdot,x)]^\top[\nu\times H] + H_e^i(\cdot,x)]^\top[\nu\times E] = -[E_e^s(\cdot,x)]^\top[\nu\times H] - [H_e^s(\cdot,x)]^\top[\nu\times E]$$

on ∂D. Hence, using the Gauss divergence theorem (see Lemma 2.1 in [8]) and the Silver–Müller radiation condition, (3.7) implies that

$$E^s(x) = \int_{\partial D} \left\{ [E_e^s(\cdot,x)]^\top [\nu \times H^i] + [H_e^s(\cdot,x)]^\top [\nu \times E^i] \right\} ds \tag{3.8}$$

and then again via the Gauss divergence theorem

$$E^s(x) = \int_{\partial D_h} \left\{ [E_e^s(\cdot,x)]^\top [\nu_h \times H^i] + [H_e^s(\cdot,x)]^\top [\nu_h \times E^i] \right\} ds. \tag{3.9}$$

From this, applying once again the Gauss divergence theorem and the radiation condition, we obtain that

$$E^s(x) = \int_{\partial D_h} \left\{ [E_e^s(\cdot,x)]^\top [\nu_h \times H_h] + [H_e^s(\cdot,x)]^\top [\nu_h \times E_h] \right\} ds,$$

and in view of the boundary condition for E_h, H_h we finally arrive at

$$E^s(x) = \int_{\partial D_h} \left\{ [H_e^s(\cdot,x)]^\top + \lambda\, [\nu_h \times E_e^s(\cdot,x)]^\top \right\} [\nu_h \times E_h]\, ds.$$

On the other hand, from (3.7) applied to $E_h, H_h, \partial D_h$ and the boundary condition for E_h, H_h we have that

$$E_h^s(x) = -\int_{\partial D_h} \left\{ [H_e^i(\cdot,x)]^\top + \lambda\, [\nu_h \times E_e^i(\cdot,x)]^\top \right\} [\nu_h \times E_h]\, ds \tag{3.10}$$

and the statement (3.6) follows by combining the last two equations. □

Remark 3.2. Again (3.6) remains valid for $x \notin \bar{D} \cup \bar{D}_h$ if we drop the assumption that $\bar{D} \subset D_h$ and assume that E^s, H^s can be extended as a solution to the Maxwell equations in the exterior of D_h. By the regularity results on elliptic boundary value problems this can be assured if ∂D is analytic and D_h does not differ too much from D. This follows from sections 6.1 and 6.6 in [13] by considering the impedance boundary value problem for the Maxwell equation equivalently as a boundary value problem for the vector Helmholtz equation.

In the neighborhood of ∂D we define the curvature operator R by the matrix with the columns

$$R := \left(\frac{\partial \nu}{\partial x_1}\,,\ \frac{\partial \nu}{\partial x_2}\,,\ \frac{\partial \nu}{\partial x_3} \right).$$

Then the curl operator trace on ∂D can be expressed in terms of surface operators in the form

$$\operatorname{curl} E = [\operatorname{Div}(E \times \nu)]\nu + \operatorname{Grad}(E \cdot \nu) \times \nu + \left(R + 2H_{\partial D} - \frac{\partial}{\partial \nu} \right) (E \times \nu) \quad \text{on } \partial D,$$

where we use the subscript ∂D to distinguish the mean curvature $H_{\partial D}$ from the magnetic field (see Theorem 2.5.20 in [14]). From this we derive

$$\operatorname{curl}(\nu \times E) = [\operatorname{Div} E_T]\nu + \left(R + 2H_{\partial D} - \frac{\partial}{\partial \nu} \right) E_T \quad \text{on } \partial D, \tag{3.11}$$

and, in particular, since E_T is tangential

$$\nu \cdot \operatorname{curl} E = \operatorname{Div}(E \times \nu) \quad \text{on } \partial D \tag{3.12}$$

and

$$\nu \cdot \operatorname{curl}(\nu \times E) = \operatorname{Div} E_T \quad \text{on } \partial D. \tag{3.13}$$

LEMMA 3.3. *Let ∂D be analytic and let K be a compact subset of $\mathbb{R}^3 \setminus \bar{D}$. Then*

$$E_h^s(x) - E^s(x) = -\int_{\partial D} [E_e(y,x)]^\top (NE)(y)\, ds(y) + o\left(\|h\|_{C^2(\partial D)}\right) \tag{3.14}$$

uniformly for x in K, where

$$\begin{aligned} NE :=\ & -\nu \times \{\operatorname{Grad}[(\nu \cdot H)(\nu \cdot h)] + \lambda\, E \times \operatorname{Grad}(\nu \cdot h)\} \\ & +(\nu \cdot h)\left[ik - \lambda\left(R + 2H_{\partial D} - \frac{\partial}{\partial \nu}\right)\right] E_T. \end{aligned}$$

Proof. From (3.10), with the aid of the jump relation we obtain the hypersingular boundary integral equation

$$\begin{aligned} &\nu_h(x) \times E_h(x) = 2\,\nu_h(x) \times E^i(x) \\ &-2\,\nu_h(x) \times \int_{\partial D_h} \left\{[H_e^i(\cdot,x)]^\top + \lambda\,[\nu_h \times E_e^i(\cdot,x)]^\top\right\}[\nu_h \times E_h]\, ds, \quad x \in \partial D_h. \end{aligned}$$

From this, regularizing the integral equation and employing a perturbation argument it can be deduced that

$$|\nu_h(y + h(y)) \times E_h(y + h(y)) - \nu(y) \times E(y)| \to 0, \quad \|h\|_{C^2(\partial D)} \to 0,$$

uniformly for all $y \in \partial D$. Hence, in view of (3.6), using the impedance boundary condition for E_e, H_e together with Taylor's formula and the estimate $\nu_h(y + h(y)) - \nu(y) = O(\|h\|_{C^2(\partial D)})$ (see (2.15)) it follows that

$$\begin{aligned} E_h^s(x) - E^s(x) \ =\ & -\int_{\partial D_h} \left\{[H_e(\cdot,x)]^\top + \lambda\,[\nu_h \times E_e(\cdot,x)]^\top\right\}[\nu_h \times E]\, ds \\ & +o\left(\|h\|_{C^2(\partial D)}\right) \end{aligned}$$

uniformly for x on compact subsets of $\mathbb{R}^3 \setminus \bar{D}$. From this, abbreviating $E_e = E_e(x,\cdot)$ and $H_e = H_e(x,\cdot)$, and using the Gauss divergence theorem in view of (2.17), the boundary condition for E_e, H_e, and the Maxwell equations we obtain

$$\begin{aligned} E_h^s(x) - E^s(x) \ =\ & -\lambda \int_{\partial D_h} [\nu \times E_e]^\top [(\nu_h - \nu) \times E]\, ds \\ & -\int_{D_h^*} ik \left\{H_e^\top [H + \lambda\,\nu \times E] + E_e^\top E^*\right\} \chi\, dy \\ & +o\left(\|h\|_{C^2(\partial D)}\right), \end{aligned} \tag{3.15}$$

where D_h^* and χ are defined as in the proof of Lemma 2.4 and

$$E^* := E - \frac{\lambda}{ik}\,\mathrm{curl}[\nu \times E]. \tag{3.16}$$

In (3.15), we estimate the integral over ∂D_h with the aid of (2.16) and approximate the integral over D_h^* by an integral over ∂D. Then with the aid of (2.13), Taylor's formula, and the boundary condition for E, H we obtain that

$$\begin{aligned} E_h^s(x) - E^s(x) \;=\; & \int_{\partial D} \left\{ \lambda E_e^\top [\nu \times \{E \times \mathrm{Grad}(\nu \cdot h)\}] \right. \\ & \left. - ik \left[H_e^\top \{(H \cdot \nu)\,\nu\} + E_e^\top E^* \right] (\nu \cdot h) \right\} ds \\ & + o\left(\|h\|_{C^2(\partial D)} \right). \end{aligned} \tag{3.17}$$

With the aid of Maxwell's equation for E, the vector identities (3.12) and (3.13), and the boundary condition for E, H, we observe that

$$E^* \cdot \nu = \frac{1}{ik}\,\mathrm{Div}(\nu \times H - \lambda E_T) = 0 \quad \text{on } \partial D.$$

Using the Maxwell equation for H_e, the identity (3.12), and the Gauss surface divergence theorem we can transform

$$\int_{\partial D} ik\, H_e^\top (\nu \cdot H)(\nu \cdot h)\nu\, ds = -\int_{\partial D} E_e^\top [\nu \times \mathrm{Grad}\{(\nu \cdot H)(\nu \cdot h)\}]\, ds.$$

Inserting this into (3.17) we conclude that

$$\begin{aligned} E_h^s(x) - E^s(x) \;=\; & \int_{\partial D} \left\{ E_e^\top [\nu \times \{\mathrm{Grad}[(\nu \cdot H)(\nu \cdot h)] + \lambda\, E \times \mathrm{Grad}(\nu \cdot h)\}] \right. \\ & \left. - ik\, E_e^\top [\nu \times (E^* \times \nu)(\nu \cdot h)] \right\} ds \\ & + o\left(\|h\|_{C^2(\partial D)} \right). \end{aligned}$$

From this, the statement (3.14) follows with the help of the definition (3.16) and the identity (3.11). □

The Silver–Müller radiation condition (3.3) implies an asymptotic behavior in the form of a spherical wave

$$E(x) = \frac{e^{ik|x|}}{|x|} \left\{ E_\infty(\hat{x}) + O\left(\frac{1}{|x|} \right) \right\}, \quad |x| \to \infty,$$

uniformly for all directions $\hat{x} = x/|x|$ where the tangential field E_∞ on the unit sphere Ω is known as the electric far field pattern.

For a fixed incident field E^i, H^i, we define the operator

$$F : \partial D_h \to E_{\infty,h}$$

that maps the boundary ∂D_h onto the electric far field pattern $E_{\infty,h}$ of the scattered wave for scattering from the impedance obstacle D_h. For this operator we now are ready to prove the following result on Fréchet differentiability.

THEOREM 3.4. *Let ∂D be analytic. Then the boundary to far field operator $F : C(\partial D) \to L^2(\Omega)$ is Fréchet differentiable with the Fréchet derivative given through*

$$F'(\partial D)h = E_{\infty,h},$$

where $E_{\infty,h}$ is the far field pattern of the solution E_h, H_h to the Maxwell equations in $\mathbb{R}^3 \setminus \bar{D}$ that satisfies the Silver–Müller radiation condition and the impedance boundary condition

$$\begin{aligned} &\nu \times H_h - \lambda(\nu \times E_h) \times \nu \\ &\quad = -\nu \times \{\mathrm{Grad}[(\nu \cdot H)(\nu \cdot h)] - \lambda\, E \times \mathrm{Grad}(\nu \cdot h)\} \\ &\qquad +(\nu \cdot h)\left[ik - \lambda\left(R + 2H_{\partial D} - \frac{\partial}{\partial \nu}\right)\right][(\nu \times E) \times \nu] \quad \text{on } \partial D. \end{aligned} \tag{3.18}$$

Proof. This is an immediate consequence of Lemma 3.3 and the representation formula (3.5). □

We expect that, proceeding as in [11], Theorem 3.4 can be extended to the case of C^2 boundaries.

We will conclude with a few remarks on the injectivity of the Fréchet derivative as given in the previous theorem, i.e., the characterization of the nullspace of the derivative as the space of tangential fields $\nu \cdot h = 0$. In acoustic scattering, on one hand for the Dirichlet condition this result is an immediate consequence of the characterization of the derivative via Holmgren's uniqueness theorem (see Theorem 5.15 in [2]). On the other hand, the injectivity for the Neumann problem based on the characterization of Theorem 2.5 for $\lambda = 0$ remains an open problem. For the impedance problem for large λ, i.e., for the impedance condition close to the Dirichlet case, Kress and Rundell [12] (see also [10]) have settled the injectivity of the linearization.

The ideas for proving injectivity as applied in [12] fail for the Neumann boundary condition due to the occurrence of two definite integrals with opposite signs. Unfortunately the same happens in the electromagnetic case for $\lambda = 0$. In this case, in view of (3.18) we would want to conclude from

$$\nu \times \mathrm{Grad}[(\nu \cdot H)(\nu \cdot h)] - ik\, \nu \times [E \times \nu]\,(\nu \cdot h) = 0 \quad \text{on } \partial D \tag{3.19}$$

that $\nu \cdot h = 0$. Taking the dot product of (3.19) with the conjugate complex $\bar{E}$ and integrating over

$$U := \{x \in \partial D : \nu(x) \cdot h(x) \geq 0\}$$

with the aid of (3.12), the Maxwell equations and the Gauss divergence theorem yields

$$\int_U \left\{|\nu \times E|^2 - |\nu \cdot H|^2\right\}(\nu \cdot h)\, ds = 0$$

and no further conclusions are possible from this equation.

The following example actually shows that further assumptions on the incident field are required for establishing that the nullspace of the derivative consists only of the tangential fields. Consider as the incident field the vector spherical wave function

$$E^i(x) = \mathrm{curl}\left\{x j_n(k|x|) Y_n\left(\frac{x}{|x|}\right)\right\}, \quad H^i(x) = \frac{1}{ik}\,\mathrm{curl}\, E^i(x),$$

where j_n is the spherical Bessel function of order n and Y_n a spherical harmonic of order n. Elementary computations (see Theorem 6.24 in [2]) show that for D a ball of radius R centered at the origin the scattered wave is given by

$$E^s(x) = A_n(R)\,\mathrm{curl}\left\{x h_n^{(1)}(k|x|)Y_n\left(\frac{x}{|x|}\right)\right\}, \quad H^s(x) = \frac{1}{ik}\,\mathrm{curl}\,E^s(x), \tag{3.20}$$

where $h_n^{(1)}$ is the spherical Hankel function of order n and

$$A_n(R) = -\frac{(1+ik\lambda R)j_n(kR) + k\,j_n'(kR)}{(1+ik\lambda R)h_n^{(1)}(kR) + k\,h_n^{(1)\prime}(kR)}\,. \tag{3.21}$$

The denominator in (3.21) is nonzero, since for the radiating solution

$$u_n(x) = h_n^{(1)}(k|x|)Y_n\left(\frac{x}{|x|}\right)$$

to the Helmholtz equation we have the impedance boundary values

$$\frac{\partial u_n}{\partial \nu} + (1+ikR\lambda)u_n = \left\{k\,h_n^{(1)\prime}(kR) + (1+ik\lambda R)h_n^{(1)}(kR)\right\}Y_n \tag{3.22}$$

on ∂D. Therefore, due to the uniqueness result for the acoustic impedance boundary value problem as mentioned in section 2 the right-hand side of (3.22) cannot vanish identically.

The tangential component of the total electric field on ∂D is of the form

$$\nu \times E = a_n(R)\,\mathrm{Grad}\,Y_n,$$

where

$$a_n(R) := j_n(kR) + A_n(R)\,h_n^{(1)}(kR).$$

Then, with the aid of the Maxwell equations, the identities (3.12), and

$$\mathrm{Div}\,\mathrm{Grad}\,Y_n + n(n+1)Y_n = 0$$

for $R = 1$, $\lambda = 0$, and $\nu \cdot h = 1$, the right-hand side of (3.18) becomes

$$\nu \times \mathrm{Grad}[(\nu\cdot H)(\nu\cdot h)] - ik\,\nu\times[E\times\nu]\,(\nu\cdot h) = \frac{a_n(1)}{ik}\,[n(n+1)-k^2]\,\nu\times\mathrm{Grad}\,Y_n.$$

Hence, for $k^2 = n(n+1)$ we have vanishing boundary data and consequently a vanishing Fréchet derivative for a nontangential field h.

This observation, of course, can also be based on differentiating with respect to the radius R. The far field pattern E_∞ of (3.20) is given by

$$E_\infty = \frac{A_n(R)}{k\,i^{n+1}}\;\mathrm{Grad}\,Y_n \times \nu$$

with the exterior unit normal ν to the unit sphere Ω (see Theorem 6.26 in [2]). Straightforward calculations using the differential equation and the Wronskian for the spherical Bessel and Hankel functions (see section 2.4 in [2]) shows that

$$A_n'(1) = \frac{i[k^2(1-\lambda^2) - n(n+1)]}{k[(1+ik\lambda)h_n^{(1)}(k) + k\,h_n^{(1)\prime}(k)]^2}\,,$$

i.e., the derivative vanishes for $k^2(1-\lambda^2) = n(n+1)$ leading to vanishing Fréchet derivatives for a nontangential field h if λ is real and less than one. However, according to the symmetry of the Maxwell equations, if E, H solves the impedance boundary value problem with impedance constant λ, then $H, -E$ solves the impedance boundary value problem with impedance $1/\lambda$. Hence, our example also provides vanishing Fréchet derivatives for impedance values larger than one. These counter examples suggest that for proving injectivity of the derivative, for example, for plane wave incidence, the special structure of the plane waves has to be incorporated.

Acknowledgment. This research was initiated while Houssem Haddar was visiting the University of Göttingen. The hospitality of the University of Göttingen and the support are gratefully acknowledged.

REFERENCES

[1] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, John Wiley and Sons, New York, 1983.

[2] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer-Verlag, Berlin, Heidelberg, New York, 1998.

[3] F. Hettlich, *Fréchet derivatives in inverse obstacle scattering*, Inverse Problems, 11 (1995), pp. 371–382.

[4] F. Hettlich, *Erratum: Fréchet derivatives in inverse obstacle scattering*, Inverse Problems, 14 (1998), pp. 209–210.

[5] F. Hettlich, *The Domain Derivative in Inverse Obstacle Scattering*, Habilitation, Erlangen, 1999.

[6] T. Hohage, *Iterative Methods in Inverse Obstacle Scattering: Regularization Theory of Linear and Nonlinear Exponentially Ill-Posed Problems*, dissertation, Linz University, Linz, Austria, 1999.

[7] A. Kirsch, *The domain derivative and two applications in inverse scattering theory*, Inverse Problems, 9 (1993), pp. 81–96.

[8] R. Kress, *Electromagnetic waves scattering: Specific theoretical tools*, in Scattering, E. R. Pike and P. C. Sabatier, eds., Academic Press, London, 2001, pp. 175–190.

[9] R. Kress, *Electromagnetic waves scattering: Scattering by obstacles*, in Scattering, E. R. Pike and P. C. Sabatier, eds., Academic Press, London, 2001, pp. 191–210.

[10] R. Kress, *Uniqueness in inverse obstacle scattering*, in New Analytic and Geometric Methods in Inverse Problems, K. Bingham, Y. V. Kurylev, and E. Somersalo, eds, Springer-Verlag, Berlin, Heidelberg, New York, 2004, pp. 323–336.

[11] R. Kress and L. Päivärinta, *On the far field in obstacle scattering*, SIAM J. Appl. Math., 59 (1999), pp. 1413–1426.

[12] R. Kress and W. Rundell, *Inverse scattering for shape and impedance*, Inverse Problems, 17 (2001), pp. 1075–1085.

[13] C.M. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, Berlin, Heidelberg, New York, 1966.

[14] J.C. Nédélec, *Acoustic and Electromagnetic Equations*, Springer-Verlag, Berlin, Heidelberg, New York, 2001.

[15] R. Potthast, *Fréchet differentiability of boundary integral operators in inverse acoustic scattering*, Inverse Problems, 10 (1994), pp. 431–447.

[16] R. Potthast, *Fréchet differentiability of the solution to the acoustic Neumann scattering problem with respect to the domain*, J. Inverse Ill-posed Probl., 4 (1996), pp. 67–84.

[17] R. Potthast, *Domain derivatives in electromagnetic scattering*, Math. Methods Appl. Sci., 19 (1996), pp. 1157–1175.

[18] A. Roger, *Newton Kantorovich algorithm applied to an electromagnetic inverse problem*, IEEE Trans. Antennas and Propagation, AP-29 (1981), pp. 232–238.

[19] C. Schormann, *Analytische und numerische Untersuchungen bei inversen Transmissionsproblemen zur zeitharmonischen Wellengleichung*, dissertation, University of Göttingen, Göttingen, Germany, 2000.

SIAM J. APPL. MATH.
Vol. 65, No. 1, pp. 209–229

A MATHEMATICAL MODEL OF COMPETITION FOR TWO ESSENTIAL RESOURCES IN THE UNSTIRRED CHEMOSTAT*

JIANHUA WU†, HUA NIE,†, AND GAIL S. K. WOLKOWICZ‡

Abstract. A mathematical model of competition between two species for two growth-limiting, essential (complementary) resources in the unstirred chemostat is considered. The existence of a positive steady-state solution and some of its properties are established analytically. Techniques include the maximum principle, the fixed point index, and numerical simulations. The simulations also seem to indicate that there are regions in parameter space for which a globally stable positive equilibrium occurs and that there are other regions for which the model admits bistability and even multiple positive equilibria.

Key words. chemostat, essential or complementary resources, steady-state solution, fixed point index, numerical simulation

AMS subject classifications. 35K55, 35J65, 92A17

DOI. 10.1137/S0036139903423285

1. Introduction.

An apparatus called the chemostat, used for the continuous culture of microorganisms, has played an important role in ecology. It has been thought of as a *lake in a laboratory.* See [9, 25, 29] for a description of the apparatus and the general theory.

In the basic set up, the culture vessel is assumed to be well stirred. One or more populations of microorganisms grow and/or compete exploitatively for a single, nonreproducing, growth-limiting nutrient that is supplied at a constant rate. The contents of the culture vessel are removed at the same constant rate as the medium containing the nutrient is supplied, and thus the volume of the culture vessel remains constant. Species-specific parameters can be measured one species at a time, and based on these parameters the theory predicts the qualitative outcome in advance of actual competition. In particular, the theory predicts that the species with the lowest *break-even* concentration excludes all other competitors (see [6, 14, 29]). Experiments confirmed this prediction in the case of auxotrophic bacterial strains competing for limiting tryptophan [11].

Mathematical analysis of chemostat models involving two limiting resources under the assumption that the culture vessel is *well stirred* can be found, for example, in [2, 3, 7, 13, 12, 17, 18, 19, 28]. When more than one resource is limiting, it is necessary to consider how these resources promote growth. At one extreme are resources that are sources of different essential substances that must be taken together, because each substance fulfills different physiological needs with respect to growth, for example, a

*Received by the editors February 22, 2003; accepted for publication (in revised form) February 6, 2004; published electronically September 24, 2004. This work was supported in part by the Natural Science Foundation of China, the Excellent Young Teachers Program of the Ministry of Education of China, the Foundation for University Key Teacher of the Ministry of Education of China, the Scholarship Foundation of CSC, the Institute for Mathematics and its Applications with funds provided by the NSF, and by the Natural Sciences and Engineering Research Council of Canada.

http://www.siam.org/journals/siap/65-1/42328.html

†Department of Mathematics, Shaanxi Normal University, Xi'an, Shaanxi 710062, People's Republic of China (wjhua@snnu.edu.cn).

‡Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada L8S 4K1 (wolkowic@mcmaster.ca).

carbon source and a nitrogen source. Such resources are called complementary by Leon and Tumpson [17], Rapport [22], and Baltzis and Fredrickson [4]; essential by Tilman [28]; and heterologous by Harder and Dijkhuizen [12].

The model of exploitative competition for two essential resources in the well-stirred case is given by

$$\begin{aligned} S_t &= (S^0 - S)D - \tfrac{1}{y_{s_1}} g_1(S,R)u - \tfrac{1}{y_{s_2}} g_2(S,R)v, \\ R_t &= (R^0 - R)D - \tfrac{1}{y_{r_1}} g_1(S,R)u - \tfrac{1}{y_{r_2}} g_2(S,R)v, \\ u_t &= [-D + g_1(S,R)]u, \\ v_t &= [-D + g_2(S,R)]v. \end{aligned}$$

$S(t)$, $R(t)$ denote the nutrient concentrations at time t, and $u(t)$ and $v(t)$ denote the biomass of each population in the culture vessel. $S^0 > 0$ and $R^0 > 0$ are constants that represent the input concentrations of nutrients S and R, respectively, D is the dilution rate, and y_{s_i} and y_{r_i}, $i = 1, 2$, are the corresponding growth yield constants. The response functions are denoted $g_i(S,R) = \min(p_i(S), q_i(R))$, $i = 1, 2$, where $p_i(S)$ denotes the response function of the ith population when only resource S is limiting and $q_i(R)$ denotes the response function of the ith population when only resource R is limiting. We will consider the case that the Monod model for exploitative competition for one resource is generalized to the two essential resources case, i.e., $p_i(S) = \frac{m_{s_i} S}{K_{s_i} + S}$, $q_i(R) = \frac{m_{r_i} R}{K_{r_i} + R}$, $i = 1, 2$, where m_{s_i}, m_{r_i}, K_{s_i}, K_{r_i}, are positive constants.

In this paper, we study the *unstirred* chemostat and consider *two species' competition* for *two*, growth-limiting, nonreproducing *essential* resources. Motivated by the work on the unstirred chemostat in the case of one limiting resource (see [5, 8, 15, 16, 20, 23, 24, 25, 26, 30, 31]) and in the case of two limiting resources in [32], the model takes the form of the following reaction-diffusion equations:

$$\begin{aligned} S_t &= dS_{xx} - \tfrac{1}{y_{s_1}} g_1(S,R)u - \tfrac{1}{y_{s_2}} g_2(S,R)v, && 0 < x < 1,\ t > 0, \\ R_t &= dR_{xx} - \tfrac{1}{y_{r_1}} g_1(S,R)u - \tfrac{1}{y_{r_2}} g_2(S,R)v, && 0 < x < 1,\ t > 0, \\ u_t &= du_{xx} + g_1(S,R)u, && 0 < x < 1,\ t > 0, \\ v_t &= dv_{xx} + g_2(S,R)v, && 0 < x < 1,\ t > 0, \end{aligned}$$

with boundary conditions

$$\begin{aligned} &S_x(0,t) = -S^0, && R_x(0,t) = -R^0,\ u_x(0,t) = 0,\ v_x(0,t) = 0, \\ &S_x(1,t) + \gamma S(1,t) = 0, && R_x(1,t) + \gamma R(1,t) = 0, \\ &u_x(1,t) + \gamma u(1,t) = 0, && v_x(1,t) + \gamma v(1,t) = 0. \end{aligned}$$

The boundary conditions are very intuitive. Readers may refer to [5, 16, 26] for their derivation.

These equations can be simplified using the nondimensional variables and parameters defined as follows: $\bar{S} = \frac{S}{S^0}$, $\bar{R} = \frac{R}{R^0}$, $\alpha = \frac{S^0 y_{s_1}}{R^0 y_{r_1}}$, $\beta = \frac{R^0 y_{r_2}}{S^0 y_{s_2}}$, $\bar{g}_i(\bar{S}, \bar{R}) = \min(\frac{m_{s_i}\bar{S}}{\bar{K}_{s_i} + \bar{S}}, \frac{m_{r_i}\bar{R}}{\bar{K}_{r_i} + \bar{R}})$, $i = 1, 2$, $\bar{u} = \frac{u}{y_{s_1} S^0}$, $\bar{v} = \frac{v}{y_{r_2} R^0}$, where $\bar{K}_{s_i} = \frac{K_{s_i}}{S^0}$, $\bar{K}_{r_i} = \frac{K_{r_i}}{R^0}$, $i = 1, 2$. For more convenient notation, we drop the bars on the nondimensional

variables and parameters, yielding the following model:

$$
\begin{aligned}
S_t &= dS_{xx} - g_1(S,R)u - \beta g_2(S,R)v, && 0<x<1,\ t>0,\\
R_t &= dR_{xx} - \alpha g_1(S,R)u - g_2(S,R)v, && 0<x<1,\ t>0,\\
u_t &= du_{xx} + g_1(S,R)u, && 0<x<1,\ t>0,\\
v_t &= dv_{xx} + g_2(S,R)v, && 0<x<1,\ t>0,
\end{aligned}
\tag{1}
$$

with boundary conditions

$$
\begin{aligned}
&S_x(0,t) = -1,\ \ R_x(0,t) = -1,\ \ u_x(0,t) = 0,\ v_x(0,t) = 0,\\
&S_x(1,t) + \gamma S(1,t) = 0,\ \ R_x(1,t) + \gamma R(1,t) = 0,\\
&u_x(1,t) + \gamma u(1,t) = 0,\ \ v_x(1,t) + \gamma v(1,t) = 0,
\end{aligned}
$$

and initial conditions

$$
S(x,0) = S_0(x) \ge 0,\ R(x,0) = R_0(x) \ge 0,\ u(x,0) = u_0(x) \ge 0,\ v(x,0) = v_0(x) \ge 0.
$$

Denote $\varphi_1 = S + u + \beta v,\ \varphi_2 = R + \alpha u + v$, where $\varphi_i,\ i = 1,2$, is the solution of

$$
\begin{aligned}
&\varphi_{it} = d\varphi_{ixx}, && 0<x<1,\ t>0,\\
&\varphi_{ix}(0,t) = -1,\ \ \varphi_{ix}(1,t) + \gamma\varphi_i(1,t) = 0,\\
&\varphi_i(x,0) = \varphi_{i0}(x).
\end{aligned}
$$

Then u and v satisfy

$$
\begin{aligned}
u_t &= du_{xx} + ug_1(\varphi_1 - u - \beta v, \varphi_2 - \alpha u - v), && 0<x<1,\ t>0,\\
v_t &= dv_{xx} + vg_2(\varphi_1 - u - \beta v, \varphi_2 - \alpha u - v), && 0<x<1,\ t>0,\\
&u_x(0,t) = 0,\ \ u_x(1,t) + \gamma u(1,t) = 0, && t>0,\\
&v_x(0,t) = 0,\ \ v_x(1,t) + \gamma v(1,t) = 0, && t>0.
\end{aligned}
\tag{1$'$}
$$

This paper is devoted to determining the positive solution of this two-species model of exploitative competition for *two essential resources* in the *unstirred chemostat*. Since the reaction terms are Lipschitz continuous, but not C^1, many methods used to analyze elliptic systems do not apply. This makes the analysis more difficult. Some methods used to prove the existence of the positive equilibrium in the region $D = \{(\hat{\lambda}_1, \hat{\lambda}_2) : \hat{\lambda}_1 > 1,\ \hat{\lambda}_2 > 1\}$ occupy a major portion of the paper, where $\hat{\lambda}_i,\ i = 1,2$, is defined in the next section. The main result is established in Theorem 3. The other related results are also obtained in section 2. Extensive numerical studies were run, and some conclusions are summarized in section 3. The simulations convince us that much more complex dynamics can occur in region D.

The paper is organized as follows. In section 2, the existence of a positive steady-state solution and some of its properties are established by using the maximum principle and fixed point index theory, which is closely related to bounding the principal eigenvalues of certain differential operators. Some results on extensive numerical studies are reported in section 3, complementing the mathematical results in section 2, and a number of typical figures chosen from many simulations are also given in this section.

2. The positive steady-state solution. First, we consider the steady state of system (1):

$$
\begin{aligned}
&dS_{xx} - g_1(S,R)u - \beta g_2(S,R)v = 0, && 0 < x < 1,\\
&dR_{xx} - \alpha g_1(S,R)u - g_2(S,R)v = 0, && 0 < x < 1,\\
&du_{xx} + g_1(S,R)u = 0, && 0 < x < 1,\\
&dv_{xx} + g_2(S,R)v = 0, && 0 < x < 1,
\end{aligned} \tag{2}
$$

with boundary conditions

$$
\begin{aligned}
&S_x(0) = -1,\ \ R_x(0) = -1,\ \ u_x(0) = 0,\ v_x(0) = 0,\\
&S_x(1) + \gamma S(1) = 0,\ \ R_x(1) + \gamma R(1) = 0,\ \ u_x(1) + \gamma u(1) = 0,\ \ v_x(1) + \gamma v(1) = 0.
\end{aligned}
$$

It follows that $S + u + \beta v = z$, $R + \alpha u + v = z$, where $z = z(x) = \frac{1+\gamma}{\gamma} - x$. Then u and v satisfy

$$
\begin{aligned}
&du_{xx} + ug_1(z - u - \beta v, z - \alpha u - v) = 0, && 0 < x < 1,\\
&dv_{xx} + vg_2(z - u - \beta v, z - \alpha u - v) = 0, && 0 < x < 1,\\
&u_x(0) = 0,\ \ u_x(1) + \gamma u(1) = 0,\\
&v_x(0) = 0,\ \ v_x(1) + \gamma v(1) = 0.
\end{aligned} \tag{3}
$$

Let λ_i be the principal eigenvalue and let $\phi_i(x) > 0$ on $[0,1]$, $i = 1, 2$, be the corresponding eigenfunction, normalized as $\max_{x\in[0,1]} \phi_i(x) = 1$, of the following problem:

$$
d\phi_{ixx} + \lambda_i \phi_i g_i(z,z) = 0, \quad 0 < x < 1, \quad \phi_{ix}(0) = 0, \quad \phi_{ix}(1) + \gamma\phi_i(1) = 0. \tag{4}
$$

Let $U(x)$ be the solution of

$$
\begin{aligned}
&dU_{xx} + Ug_1(z - U, z - \alpha U) = 0, && 0 < x < 1,\\
&U_x(0) = 0,\ \ U_x(1) + \gamma U(1) = 0,
\end{aligned} \tag{5}
$$

and let $U(x,t)$ be the solution of

$$
\begin{aligned}
&U_t = dU_{xx} + Ug_1(\varphi_1 - U, \varphi_2 - \alpha U), && 0 < x < 1,\ t > 0,\\
&U_x(0,t) = 0,\ \ U_x(1,t) + \gamma U(1,t) = 0,\\
&U(x,0) = U_0(x) \geq 0.
\end{aligned} \tag{6}
$$

From Lemmas 2.2–2.4 and Theorem 2.5 in [32] we have the following lemma.

LEMMA 1. *If $\lambda_1 < 1$, then there exists a unique positive solution $U(x)$ of (5), satisfying $0 < U < \min\{1, \frac{1}{\alpha}\}z$ on $[0,1]$. If $\lambda_1 \geq 1$, the only nonnegative solution of (5) is $U = 0$. Furthermore, $\lim_{t\to\infty} U(x,t) = U(x)$ if $\lambda_1 < 1$, and $\lim_{t\to\infty} U(x,t) = 0$ if $\lambda_1 > 1$.*

Remark 1. If $\lambda_2 < 1$, a similar result holds for $V(x)$, where $V(x)$ is the unique positive solution of

$$
\begin{aligned}
&dV_{xx} + Vg_2(z - \beta V, z - V) = 0, && 0 < x < 1,\\
&V_x(0) = 0,\ \ V_x(1) + \gamma V(1) = 0.
\end{aligned}
$$

Since we are only concerned with the nonnegative steady-state solutions of (3), there is no loss of generality if we redefine

$$p_i(S) = \begin{cases} \frac{m_{s_i} S}{K_{s_i}+S}, & S \geq 0, \\ 0, & S < 0, \end{cases} \quad q_i(R) = \begin{cases} \frac{m_{r_i} R}{K_{r_i}+R}, & R \geq 0, \\ 0, & R < 0. \end{cases}$$

LEMMA 2. *Suppose* (u, v) *is the nonnegative solution of* (3). *Then* (i) $u > 0$ *or* $u \equiv 0$, *and* $v > 0$ *or* $v \equiv 0$; (ii) $u + \beta v < z$, $\alpha u + v < z$; (iii) $u \leq U, v \leq V$. *Moreover,* $u < U$ *or* $u \equiv U$, *and* $v < V$ *or* $v \equiv V$.

Proof. (i) This part can be proved by the maximum principle, in the usual way, and the details are omitted here.

(ii) Define $w = u + \beta v - z$. Note that by (3) it follows that

$$dw_{xx} + ug_1(-w, z - \alpha u - v) + \beta v g_2(-w, z - \alpha u - v) = 0,$$
$$w_x(0) = 1 \quad \text{and} \quad w_x(1) + \gamma w(1) = 0.$$

First we show that $w \leq 0$ on $[0, 1]$. Suppose not. If $w(1) > 0$, then $w_x(1) < 0$. Therefore, there exists $a \in [0, 1)$ so that for all $x \in (a, 1]$, $w(x) > 0$, and either $a = 0$ or $w(a) = 0$. But then for all $x \in [a, 1]$, $w_{xx} = 0$ and so $w_x(x) = w_x(1) < 0$, i.e., $w(x)$ is decreasing there. Since $w_x(0) = 1 > 0$, $a \neq 0$. But $a > 0$ is also impossible since then $w(a) = 0$, $w(x)$ is decreasing in $[a, 1]$, and $w(1) > 0$. Therefore, $w(1) \leq 0$. Next, assume there exists $\bar{x} \in [0, 1)$ with $w(\bar{x}) > 0$. Then there exist $\delta_1 \geq 0$ and $\delta_2 > 0$ such that $w(x) > 0$ for all $x \in (\bar{x} - \delta_1, \bar{x} + \delta_2) \subset (0, 1)$, $w(\bar{x} + \delta_2) = 0$, and either $\bar{x} - \delta_1 = 0$ or $w(\bar{x} - \delta_1) = 0$. But then for all $x \in [\bar{x} - \delta_1, \bar{x} + \delta_2]$, $w_{xx}(x) = 0$ and so $w_x(x)$ is constant. Since $w(\bar{x} + \delta_2) = 0$, it follows that $w_x(\bar{x} + \delta_2) \leq 0$, and so $w(x)$ is nonincreasing on $[\bar{x} - \delta_1, \bar{x} + \delta_2]$. Then $\bar{x} - \delta_1 \neq 0$, since $w_x(0) = 1$, and so $w(\bar{x} - \delta_1) = 0$. Therefore, $w(x) \equiv 0$ on $[\bar{x} - \delta_1, \bar{x} + \delta_2]$, a contradiction. Hence, $u + \beta v \leq z$ on $[0, 1]$. That $\alpha u + v \leq z$ follows similarly. It is easy to see that $u + \beta v \not\equiv z$, $\alpha u + v \not\equiv z$; otherwise we have $du_{xx} = 0$, $dv_{xx} = 0$, with the usual boundary condition, which gives $u \equiv 0$, $v \equiv 0$, a contradiction. Let $w_1 = z - u - \beta v$, $w_2 = z - \alpha u - v$. Then $w_i \geq 0, \not\equiv 0$, and w_1 satisfies

$$-dw_{1xx} + ug_1(w_1, w_2) + \beta v g_2(w_1, w_2) = 0,$$
$$w_{1x}(0) = -1, \quad w_{1x}(1) + \gamma w_1(1) = 0,$$

which leads to

$$-dw_{1xx} + w_1 \left(\frac{m_{s_1} u}{K_{s_1}+w_1} + \frac{m_{s_2} \beta v}{K_{s_2}+w_1} \right) \geq 0,$$
$$w_{1x}(0) = -1, \quad w_{1x}(1) + \gamma w_1(1) = 0.$$

If $w_1(x_0) = 0$ for some point $x_0 \in [0, 1]$, by applying the strong maximum principle (see [21]) we obtain a contradiction. Hence $w_1 > 0$ on $[0, 1]$. The proof that $w_2 > 0$ on $[0, 1]$ is similar.

(iii) It follows by the monotone method and the uniqueness of U that $u \leq U \leq \min\{1, \frac{1}{\alpha}\} z$. By the Lipschitz continuity of $g_1(S, R)$, there exists a constant $L > 0$, such that $0 \leq g_1(z - u, z - \alpha u) - g_1(z - U, z - \alpha U) \leq L(U - u)$. Let $\hat{U} = U - u$. Then $\hat{U} \geq 0$ satisfies

$$d\hat{U}_{xx} + \hat{U}[g_1(z - U, z - \alpha U) - uL] \leq 0, \quad 0 < x < 1,$$
$$\hat{U}_x(0) = 0, \quad \hat{U}_x(1) + \gamma \hat{U}(1) = 0.$$

If $\hat{U} \not\equiv 0$, then the maximum principle leads to $\hat{U} > 0$. Thus either $u < U$ or $u \equiv U$. The proof for v is similar. ∎

Remark 2. It follows from Lemmas 1 and 2 that, for $\lambda_i \geq 1$, $i = 1, 2$, the only nonnegative solution of (3) is $(0, 0)$. In order to guarantee the existence of a positive solution of (3), we must assume that $\lambda_i < 1$ for $i = 1, 2$.

Let $\hat{\lambda}_i$ be the principal eigenvalues and let $\hat{\phi}_i(x) > 0,\ x \in [0, 1],\ i = 1, 2$, be the corresponding eigenfunctions of the problem

$$d\hat{\phi}_{1xx} + \hat{\lambda}_1\hat{\phi}_1 g_1(z - \beta V, z - V) = 0, \quad 0 < x < 1, \quad \hat{\phi}_{1x}(0) = 0,\ \ \hat{\phi}_{1x}(1) + \gamma\hat{\phi}_1(1) = 0,$$

$$d\hat{\phi}_{2xx} + \hat{\lambda}_2\hat{\phi}_2 g_2(z - U, z - \alpha U) = 0, \quad 0 < x < 1, \quad \hat{\phi}_{2x}(0) = 0,\ \ \hat{\phi}_{2x}(1) + \gamma\hat{\phi}_2(1) = 0.$$

THEOREM 1. *Suppose* $\hat{\lambda}_i < 1$ *for* $i = 1, 2$. *Then there exists a positive steady-state solution* (u, v) *of* (3) *satisfying* $0 < u(x) < U(x),\ 0 < v(x) < V(x)$ *for* $x \in [0, 1]$.

Proof. It is easy to check that (U, V) is the sup-solution of (3). Let $(\underline{u}, \underline{v}) = (\delta\hat{\phi}_1, \delta\hat{\phi}_2)(\delta > 0)$. Then for δ sufficiently small, we have

$$\begin{aligned}
&d\underline{u}_{xx} + \underline{u}g_1(z - \underline{u} - \beta V, z - \alpha\underline{u} - V)\\
&= [\underline{u}g_1(z - \underline{u} - \beta V, z - \alpha\underline{u} - V) - \hat{\lambda}_1\underline{u}g_1(z - \beta V, z - V)]\\
&= \underline{u}[(1 - \hat{\lambda}_1)g_1(z - \beta V, z - V)\\
&\qquad + (g_1(z - \underline{u} - \beta V, z - \alpha\underline{u} - V) - g_1(z - \beta V, z - V))] > 0.
\end{aligned}$$

Hence there exists a solution (u, v) of (3) satisfying $(\delta\hat{\phi}_1, \delta\hat{\phi}_2) \leq (u, v) \leq (U, V)$ for small δ. By Lemma 2 we obtain the strict inequalities in Theorem 1. □

Now we consider the special case that $g_1 = g_2 = g$, and we find that there exist infinitely many positive solutions of (3).

THEOREM 2. *Suppose that* $\lambda_i < 1$ *for* $i = 1, 2$ *and* $g_1 = g_2 = g$. *Then there exist infinitely many positive solutions* (u_ρ, v_ρ) $(\rho > 0)$ *of* (3) *satisfying* $0 < v_\rho \leq \min\{\frac{1}{\rho+\beta}, \frac{1}{\alpha\rho+1}\}z,\ u_\rho = \rho v_\rho$.

Proof. Set $\omega = \frac{u}{v}$. Then ω satisfies

$$-d\omega_{xx} - \frac{2dv_x}{v}\omega_x = 0, \quad \omega_x(0) = \omega_x(1) = 0.$$

By the maximum principle it follows that $\omega \equiv \rho$, a positive constant, i.e., $u = \rho v$. Thus v satisfies

$$dv_{xx} + vg(z - (\rho + \beta)v, z - (\alpha\rho + 1)v) = 0, \quad v_x(0) = 0,\ \ v_x(1) + \gamma v_x(1) = 0.$$

For $\rho > 0$ fixed, arguing as for the existence of U or V, and noting that $\lambda_2 < 1$, it follows that there exists a unique positive solution of the above problem, say, v_ρ, satisfying $0 < v_\rho \leq \min\{\frac{1}{\rho+\beta}, \frac{1}{\alpha\rho+1}\}z$. Thus (u_ρ, v_ρ) $(\rho > 0)$, where $u_\rho = \rho v_\rho$, is the positive solution of (3). This completes the proof. □

Remark 3. Suppose that $g_1 \leq g_2$, $g_1 \not\equiv g_2$ or $g_1 \geq g_2$, $g_1 \not\equiv g_2$. Then there exists no positive solution of (3). This conclusion is consistent with the analysis in [9] for the pure and simple competition model. In fact, suppose $u > 0,\ v > 0$ satisfy (3). We consider the first case, since the second case can be proved similarly. Denoting $\omega = \frac{u}{v}$, we have

$$\begin{aligned}
&-d\omega_{xx} - \frac{2dv_x}{v}\omega_x + \omega[g_2(z - u - \beta v, z - \alpha u - v) - g_1(z - u - \beta v, z - \alpha u - v)] = 0,\\
&\omega_x(0) = \omega_x(1) = 0.
\end{aligned}$$

Then $\omega =$ constant, and hence $\omega = 0$, a contradiction.

THEOREM 3. *Suppose* $\lambda_i < 1$ *and* $\hat{\lambda}_i > 1$ *for* $i = 1, 2$. *Then there exists a positive solution* (u, v) *of* (3).

Proof. Let $C_B[0,1] = \{u(x) \in C[0,1] : u_x(0) = 0,\ u_x(1) + \gamma u(1) = 0\}$ be the Banach space, with the usual maximum norm, denoted by $\|\cdot\|$, $X = C_B[0,1] \times C_B[0,1]$, $K = C_B^+[0,1] \times C_B^+[0,1]$, the positive cone of X. Let $N = (-d\Delta)^{-1}$, the inverse operator of $-d\Delta$ in $C_B[0,1]$. Then system (3) can be written as

$$u - N(ug_1(z - u - \beta v, z - \alpha u - v)) = 0,$$
$$v - N(vg_2(z - u - \beta v, z - \alpha u - v)) = 0.$$

Let $T(u,v) = (N(ug_1(z - u - \beta v, z - \alpha u - v)), N(vg_2(z - u - \beta v, z - \alpha u - v)))$. Then the fixed points of T in K are the corresponding nonnegative solutions of (3). Define $D = \{(u,v) \in K : \|u\| + \|v\| \le R_0\}$, where $R_0 = 2\max\{1, \frac{1}{\alpha}, \frac{1}{\beta}\}\|z\|$, and let $\dot{D}$ denote the interior of D in K. Since the proof is long, we divide it into three lemmas.

LEMMA 3. *For* $\lambda \ge 1$, *the equation* $T(u,v) = \lambda(u,v)$ *has no solution in* K *satisfying* $\|u\| + \|v\| = R_0$.

Proof. Suppose $(u,v) \in K$ satisfies $T(u,v) = \lambda(u,v)$. Then we have

$$du_{xx} + \lambda^{-1} ug_1(z - u - \beta v, z - \alpha u - v) = 0,$$
$$dv_{xx} + \lambda^{-1} vg_2(z - u - \beta v, z - \alpha u - v) = 0,$$

with the boundary conditions as above. As in the proof of Lemma 2, it follows that $u + \beta v < z, \alpha u + v < z$. Thus $u + v < \max\{1, \frac{1}{\alpha}, \frac{1}{\beta}\}z$. Hence there exists no fixed point of $T(u,v) = \lambda(u,v)$ in K satisfying $\|u\| + \|v\| = R_0$. □

Remark 4. It follows from Lemma 12.1 in [1] that $index_K(T, \dot{D}) = 1$.

Let $P_\sigma(0,0) = \{(u,v) \in K : \|u\| + \|v\| < \sigma\}$ be the neighborhood of $(0,0)$ in K with radius σ.

LEMMA 4. *For* $\sigma > 0$ *small enough,* $index_K(T, P_\sigma(0,0)) = 0$.

Proof. Given $\epsilon_0 > 0$ sufficiently small, noting the definition of U, V, we can take $0 < \sigma < \sigma_0 \ll 1$ such that $\frac{\sigma}{\gamma} < \min\{U - \epsilon_0, V - \epsilon_0\}$. Denote $S_\sigma^+ = \{(u,v) \in K : \|u\| + \|v\| = \frac{\sigma}{\gamma}\}$. Thus $\|u\| \le \sigma z,\ \|v\| \le \sigma z$ whenever $(u,v) \in S_\sigma^+$.

Let $\psi = (2 + \gamma) - \gamma x^2$. Then $\psi > 0$ on $[0,1]$ and satisfies

$$\psi_{xx} < 0, \quad 0 < x < 1, \quad \psi_x(0) = 0, \quad \psi_x(1) + \gamma\psi(1) = 0.$$

Take $p = (\psi, \psi)(\in K)$. We show next (by contradiction) that for $\lambda \ge 0$, $(u,v) - T(u,v) = \lambda(\psi, \psi)$ has no solution on S_σ^+ for small σ. Assume that this problem has a solution (u,v) on S_σ^+. Then (u,v) satisfies

$$du_{xx} + ug_1(z - u - \beta v, z - \alpha u - v) = d\lambda\psi_{xx}, \quad 0 < x < 1,$$
$$dv_{xx} + vg_2(z - u - \beta v, z - \alpha u - v) = d\lambda\psi_{xx}, \quad 0 < x < 1.$$

Hence by the definition of ψ, we have

$$du_{xx} + ug_1((1 - \sigma\beta)z - u, (1 - \sigma)z - \alpha u) \le 0, \quad 0 < x < 1,$$
$$dv_{xx} + vg_2((1 - \sigma)z - \beta v, (1 - \sigma\alpha)z - v) \le 0, \quad 0 < x < 1.$$

Since $\lambda_i < 1$, we can take sufficiently small σ, say, $\sigma < \sigma_1 \ll 1$, such that $\lambda_1(g_1((1 - \sigma\beta)z, (1 - \sigma)z)) < 1,\ \lambda_2(g_2((1 - \sigma)z, (1 - \sigma\alpha)z)) < 1$, where $\lambda_1(g_1((1 - \sigma\beta)z, (1 - \sigma)z))$, $\lambda_2(g_2((1 - \sigma)z, (1 - \sigma\alpha)z))$ are the principal eigenvalues of (4) with g_1 and g_2

replaced by $g_1((1-\sigma\beta)z,(1-\sigma)z)$ and $g_2((1-\sigma)z,(1-\sigma\alpha)z)$, respectively. As in the proof of Lemma 3.2 in [31] we can prove the existence and uniqueness of U^*, V^* of the following problem:

$$\begin{aligned} dU^*_{xx} + U^* g_1((1-\sigma\beta)z - U^*, (1-\sigma)z - \alpha U^*) = 0, \qquad 0 < x < 1,\\ dV^*_{xx} + V^* g_2((1-\sigma)z - \beta V^*, (1-\sigma\alpha)z - V^*) = 0, \qquad 0 < x < 1, \end{aligned}$$

with the usual boundary conditions. By an L^p estimate and the Sobolev embedding theorem (see [27]), we proceed as in the proof of Theorem 2.5 in [32] to obtain

$$\lim_{\sigma\to 0} U^* = U, \ \lim_{\sigma\to 0} V^* = V.$$

Thus there exists $\sigma_2 > 0$, such that for $\sigma < \sigma_2$, $U^* > U - \epsilon_0$, $V^* > V - \epsilon_0$. It follows from the monotone method and the uniqueness of U^*, V^* that $u \geq U^*$, $v \geq V^*$. Now take $\sigma < \bar{\sigma} = \min\{\sigma_0, \sigma_1, \sigma_2\}$. Then for $\sigma < \bar{\sigma}$, we have $u > \frac{\sigma}{\gamma}, v > \frac{\sigma}{\gamma}$, which contradicts $(u,v) \in S^+_\sigma$. Lemma 12.1 of [1] can be applied to complete the proof of this lemma. □

Let $O^+(U,0)$ be a small neighborhood of $(U,0)$ in K. Then we have the following lemma.

LEMMA 5. *Suppose that T has no fixed point in $\dot{D}$. Then $index_K(T, O^+(U,0)) = 1$ if $\hat{\lambda}_2 > 1$, $\lambda_1 < 1$.*

Proof. Define $T(\theta)(u,v) = (N(ug_1(z-u-\theta\beta v, z-\alpha u-\theta v)), N(vg_2(z-u-\theta\beta v, z-\alpha u-\theta v)))$. It follows from $(u,v) = T(\theta)(u,v)$ that

$$\begin{aligned} du_{xx} + ug_1(z-u-\theta\beta v, z-\alpha u-\theta v) = 0,\\ dv_{xx} + vg_2(z-u-\theta\beta v, z-\alpha u-\theta v) = 0. \end{aligned}$$

If (u,v) is a fixed point of $T(\theta)$ on $\partial O^+(U,0)$, the boundary of $O^+(U,0)$ in K, it is easy to see that $u > 0$, $v \geq 0$. Furthermore, we have $v > 0$; otherwise we have $(u,v) = (U,0)$, contradicting $(u,v) \in \partial O^+(U,0)$. We claim that for $\theta \in [0,1]$, $T(\theta)$ has no fixed point on $\partial O^+(U,0)$. Otherwise, for $\theta = 0$, by noting $\hat{\lambda}_2 > 1$ and $\lambda_1 < 1$, we find $u = U$, $v = 0$, a contradiction; for $\theta > 0$, this implies that $(u, \theta v) > (0,0)$ is a fixed point of T in $\dot{D}$, contradicting a hypothesis of this lemma. It follows from the homotopy invariance of topological degree that

$$index_K(T, O^+(U,0)) = index_K(T(1), O^+(U,0)) = index_K(T(0), O^+(U,0)), \tag{7}$$

where $T(0)(u,v) = (N(ug_1(z-u, z-\alpha u)), N(vg_2(z-u, z-\alpha u)))$.

The fixed point (u,v) of $T(0)$ in $O^+(U,0)$ satisfies

$$\begin{aligned} du_{xx} + ug_1(z-u, z-\alpha u) = 0, \quad 0 < x < 1,\\ dv_{xx} + vg_2(z-u, z-\alpha u) = 0, \quad 0 < x < 1, \end{aligned} \tag{8}$$

with the boundary conditions

$$u_x(0) = 0, \ \ v_x(0) = 0, \ \ u_x(1) + \gamma u(1) = 0, \ \ v_x(1) + \gamma v(1) = 0.$$

Since $\lambda_1 < 1$, we have $u = U$. Noting $\hat{\lambda}_2 > 1$, we determine that the principal eigenvalue λ'_2 of the following problem is negative:

$$d\phi'_{xx} + \phi' g_2(z-U, z-\alpha U) = \lambda'_2 \phi', \quad \phi'_x(0) = 0, \quad \phi'_x(1) + \gamma\phi'(1) = 0.$$

Substituting $u = U$ into the second equation of (8), we have $v = 0$. Hence $(U, 0)$ is the unique fixed point of $T(0)$ in $O^+(U, 0)$; thus

$$index_K(T(0), O^+(U, 0)) = index_K(T(0), (U, 0)). \tag{9}$$

Let $I(\theta)$ $(\theta \in [0, 1])$ be defined by $I(\theta)(u, v) = (N(ug_1(z - u, z - \alpha u)), N(vg_2(z - (\theta U + (1 - \theta)u), z - \alpha(\theta U + (1 - \theta)u))))$. Then $(u, v) = I(\theta)(u, v)$ satisfies

$$\begin{aligned} &du_{xx} + ug_1(z - u, z - \alpha u) = 0, && 0 < x < 1, \\ &dv_{xx} + vg_2(z - (\theta U + (1 - \theta)u), z - \alpha(\theta U + (1 - \theta)u)) = 0, && 0 < x < 1, \end{aligned} \tag{10}$$

with the usual boundary conditions. We claim that $I(\theta)$ has no fixed point on $\partial O^+(U, 0)$ in K. Otherwise, from the first equation of (10), we have $u = U$, and substituting this into the second equation of (10), we find $v = 0$, so the only fixed point of $I(\theta)$ on $\partial O^+(U, 0)$ is $(U, 0)$, a contradiction. By the definition of $I(\theta)$, we obtain

$$T(0) = I(0), \quad I(1) = T_1 \times T_2, \tag{11}$$

where $T_1 u = N(ug_1(z - u, z - \alpha u))$, $T_2 v = N(vg_2(z - U, z - \alpha U))$, $(T_1 \times T_2)(u, v) = (T_1 u, T_2 v)$. $(u, v) = I(1)(u, v)$ satisfies

$$\begin{aligned} &du_{xx} + ug_1(z - u, z - \alpha u) = 0, && 0 < x < 1, \\ &dv_{xx} + vg_2(z - U, z - \alpha U) = 0, && 0 < x < 1. \end{aligned}$$

It follows from (7)–(11) and the product theorem for fixed points (see [33]) that

$$\begin{aligned} index_K(T(0), (U, 0)) &= index_K(I(0), (U, 0)) = index_K(I(1), (U, 0)) \\ &= index_{C_B}(T_1, U) \cdot index_{C_B^+}(T_2, 0). \end{aligned} \tag{12}$$

Since T_2 is a linear compact operator and $\hat{\lambda}_2 > 1$, then T_2 has no eigenvalue > 1 with positive eigenfunction in C_B^+. It follows from Lemma 13.1 of [1] that $index_{C_B^+}(T_2, 0) = 1$.

We show next that $index_{C_B}(T_1, U) = 1$. Let $\tau = 2 \min\{1, \frac{1}{\alpha}\}\|z\|$, $P_\tau = \{u \in C_B^+ : \|u\| \le \tau\}$, $\partial P_\tau = \{u \in C_B^+ : \|u\| = \tau\}$. For $\lambda \ge 1$, if $T_1 u = \lambda u$, $d\lambda u_{xx} + ug_1(z - u, z - \alpha u) = 0$. Arguing as in the proof of Lemma 1, we have $u \le \min\{1, \frac{1}{\alpha}\}z < \tau$. Hence for $\lambda \ge 1$, $T_1 u = \lambda u$ has no solution on ∂P_τ. It follows from Lemma 12.1 of [1] that $index_{C_B^+}(T_1, P_\tau) = 1$. Let $0 < \tau_0 \le \frac{1}{2} \min_{[0,1]}\{U(x)\}$. Suppose that for $\lambda \ge 0$, $p = \psi(x)$, such that $u - T_1 u = \lambda p$ has a solution u on ∂P_{τ_0}, where $\psi(x)$ is defined as in the proof of Lemma 4. Then, $du_{xx} + ug_1(z - u, z - \alpha u) = d\lambda\psi_{xx} \le 0$. Thus u is a sup-solution of (5). From the monotone method and the uniqueness of U it follows that $u \ge U$, a contradiction to $\|u\| = \tau_0$. Hence, $index_{C_B^+}(T_1, P_{\tau_0}) = 0$. Since $u = U$ is the unique fixed point of T_1 in $P_\tau \backslash \bar{P}_{\tau_0}$, we obtain $index_{C_B}(T_1, U) = index_{C_B^+}(T_1, P_\tau \backslash \bar{P}_{\tau_0}) = index_{C_B^+}(T_1, P_\tau) - index_{C_B^+}(T_1, P_{\tau_0}) = 1$.

Combining the above result with equations (7), (9), and (12), it follows that $index_K(T, O^+(U, 0)) = 1$. □

Remark 5. Suppose that T has no fixed point in $\dot{D}$. We can proceed as above to obtain $index_K(T, O^+(0, V)) = 1$ if $\hat{\lambda}_1 > 1$, $\lambda_2 < 1$.

Now we turn to the proof of Theorem 3. Suppose that T has no fixed point in $\dot{D}$. Then the following equation holds:

$$index_K(T, \dot{D}) = index_K(T, O^+(0, 0)) + index_K(T, O^+(U, 0)) + index_K(T, O^+(0, V)),$$

contradicting Lemmas 3–5. This completes the proof of Theorem 3. □

Noting Lemma 1, and using the same process as in the proof of Theorem 3.6 in [15], we have the following theorem.

THEOREM 4. *If* $\lambda_1 > 1$ *and* $\lambda_2 > 1$, *then the solution of system* (1) *satisfies*

$$\lim_{t\to\infty}(S,R,U,V) = (z,z,0,0).$$

If $\lambda_1 > 1$ *and* $\lambda_2 < 1$, *then the solution of system* (1) *satisfies*

$$\lim_{t\to\infty}(S,R,U,V) = (z-\beta V, z-V, 0, V).$$

If $\lambda_1 < 1$ *and* $\lambda_2 > 1$, *then the solution of system* (1) *satisfies*

$$\lim_{t\to\infty}(S,R,U,V) = (z-U, z-\alpha U, U, 0).$$

THEOREM 5. *Suppose* $\hat{\lambda}_i < 1$, $i = 1,2$. *Then the solution of system* (1) *is uniformly persistent* ([10]).

Proof. It follows from system (1′) that $v(x,t) \le V(x,t)$, where $V(x,t)$ is the solution of the problem

$$\begin{aligned} &V_t = dV_{xx} + Vg_2(\varphi_1 - \beta V, \varphi_2 - V), \quad 0 < x < 1, \quad t > 0,\\ &V_x(0,t) = 0, \quad V_x(1,t) + \gamma V(1,t) = 0,\\ &V(x,0) = v_0(x). \end{aligned}$$

Since $\hat{\lambda}_2 < 1$, then $\lambda_2 < 1$. We can proceed as in Theorem 2.5 in [32] to show that if $\lambda_2 < 1$, then $\lim_{t\to\infty} V(x,t) = V(x)$, where $V(x) < \min\{1, \frac{1}{\beta}\}z$ is the unique positive solution of

$$\begin{aligned} &dV_{xx} + Vg_2(z-\beta V, z-V) = 0, \quad 0 < x < 1,\\ &V_x(0) = 0, \;\; V_x(1) + \gamma V(1) = 0. \end{aligned}$$

Since $\hat{\lambda}_1 < 1$, we can take $0 < \epsilon \ll 1$, such that for the following principal eigenvalue $\tilde{\lambda}_1 < 1$,

$$\begin{aligned} &d\tilde{\phi}_{1xx} + \tilde{\lambda}_1\tilde{\phi}_1 g_1((1-\epsilon(1+\beta)/2)z - \beta V(x), (1-\epsilon)z - V(x)) = 0, \quad 0 < x < 1,\\ &\tilde{\phi}_{1x}(0) = 0, \quad \tilde{\phi}_{1x}(1) + \tilde{\phi}_1(1) = 0. \end{aligned}$$

There exists $\tau' > 0$ such that for $x \in [0,1]$, $t \ge \tau'$, the following inequalities hold:

$$\varphi_1 \ge z - (\epsilon/2)z, \quad \varphi_2 \ge z - (\epsilon/2)z, \quad v(x,t) \le V(x) + (\epsilon/2)z.$$

Using the comparison theorem, it follows that for $t \ge \tau'$, $u(x,t) \ge \underline{u}(x,t)$, where $\underline{u}(x,t)$ is the solution of

$$\begin{aligned} &\underline{u}_t = d\underline{u}_{xx} + \underline{u}g_1((1-(\epsilon(1+\beta)/2))z - \underline{u} - \beta V(x), \quad (1-\epsilon)z - \alpha\underline{u} - V(x)),\\ &\qquad\qquad 0 < x < 1, t > \tau',\\ &\underline{u}_x(0,t) = 0, \quad \underline{u}_x(1,t) + \gamma\underline{u}(1,t) = 0,\\ &\underline{u}(x,\tau') = \min\{(1-\epsilon(1+\beta)/2)z - \beta V(x), \quad ((1-\epsilon)z - V(x))/\alpha, u(x,\tau')\}. \end{aligned}$$

Noting $\tilde{\lambda}_1 < 1$, we have $\lim_{t\to\infty} \underline{u}(x,t) = \underline{u}_\epsilon(x)$, where $\underline{u}_\epsilon(x)$ is the unique positive solution of

$$d\underline{u}_{\epsilon xx} + \underline{u}_\epsilon g_1((1-\epsilon(1+\beta)/2)z - \underline{u}_\epsilon - \beta V(x), (1-\epsilon)z - \alpha\underline{u}_\epsilon - V(x)) = 0$$

with the usual boundary conditions. It follows from an L^p estimate and the embedding theorem (see [27]) that $\lim_{\epsilon\to 0} \underline{u}_\epsilon(x) = \hat{u}(x)$, where $\hat{u}(x)$ is the unique positive solution of the following problem on $[0,1]$:

$$d\hat{u}_{xx} + \hat{u}g_1(z - \hat{u} - \beta V(x), z - \alpha\hat{u} - V(x)) = 0 \tag{13}$$

with the usual boundary conditions. A similar result holds for v if $\hat{\lambda}_2 < 1$. Hence, there exist constants $\eta_1 > 0,\ \tau_1 \geq \tau'$ such that $u(x,t) \geq \eta_1,\ v(x,t) \geq \eta_1$ for $x \in [0,1],\ t \geq \tau_1$.

By the equation of S in system (1) and the definition of $g_i,\ i = 1,2$, we have

$$\begin{aligned} S_t &= dS_{xx} - g_1(S,R)u - \beta g_2(S,R)v \\ &\geq dS_{xx} - \max\left\{\tfrac{m_{s_1}}{K_{s_1}}, \tfrac{m_{s_2}}{K_{s_2}}\right\} S(u+\beta v). \end{aligned}$$

Then there exists $\tau'' > 0$, and for $t \geq \tau''$, the following inequality holds:

$$S_t \geq dS_{xx} - \max\left\{\frac{m_{s_1}}{K_{s_1}}, \frac{m_{s_2}}{K_{s_2}}\right\} S(z+\varepsilon - S).$$

A similar result holds for R. Thus we can proceed as in Lemma 3.8 in [15] to show that there exist $\eta_2 > 0,\ \tau_2 > 0$ such that $S(x,t) \geq \eta_2,\ R(x,t) \geq \eta_2$ for $x \in [0,1],\ t \geq \tau_2$. Denote $\tau = \max\{\tau_1, \tau_2\},\ \eta = \max\{\eta_1, \eta_2\}$. Then we have $S \geq \eta,\ R \geq \eta,\ u \geq \eta,\ v \geq \eta$ for $x \in [0,1],\ t \geq \tau$. This completes the proof. □

3. Numerical simulations. The goal of this section is to present the results of numerical simulations that complement the analytic results of the previous section. The simulations reported below represent a small fraction of those made. We wish to make a few general comments based on our observations. First, in most simulations performed, convergence to equilibrium was observed. Second, competitive exclusion, the elimination of one population by another, was observed. Finally, nonuniqueness of the positive equilibrium and bistability of the semitrivial equilibrium were observed. Our simulations are consistent with the analytic results of the previous sections. Furthermore, the simulations reveal that much more complicated dynamics are also possible in the region D defined below. Our numerical simulations also seemed to indicate that coexistence is more likely in the case of competition for two limiting complementary resources in the unstirred chemostat, than in the case of competition for a single limiting resource in the unstirred chemostat (see [26]).

Define $A = \{(\hat{\lambda}_1, \hat{\lambda}_2) : 0 < \hat{\lambda}_1 < 1,\ 0 < \hat{\lambda}_2 < 1\}$, $B = \{(\hat{\lambda}_1, \hat{\lambda}_2) : 0 < \hat{\lambda}_1 < 1,\ \hat{\lambda}_2 > 1\}$, $C = \{(\hat{\lambda}_1, \hat{\lambda}_2) : \hat{\lambda}_1 > 1,\ 0 < \hat{\lambda}_2 < 1\}$, and $D = \{(\hat{\lambda}_1, \hat{\lambda}_2) : \hat{\lambda}_1 > 1,\ \hat{\lambda}_2 > 1\}$.

Our numerical simulations seem to indicate the following:

(1) Coexistence in the form of a positive equilibrium can be observed when $(\hat{\lambda}_1, \hat{\lambda}_2) \in A \cup B \cup C$ (see Figures 1–2 and Tables 1–2), and apparently a globally stable positive equilibrium can always be observed when $(\hat{\lambda}_1, \hat{\lambda}_2) \in A$;

(2) Competitive exclusion in the form of an apparently globally stable semitrivial positive equilibrium can occur when $(\hat{\lambda}_1, \hat{\lambda}_2) \in B \cup C$ (see Figures 1–2 and Tables 1–2);

(3) Both stable and unstable positive equilibria can exist, and there can be bistability with two stable semitrivial equilibria and an unstable positive equilibrium, resulting in initial condition dependent outcomes when $(\hat{\lambda}_1, \hat{\lambda}_2) \in D$ (see Figures 3–4);

(4) Existence of multiple stable and/or unstable positive equilibria can be observed when $(\hat{\lambda}_1, \hat{\lambda}_2) \in D$ (see Figures 4–5);

(5) The parameters have an apparent effect on the density of both organisms, i.e., the density u can be nondecreasing and the density v can be nonincreasing as α increases (see Figures 6(a)–6(c)). Similar results for v and u hold as β increases. But the density of both organisms can decrease as γ increases (see Figures 6(b) and 6(d)).

Now we describe an indirect method used for determining either $\hat{\lambda}_i > 1$ or $\hat{\lambda}_i < 1$ from numerical simulations. The method will be described for determining the sign of $\hat{\lambda}_1 - 1$ only, since the other case is similar. Consider the following system:

$$
\begin{aligned}
&u_t = du_{xx} + ug_1(z - u - \beta v, z - \alpha u - v), && 0 < x < 1,\ t > 0,\\
&v_t = dv_{xx} + vg_2(z - \beta v, z - v), && 0 < x < 1,\ t > 0,\\
&u_x(0,t) = 0,\ \ u_x(1,t) + \gamma u(1,t) = 0,\\
&v_x(0,t) = 0,\ \ v_x(1,t) + \gamma v(1,t) = 0,\\
&u(x,0) = u_0(x) \geq 0,\ \ v(x,0) = v_0(x) \geq 0, \not\equiv 0,
\end{aligned}
\tag{14}
$$

where $u_0 + \beta v_0 \leq z,\ \alpha u_0 + v_0 \leq z$. Taking initial conditions characterized by a very small density of u_0, we can prove and observe numerically that v rapidly approaches the equilibrium $V(x)$. Hence for large times, t, we take $v(x,t)$ as $V(x)$ in the first equation of (14). Then we have

$$
u_t = du_{xx} + ug_1(z - u - \beta V, z - \alpha u - V), \quad 0 < x < 1,\ t > 0 \tag{15}
$$

with the usual boundary and initial conditions. We can use the comparison theorem and the Liapunov function method to prove that the solution $u(x,t)$ of (15) satisfies $\lim_{t\to\infty} u(x,t) = 0$ if $\hat{\lambda}_1 \geq 1$ and $\lim_{t\to\infty} u(x,t) = \hat{u}$ if $\hat{\lambda}_1 < 1$, where $\hat{u}$ is the unique positive solution of (13). Therefore, what happens to u depends essentially on the sign of $\hat{\lambda}_1 - 1$. If $\hat{\lambda}_1 \geq 1$, we observed the decay of the solution u of (15) to very small values; if $\hat{\lambda}_1 < 1$, we observed the growth of the solution u of (15) to the value of the solution of (13). Therefore, we can determine the sign of $\hat{\lambda}_1 - 1$ numerically by observing whether there is decay to very small values or growth to the value of the solution of (13).

We next simulate the corresponding time-dependent system of (3), which determines the limiting system of (1):

$$
\begin{aligned}
&u_t = du_{xx} + ug_1(z - u - \beta v, z - \alpha u - v), && 0 < x < 1,\ t > 0,\\
&v_t = dv_{xx} + vg_2(z - u - \beta v, z - \alpha u - v), && 0 < x < 1,\ t > 0,\\
&u_x(0,t) = 0,\ \ u_x(1,t) + \gamma u(1,t) = 0,\\
&v_x(0,t) = 0,\ \ v_x(1,t) + \gamma v(1,t) = 0,\\
&u(x,0) = u_0(x) \geq 0,\ \ v(x,0) = v_0(x) \geq 0.
\end{aligned}
$$

We have chosen to discretize the spatial variables in the above system using a second-order finite-difference scheme. The derivative terms in the boundary conditions are approximated using second-order centered differencing. The temporal variable is approximated using the Crank–Nicholson method. In all of the simulations the domain is divided uniformly into 40 cells.

TABLE 1

The equilibria corresponding to the parameters given in Tables 1 *and* 2 *are shown in Figures* 1 *and* 2.

m_s	$\hat{\lambda}_1$	$\hat{\lambda}_2$	Area	Equilibrium
$(2.06, 2.04)$	< 1	> 1	B	$(U, 0)$
$(2.045, 2.055)$	< 1	> 1	B	Coexistence
$(2.04, 2.06)$	< 1	< 1	A	Coexistence
$(1.8, 2.3)$	< 1	< 1	A	Coexistence
$(1.6, 2.5)$	< 1	< 1	A	Coexistence
$(1.55, 2.55)$	> 1	< 1	C	$(0, V)$

TABLE 2

m_s	$\hat{\lambda}_1$	$\hat{\lambda}_2$	Area	Equilibrium
$(4.55, 4.45)$	< 1	> 1	B	$(U, 0)$
$(4.48, 4.52)$	< 1	> 1	B	Coexistence
$(4.45, 4.55)$	< 1	< 1	A	Coexistence
$(2.2, 6.8)$	< 1	< 1	A	Coexistence
$(2.1, 6.9)$	> 1	< 1	C	Coexistence
$(1.95, 7.05)$	> 1	< 1	C	$(0, V)$

In the following, we denote $m_s = (m_{s_1}, m_{s_2})$, $m_r = (m_{r_1}, m_{r_2})$, $K_s = (K_{s_1}, K_{s_2})$, $K_r = (K_{r_1}, K_{r_2})$.

Coexistence and competitive exclusion. In Figures 1 and 2 a sequence of simulations is reported where different growth rates were used, but all of the other parameter values remain fixed. The parameter values used were $\alpha = \beta = 0.5, \gamma = 1, K_s = (1, 1), K_r = (1, 1.2)$. In Figure 1 and Table 1, $m_r = (3, 3)$, and in Figure 2 and Table 2, $m_r = (6, 6)$. In Figure 1, m_s took the values indicated in Table 1, and in Figure 2, m_s took the values indicated in Table 2. In each simulation in Figures 1 and 2, the densities were plotted at the final time, $t = 1000$. This appeared to be long enough to allow the solutions to be very close to steady state. A similar procedure was used in the other figures. We observed from Figures 1 and 2, as well as from many other simulations, that at the highest growth rate of u and the lowest growth rate of v, u is dominant with v barely present for any initial conditions. In this case, we checked that $(\hat{\lambda}_1, \hat{\lambda}_2) \in B$. As the growth rate of u is decreased or the growth rate of v is increased, the amount of v increases at the expense of the amount of u. Both organisms coexist at a positive equilibrium. In this case, we also checked that $(\hat{\lambda}_1, \hat{\lambda}_2) \in A$ or B or C. All the simulations show that the coexistence is unique and an apparently globally stable positive equilibrium exists when $(\hat{\lambda}_1, \hat{\lambda}_2) \in A$. As the growth rate of u is further decreased or the growth rate of v is further increased, v is dominant with u barely present for any initial condition. In addition, we checked $(\hat{\lambda}_1, \hat{\lambda}_2) \in C$ in this case. Coexistence in the form of a positive equilibrium can occur when $(\hat{\lambda}_1, \hat{\lambda}_2) \in A$ or B or C. The nonexistence of a positive equilibrium can also occur when $(\hat{\lambda}_1, \hat{\lambda}_2) \in B$ or C.

Bistability and the existence of positive equilibria. (i) In Figure 3, we provide numerical evidence of bistability; i.e., each of the two semitrivial equilibria is stable to invasion by its rival and attracts solutions corresponding to nearby initial data. As well, an unstable positive equilibrium is observed. We took $m_s = (3, 2)$, $m_r = (2.4, 3.6)$, and the other parameter values as in Figure 1. In this case we checked that $(\hat{\lambda}_1, \hat{\lambda}_2) \in D$. The simulations in Figures 3(a) and 3(b) show a plot of the L_1 norms of u and of v versus time t. In Figure 3(a) the initial conditions used were $u_0 = 0.5$ and

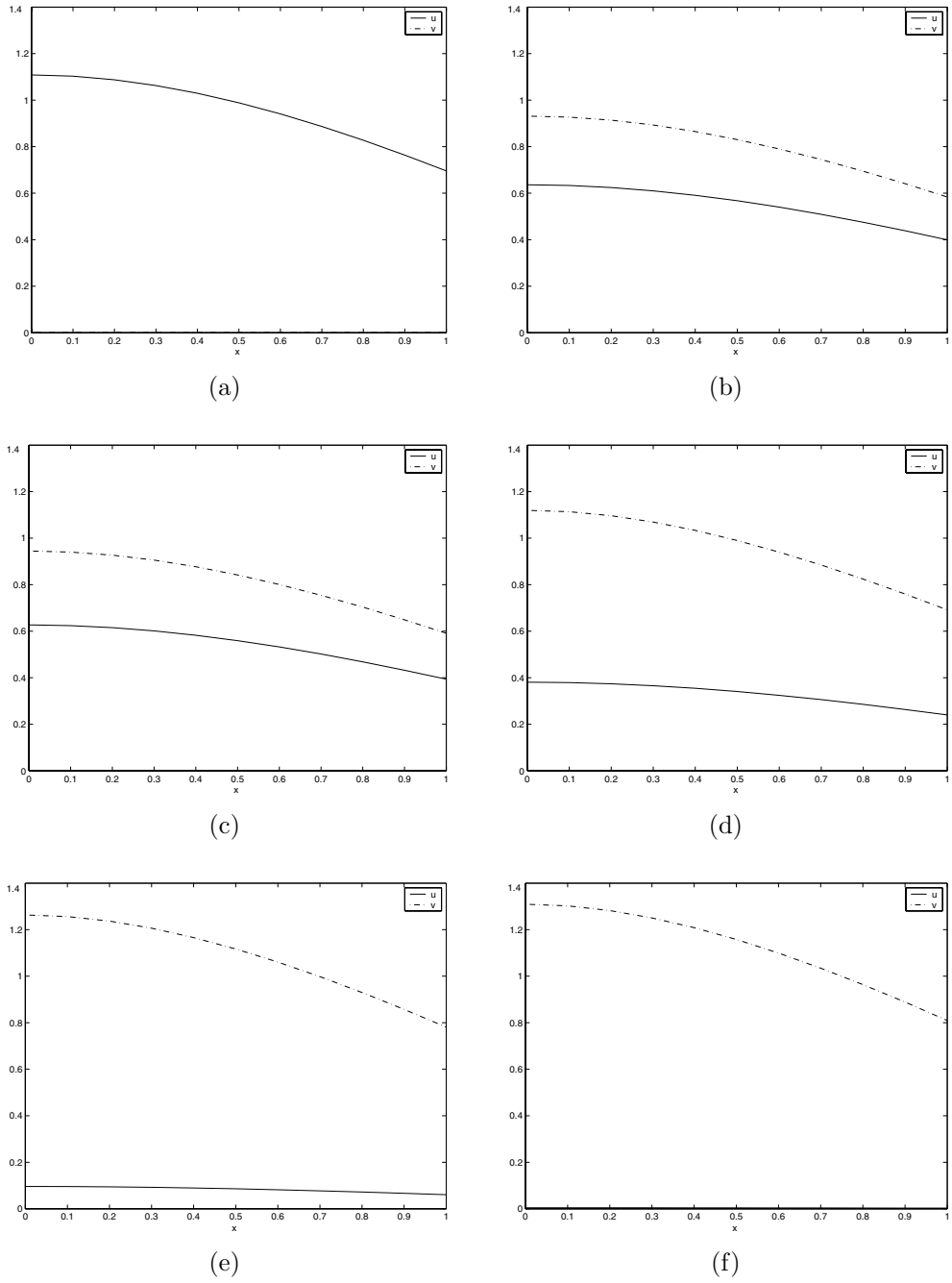


FIG. 1. *Equilibria for* $m_r = (3, 3)$ *and the different values of* m_s *from Table* 1 *(in the order given in that table). The other parameters used are* $K_s = (1, 1)$, $K_r = (1, 1.2)$, $\alpha = \beta = 0.5$, *and* $\gamma = 1$.

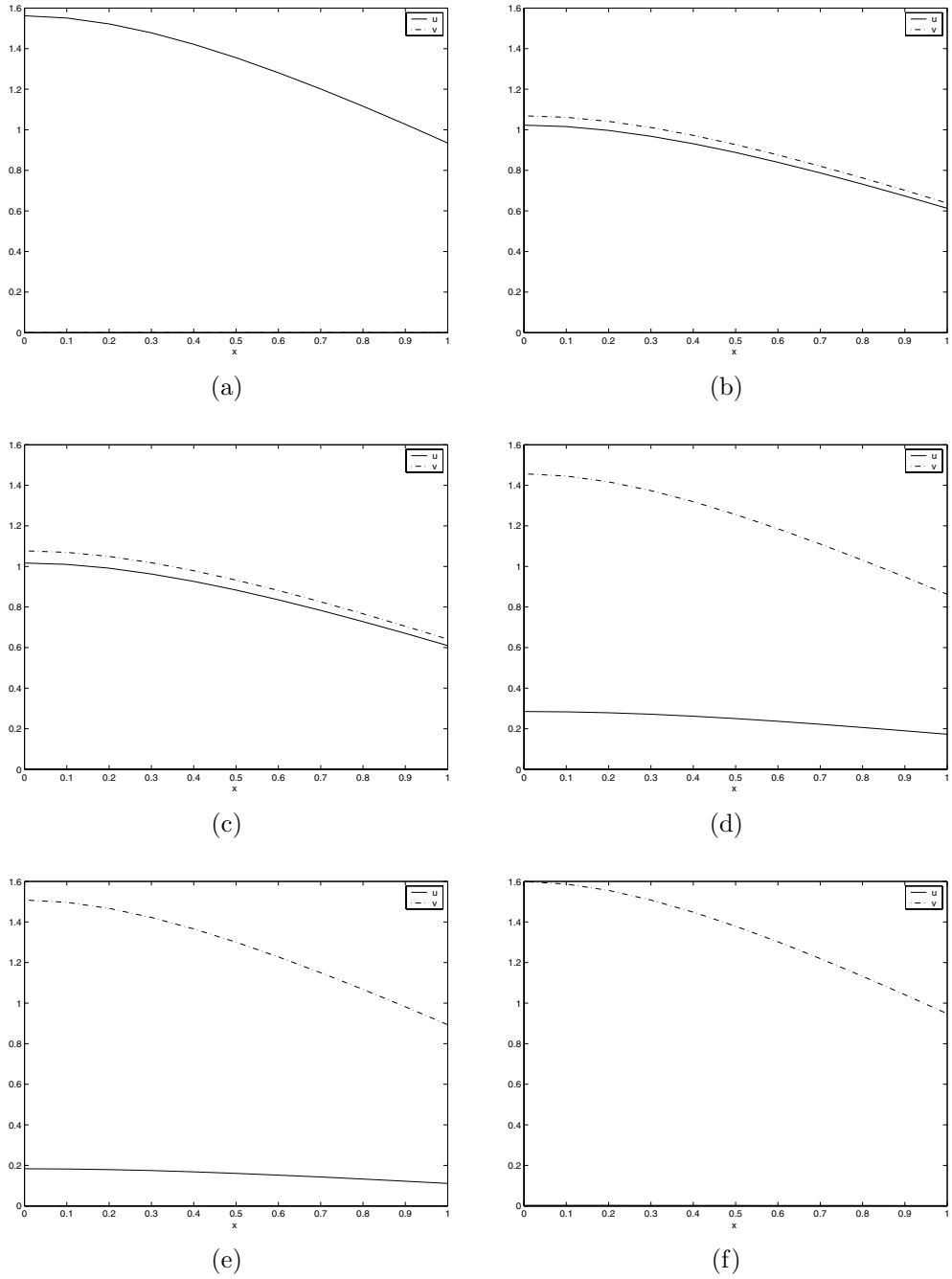


FIG. 2. *Equilibria for* $m_r = (6, 6)$ *and the different values of* m_s *from Table* 2 *(in the order given in that table). All the other parameters are the same as those in Figure* 1.

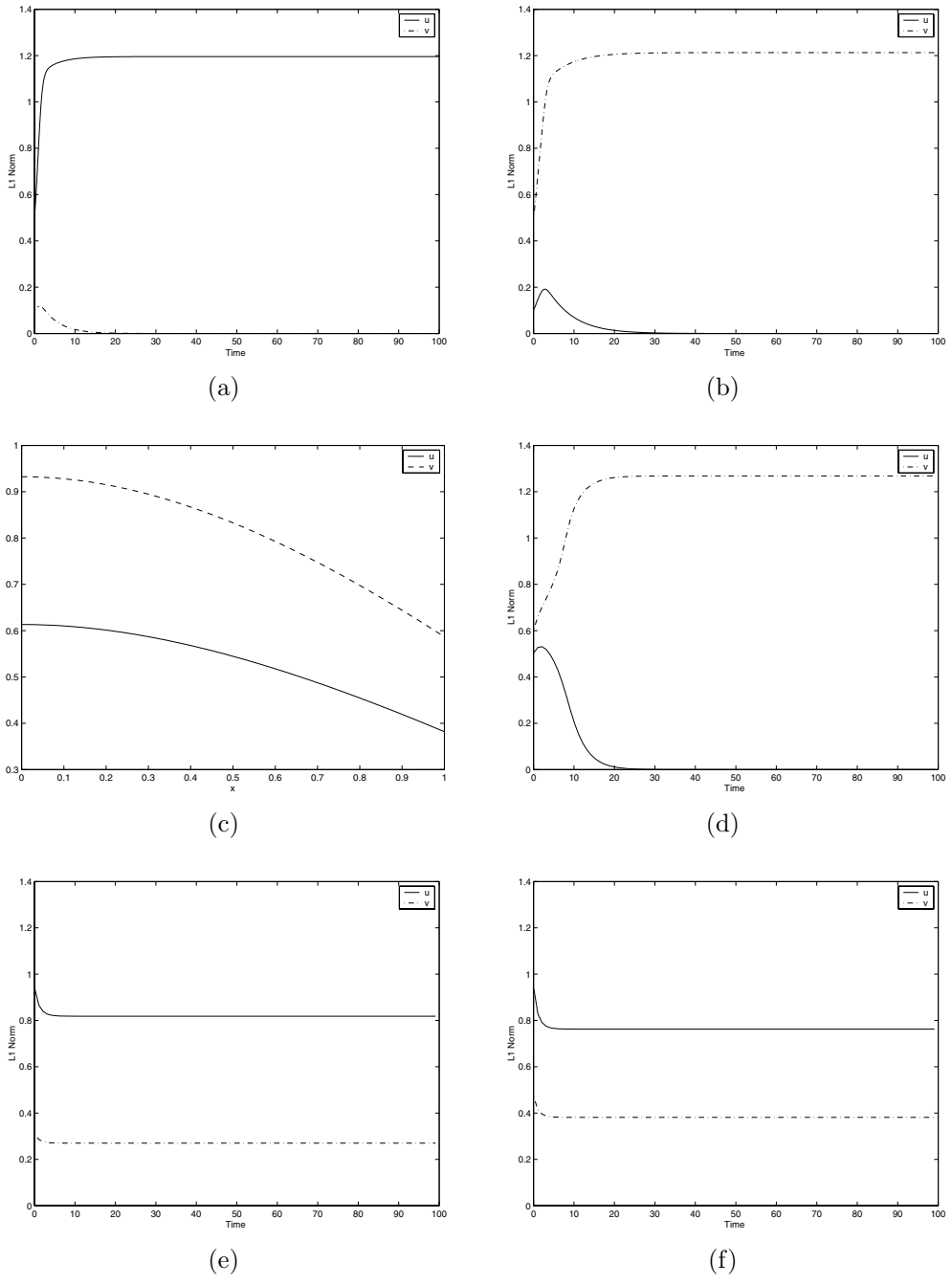


FIG. 3. *Convergence to equilibria. All parameters except m_s and m_r are the same as those in Figure* 1. *In* (a)–(c) $m_s = (3, 2)$ *and* $m_r = (2.4, 3.6)$. *In* (a)–(b) *the* L_1 *norms of* u *and of* v *versus time are shown for two semitrivial equilibria. In* (c) *a plot of the positive equilibrium for each* $x \in [0, 1]$ *is shown. In* (d)–(f) $m_s = (2, 2)$ *and* $m_r = (2, 4)$. *The* L_1 *norms of* u *and of* v *versus time are shown for several different equilibria.*

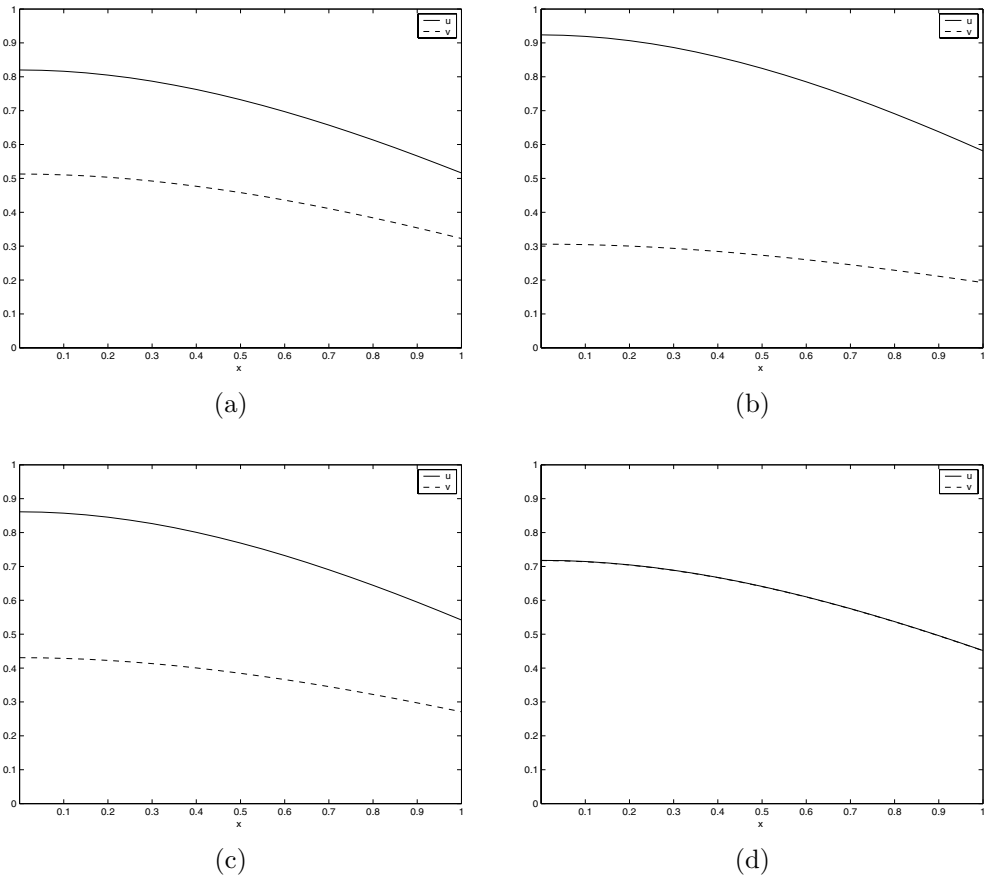


FIG. 4. *Several positive equilibria. All parameters are the same as those in Figure* 1*, except* $m_s = (2, 2)$ *and* $m_r = (2, 4)$*. Note that* u *and* v *are indistinguishable in* (d).

$v_0 = 0.1$. In Figure 3(b) the initial conditions used were $u_0 = 0.1$ and $v_0 = 0.5$. The positive equilibrium is plotted in Figure 3(c). After many simulations, in this case we believe that this is the only positive equilibrium and that it is unstable.

(ii) In Figures 3(d)–(f), we took $m_s = (2, 2)$ and $m_r = (2, 4)$. All other parameters are the same as in Figure 1. Both semitrivial equilibria are stable. Only one of them is shown (see Figure 3(d)). As well, nonuniqueness and stability of more than one positive equilibrium are observed. In this case, we checked that $(\hat{\lambda}_1, \hat{\lambda}_2) \in D$.

Existence of multiple positive equilibria. Based on extensive simulations, we believe that much more complicated dynamical behavior can occur when $(\hat{\lambda}_1, \hat{\lambda}_2) \in D$.

(i) In Figure 4 we took the same parameters as in Figures 3(d)–(f) and used continuation (numerical analysis) to find the equilibria. Simulations (not shown) seem to indicate that there are at least four positive equilibria in this case, and strongly suggest that one of them is unstable (see Figure 4(d), where u and v are indistinguishable), and that the other three are stable (see Figures 4(a)–(c)). Note that the equilibria depicted in Figures 4(b)–(c) correspond to those in Figures 3(e)–(f), respectively.

(ii) In Figures 5(a)–(c) we took $m_s = (1.5, 1.8)$ and $m_r = (1.75, 1.42)$. In Figures 5(d)–(f) we took $m_s = (2.1, 2.75)$ and $m_r = (2.8, 2.13)$. The other parameters

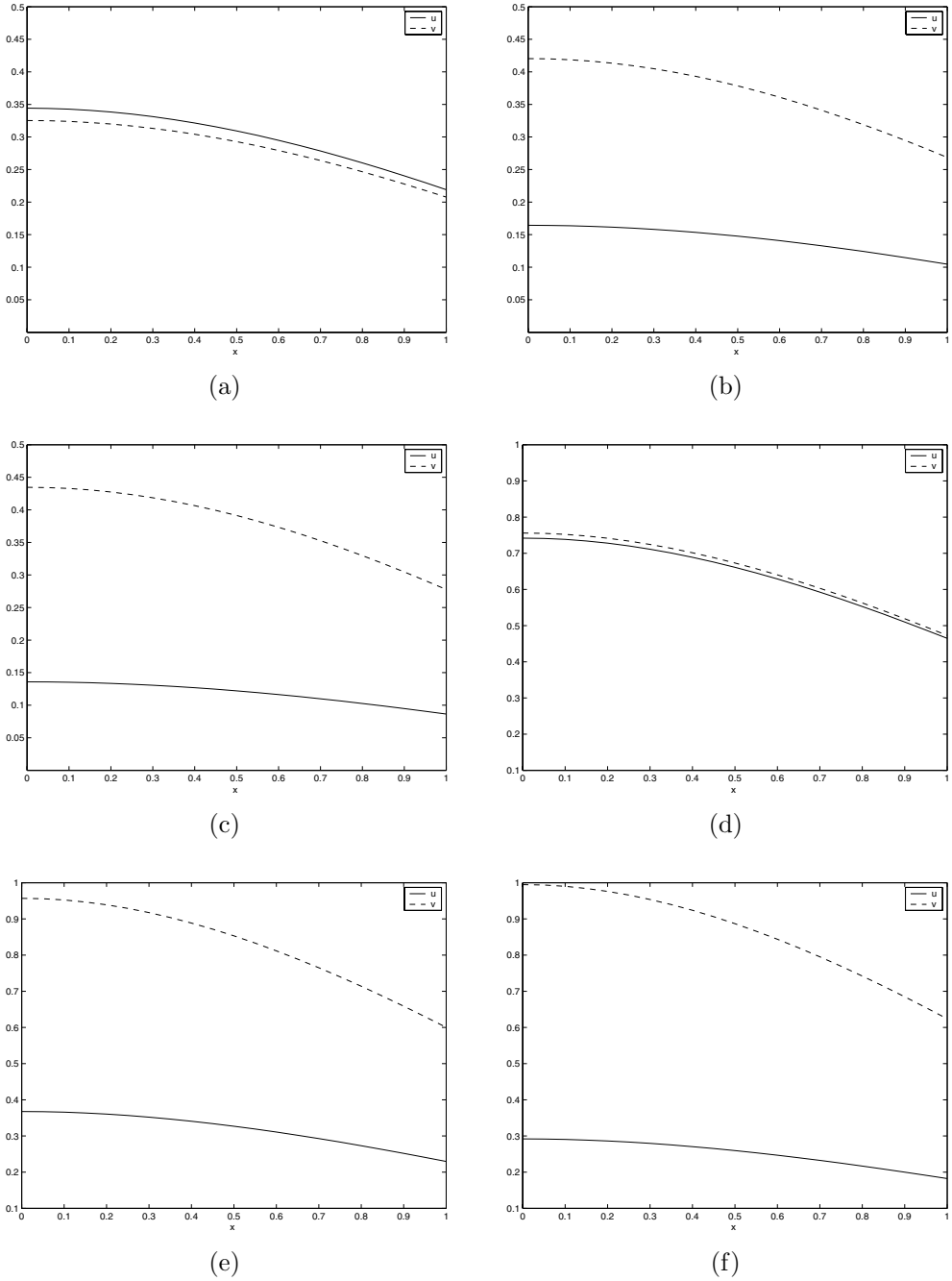


FIG. 5. *Positive equilibria. All parameters are the same as those in Figure* 1*, except that in* (a)–(c) $m_s = (1.5, 1.8)$ *and* $m_r = (1.75, 1.42)$*, and in* (d)–(f) $m_s = (2.1, 2.75)$ *and* $m_r = (2.8, 2.13)$.

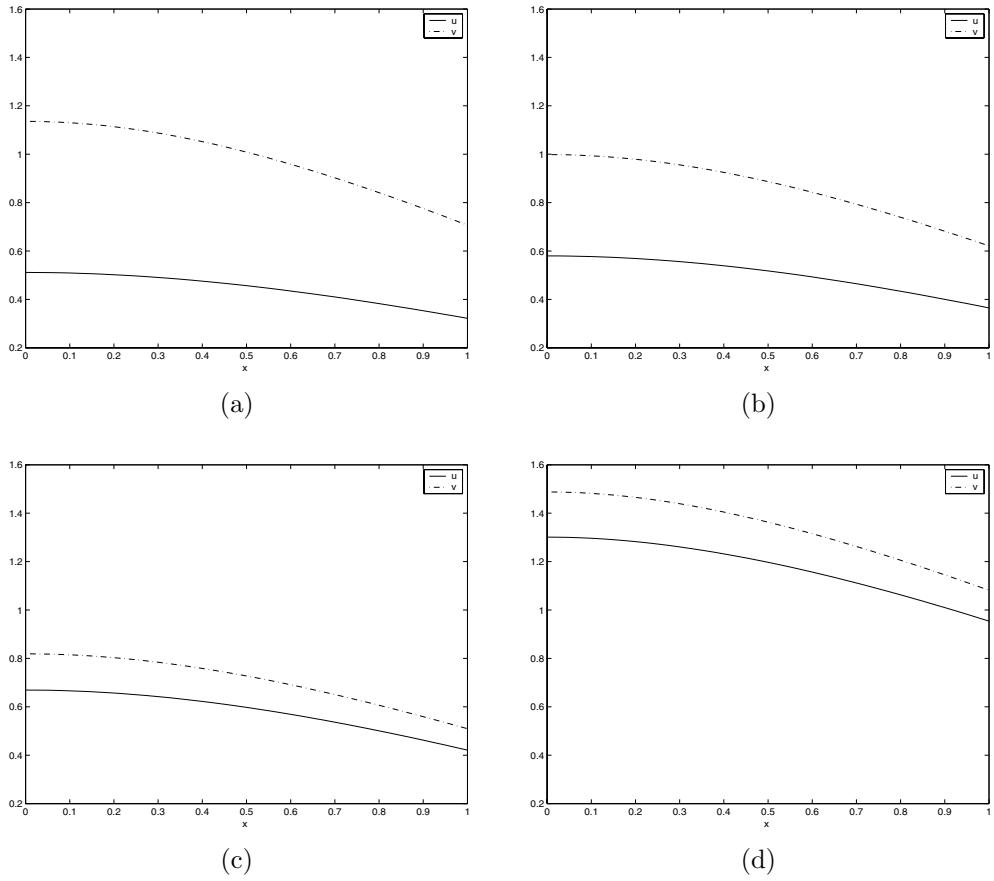


FIG. 6. *Parameters α and γ have an effect on the densities of the equilibria. In* (a)–(d) $m_s = (2, 2.1)$, $m_r = (3, 3)$, $K_s = (1, 1)$, $K_r = (1, 1.2)$, *and* $\beta = 0.5$. *In* (a)–(c) $\gamma = 1$, *and in* (d) $\gamma = 0.6$. *In* (a) $\alpha = 0.3$, *in* (b) $\alpha = 0.5$, *in* (c) $\alpha = 0.7$, *and in* (d) $\alpha = 0.5$.

were taken as in Figure 1. In both cases, we checked that $(\hat{\lambda}_1, \hat{\lambda}_2) \in D$. For each case, we used numerical analysis to find the three positive equilibria depicted in Figure 5. Subsequent simulations strongly indicated that all these positive equilibria are unstable.

Effects of the parameters. In Figure 6 a sequence of simulations shows that the parameters α, β, γ have an apparent effect on the density of both populations. Parameter values taken are $m_s = (2, 2.1)$, $m_r = (3, 3)$, $K_s = (1, 1)$, $K_r = (1, 1.2)$. Values for α, β, and γ are given in the caption of Figure 6. The initial data are $u_0 = 1$ and $v_0 = 1$. We observed that the density of u can be nondecreasing and the density of v can be nonincreasing as α increases (see Figures 6(a)–6(c)). A similar result holds for v and u as β increases. We also observed that the density of both u and v can decrease as γ increases (see Figures 6(b) and 6(d)).

Acknowledgment. The authors would like to express their sincere thanks to the anonymous referees of this paper for their careful reading and valuable suggestions leading to an improvement of the paper.

REFERENCES

[1] H. Amann, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620–709.

[2] M. M. Ballyk and G. S. K. Wolkowicz, *An examination of the thresholds of enrichment: A resource-based growth model*, J. Math. Biol., 33 (1995), pp. 435–457.

[3] M. M. Ballyk and G. S. K. Wolkowicz, *Exploitative competition in the chemostat for two perfectly substitutable resources*, Math. Biosci., 118 (1993), pp. 127–180.

[4] B. C. Baltzis and A. G. Fredrickson, *Limitation of growth rate by two complementary nutrients: Some elementary and neglected considerations*, Biotechnol. Bioeng., 31 (1988), pp. 75–86.

[5] J. V. Baxley and H. B. Thompson, *Nonlinear boundary value problems and competition in the chemostat*, Nonlinear Anal., 22 (1994), pp. 1329–1344.

[6] G. J. Butler and G. S. K. Wolkowicz, *A mathematical model of the chemostat with a general class of functions describing nutrient uptake*, SIAM J. Appl. Math., 45 (1985), pp. 138–151.

[7] G. J. Butler and G. S. K. Wolkowicz, *Exploitative competition in a chemostat for two complementary, and possibly inhibitory, resources*, Math. Biosci., 83 (1987), pp. 1–48.

[8] L. Dung and H. L. Smith, *A parabolic system modelling microbial competition in an unmixed bio-reactor*, J. Differential Equations, 130 (1996), pp. 59–91.

[9] A. G. Fredrickson and G. Stephanopoulos, *Microbial competition*, Science, 213 (1981), pp. 972–979.

[10] J. Hale and P. Waltman, *Persistence in infinite-dimensional systems*, SIAM J. Math. Anal., 20 (1989), pp. 388–395.

[11] S. R. Hansen and S. P. Hubbell, *Single nutrient microbial competition: Agreement between experimental and theoretical forecast outcomes*, Science, 207 (1980), pp. 1491–1493.

[12] W. Harder and L. Dijkhuizen, *Strategies of mixed substrate utilization in microorganisms*, Philos. Trans. R. Soc. London B., 297 (1982), pp. 459–480.

[13] S.-B. Hsu, K.-S. Cheng, and S. P. Hubbell, *Exploitative competition of microorganisms for two complementary nutrients in continuous cultures*, SIAM J. Appl. Math., 41 (1981), pp. 422–444.

[14] S. B. Hsu, S. Hubbell, and P. Waltman, *A mathematical theory for single-nutrient competition in continuous cultures of micro-organisms*, SIAM J. Appl. Math., 32 (1977), pp. 366–383.

[15] S.-B. Hsu and P. Waltman, *On a system of reaction-diffusion equations arising from competition in an unstirred chemostat*, SIAM J. Appl. Math., 53 (1993), pp. 1026–1044.

[16] S. B. Hsu, H. L. Smith, and P. Waltman, *Dynamics of competition in the unstirred chemostat*, Canad. Appl. Math. Quart., 2 (1994), pp. 461–483.

[17] J. A. Leon and D. B. Tumpson, *Competition between two species for two complementary or substitutable resources*, J. Theoret. Biol., 50 (1975), pp. 185–201.

[18] B. Li and H. Smith, *How many species can two essential resources support?*, SIAM J. Appl. Math., 62 (2001), pp. 336–366.

[19] B. Li, G. S. K. Wolkowicz, and Y. Kuang, *Global asymptotic behavior of a chemostat model with two perfectly complementary resources and distributed delay*, SIAM J. Appl. Math., 60 (2000), pp. 2058–2086.

[20] S. S. Pilyugin and P. Waltman, *Competition in the unstirred chemostat with periodic input and washout*, SIAM J. Appl. Math., 59 (1999), pp. 1157–1177.

[21] M. H. Protter and H. F. Weinberger, *Maximum Principle in Differential Equations*, Springer-Verlag, New York, 1984.

[22] D. J. Rapport, *An optimization model of food selection*, Am. Nat., 105 (1971), pp. 575–587.

[23] H. L. Smith and P. Waltman, *Competition in an unstirred multi-dimensional chemostat*, in Differential Equations and Applications to Biology and to Industry (Claremont, CA, 1994), M. Martelli et al., eds., World Scientific, River Edge, NJ, 1996, pp. 475–486.

[24] H. L. Smith, *An application of monotone systems theory to a model of microbial competition*, in Differential Equations and Control Theory (Wuhan, 1994), Lecture Notes in Pure and Appl. Math. 176, Z. Deng et al., eds., Marcel Dekker, New York, pp. 293–307.

[25] H. L. Smith and P. Waltman, *The Theory of the Chemostat*, Cambridge University Press, Cambridge, UK, 1995.

[26] J. W.-H. So and P. Waltman, *A nonlinear boundary value problem arising from competition in the chemostat*, Appl. Math. Comput., 32 (1989), pp. 169–183.

[27] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

[28] D. Tilman, *Resource Competition and Community Structure*, Princeton University Press, Princeton, NJ, 1982.

[29] P. Waltman, S. P. Hubbell, and S. B. Hsu, *Theoretical and experimental investigation of microbial competition in continuous culture*, in Modelling and Differential Equations, T. Burton, ed., Marcel Dekker, New York, 1980.

[30] J. H. Wu, *Stability of steady-state solutions of the competition model in the chemostat*, Systems Sci. Math. Sci., 7 (1994), pp. 256–260.

[31] J. H. Wu, *Global bifurcation of coexistence state for the competition model in the chemostat*, Nonlinear Anal., 39 (2000), pp. 817–835.

[32] J. H. Wu, and G. S. K. Wolkowicz, *A system of resource-based growth models with two resources in the unstirred chemostat*, J. Differential Equations, 172 (2001), pp. 300–332.

[33] E. Zeidler, *Nonlinear Functional Analysis and its Application* I: *Fixed-Point Theorems*, Springer-Verlag, New York, 1985.

SIAM J. APPL. MATH.
Vol. 65, No. 1, pp. 230–251

INSTABILITIES OF WAVY PATTERNS GOVERNED BY COUPLED BURGERS EQUATIONS*

E. A. GLASMAN†, A. A. GOLOVIN‡, AND A. A. NEPOMNYASHCHY§

Abstract. Wave dynamics in a system of coupled Burgers equations is studied. This model describes the long-wave nonlinear evolution of an oscillatory pattern-forming system in the presence of the Goldstone mode caused by the translation symmetry, for example, the oscillatory instability of a propagating combustion front. It is shown that the system of coupled Burgers equations reveals several new types of instabilities, which are studied both analytically and numerically. In some limiting cases, secondary amplitude equations governing these instabilities are derived. The nonlinear development of these instabilities is studied by numerical simulations.

Key words. waves, Burgers equation, instability, combustion

AMS subject classifications. 70K50, 34A34, 74J99, 35Q53, 80A25

DOI. 10.1137/S0036139903432913

1. Introduction.

In the last decades, the spontaneous formation of spatially inhomogeneous patterns was a subject of extensive investigations [1]. A deep understanding of various phenomena was achieved by studying nonlinear *amplitude equations* valid near the onset of instability. The role of these equations in the description of bifurcations in distributed systems is similar to the role of normal forms in the description of bifurcations in systems with a finite number of degrees of freedom. In the case of *short-wavelength instabilities* (i.e., those occurring with a nonzero wavenumber at the threshold), a complex Ginzburg–Landau (CGL) equation [2, 3, 4, 5] turns out to be a generic equation for an envelope function governing the evolution of patterns near the instability threshold. In the case of *long-wavelength instabilities* (i.e., those occurring with a zero wavenumber at the threshold) the situation is much more intricate. Depending on the features of the linear dispersion relation and the symmetries of the problem, the nonlinear dynamics of patterns is governed by different kinds of amplitude equations [6]. Among them are the CGL equation [7], the Kuramoto–Sivashinsky equation [8, 9, 10, 11], the perturbed Korteweg–de Vries equation [12, 13, 14], and others.

A basic problem, which is typically solved by means of amplitude equations, is the stability of spatially periodic waves generated by the primary instability. In many cases, there exists a *stability interval* of periodic waves ("Busse balloon"). Specifically, such an interval was revealed for wavy patterns governed by the Ginzburg–Landau equation [3, 15], the Kuramoto–Sivashinsky equation [8, 16, 17], and the perturbed Korteweg–de Vries equation [18, 19]. The boundaries of this interval are determined by long-wavelength *modulational instabilities*. The growth rate σ of these instabilities usually depends on the perturbation wavenumber k as $\sigma = O(k^2)$ ("negative

*Received by the editors August 4, 2003; accepted for publication (in revised form) March 2, 2004; published electronically October 8, 2004. This work was supported by the U.S.–Israel Binational Science Foundation, grant 9800086.

http://www.siam.org/journals/siap/65-1/43291.html

†Department of Mathematics, Technion-Israel Institute of Technology, Haifa 32000, Israel (jenny@gugelglasman.com).

‡Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL 60208 (a-golovin@northwestern.edu).

§Minerva Center for Nonlinear Physics of Complex Systems, Technion-Israel Institute of Technology, Haifa 32000, Israel (nepom@math.technion.ac.il).

viscosity"), but sometimes as $\sigma = O(k)$ (see [6, 17, 19]). The nonlinear development of modulational instabilities is governed by secondary amplitude equations (see [4, 10, 14, 20]). It leads either to a change of the pattern wavelength or to the appearance of more complicated, e.g., spatially quasi-periodic, or chaotic, dynamics.

Recently, a new class of pattern-forming systems started attracting considerable attention. The dynamics of these systems is characterized by the nonlinear interaction of an unstable mode coupled with a *stable* but slow Goldstone mode (see [21, 22] for review) associated with the symmetry of the problem or a conservation law. An interesting example of such a system in which the long-wave oscillatory mode is coupled to the slow Goldstone mode caused by the translation symmetry is a uniformly propagating planar flame front governed by sequential reactions [23, 24]. The nonlinear dynamics of this system near the instability threshold is governed by two coupled equations: a CGL equation for the frontal oscillation amplitude P and a Burgers equation for the frontal deformation Q [23],

$$\begin{aligned} \partial_t P + \nabla P \cdot \nabla Q &= P + (1+iu)\nabla^2 P - (1+iv)|P|^2 P - sP\nabla^2 Q, \\ \partial_t Q &= m\nabla^2 Q - \frac{1}{2}|\nabla Q|^2 - w|P|^2, \end{aligned} \tag{1.1}$$

where $s = s_r + is_i$ is a complex constant, and u, v, m, and w are real constants. The parameter m is assumed to be positive; otherwise the deformational mode is intrinsically unstable and should be described by the Kuramoto–Sivashinsky equation rather than by the Burgers equation [23]. Although the system of equations (1.1) was derived for a particular physical system it should be considered as *generic* for a class of systems that have left-right symmetry and the same linear dispersion relation and are characterized by the presence of the Goldstone mode associated with the translation symmetry. Evolution equations for the interaction between an unstable mode and the zero mode in systems with other symmetries were studied, for example, in [21, 22, 25].

In [23], the stability of spatially uniform pulsations of the front was investigated, and some new types of nonlinear dynamics were revealed, including modulated standing waves, blinking states, and intermittent states. In [24] it was found that the long-wave instabilities of spatially periodic patterns, essentially different from the usual Benjamin–Feir instability, are of major importance. They can produce either modulated waves or oscillating cells separated by domain walls.

In the present paper we investigate long waves governed by the system (1.1). These waves are generated by the long-wavelength instabilities caused by the coupling between the phase $\theta = \arg(P)$ and the slow mode Q. We will show that in the long-wave limit, the problem is governed, to the leading order, by two *coupled Burgers equations* and has an additional symmetry with respect to the scaling transformation. A similar approach was used in [26] for the investigation of a secondary Hopf bifurcation of a stationary periodic structure possessing a translation symmetry. Note that several other systems of coupled Burgers-type equations were studied in different contexts in [27, 28, 29, 30, 31, 32]. Linear stability analysis presented below reveals a new type of modulational instability of traveling waves, which does not have its analogy for homogeneous oscillations. It leads to an unusual situation when the Busse balloon includes the single wavenumber $k_0 = 0$ rather than an interval of wavenumbers. Depending on the parameters, this instability is characterized by the growth rate dependence $\sigma = O(k^2)$ or $\sigma = O(k)$. In some limiting cases, we derive secondary amplitude equations governing both types of instabilities. The nonlinear development of these instabilities is studied by numerical simulations.

2. Instabilities of patterns governed by coupled Burgers equations.

2.1. Derivation of the system of coupled Burgers equations. Consider *large-scale* solutions of the system (1.1). Define $P = \rho \exp(i\theta)$ and rewrite the system (1.1) in the form

$$\rho_t + \nabla\rho\cdot\nabla Q = \nabla^2\rho - \rho|\nabla\theta|^2 - u(2\nabla\rho\cdot\nabla\theta + \rho\nabla^2\theta) + \rho(1 - \rho^2 - s_r\nabla^2 Q), \tag{2.1}$$

$$\rho\theta_t + \rho\nabla\theta\cdot\nabla Q = u(\nabla^2\rho - \rho|\nabla\theta|^2) + 2\nabla\rho\cdot\nabla\theta + \rho\nabla^2\theta - \rho(v\rho^2 + s_i\nabla^2 Q), \tag{2.2}$$

$$Q_t = m\nabla^2 Q - \frac{1}{2}|\nabla Q|^2 - w\rho^2. \tag{2.3}$$

The system (2.1)–(2.3) has the solution

$$\rho = 1, \qquad \theta = -vt, \qquad Q = -wt, \tag{2.4}$$

which corresponds to a planar front moving with the velocity $-w$ and pulsating with the frequency v.

In the case of a pure CGL equation, it is known that the behavior of long-wave solutions is of major importance. They are responsible for the Benjamin–Feir instability of homogeneous oscillations [3, 4, 7] that leads to the development of spatiotemporal chaos ("phase turbulence") [10]. The goal of the present section is to analyze the behavior of long waves governed by the system (2.1)–(2.3).

Introduce a formal small parameter $\epsilon \ll 1$, and assume that the characteristic wavenumbers in the Fourier expansion of the solution are small, $k = O(\epsilon)$, i.e., the solution slowly depends on the spatial coordinates. Also, we assume that the solution slowly evolves in time. Define new variables, $\mathbf{X} = \epsilon\mathbf{x}$, $T = \epsilon^2 t$, so that $\partial_t \to \partial_t + \epsilon^2\partial_T$, $\nabla \to \epsilon\nabla_{\mathbf{X}}$, and consider

$$\rho = \rho_0 + \epsilon^2\tilde\rho(\mathbf{X}, T), \qquad \theta = -\omega_0 t + \tilde\theta(\mathbf{X}, T), \qquad Q = c_0 t + \tilde Q(\mathbf{X}, T), \tag{2.5}$$

where the functions $\tilde\rho$, $\tilde\theta$, and $\tilde Q$ are expanded in series as

$$\tilde\rho = \tilde\rho_0 + \epsilon^2\tilde\rho_2 + \ldots, \qquad \tilde\theta = \tilde\theta_0 + \epsilon^2\tilde\theta_2 + \ldots, \qquad \tilde Q = \tilde Q_0 + \epsilon^2\tilde Q_2 + \ldots. \tag{2.6}$$

Substitute (2.5)–(2.6), together with (2.4) into (2.1)–(2.3) and equate the terms of the same order in ϵ. In the leading order one obtains $\rho_0 = 1$, $\omega_0 = v$, $c_0 = -w$, which coincides with (2.4). In the next order, one obtains

$$(\nabla\tilde\theta)^2 + u\nabla^2\tilde\theta + 2\tilde\rho_0 + s_r\nabla^2\tilde Q = 0, \tag{2.7}$$

$$-\omega_0\tilde\rho_0 + \tilde\theta_T + \nabla\tilde\theta\cdot\nabla\tilde Q = -u(\nabla\tilde\theta)^2 + \nabla^2\tilde\theta - (2v\tilde\rho_0 + s_i\nabla^2\tilde Q) - v\tilde\rho_0\,, \tag{2.8}$$

$$\tilde Q_T = m\nabla^2\tilde Q - \frac{1}{2}(\nabla\tilde Q)^2 - 2w\tilde\rho_0\,, \tag{2.9}$$

where the subscripts are omitted in $\tilde\theta_0$, $\tilde Q_0$, and $\nabla_{\mathbf{X}}$. From (2.7) one gets

$$\tilde\rho_0 = -\frac{1}{2}\left[(\nabla\tilde\theta)^2 + u\nabla^2\tilde\theta + s_r\nabla^2\tilde Q\right]. \tag{2.10}$$

Substitute (2.10) into (2.8), (2.9) to obtain the *coupled Burgers equations* for $\tilde\theta$ and $\tilde Q$:

$$\begin{aligned} \tilde\theta_T &= A\nabla^2\tilde\theta + C(\nabla\tilde\theta)^2 + B\nabla^2\tilde Q - \nabla\tilde\theta\cdot\nabla\tilde Q, \\ \tilde Q_T &= D\nabla^2\tilde Q - \frac{1}{2}(\nabla\tilde Q)^2 + w(\nabla\tilde\theta)^2 + E\nabla^2\tilde\theta, \end{aligned} \tag{2.11}$$

where

$$A = 1 + uv, \quad B = vs_r - s_i, \quad C = (v - u), \quad D = m + ws_r, \quad E = wu, \tag{2.12}$$

and the subscripts have been dropped.

The system (2.11) describes the interaction between long-wave phase disturbances and long-wave modulations of the zero mode; in the case of a flame instability the latter corresponds to the large-scale deformations of the flame front. Note that for $w = 0$ the equation for Q in the system (2.11) is decoupled from the equation for the phase disturbances. It is important that the system (2.11), unlike the original system (1.1), is invariant with respect to (the group of) scaling transformations $\mathbf{X} \to \mathbf{X}' = \alpha\mathbf{X}$, $T \to T' = \alpha^2 T$. Note also the specific nonlinear coupling in the system (2.11); it is related to the particular form of the system (1.1). In other systems the coupling can be different [27, 28, 29, 30, 31, 32].

2.2. Linear stability of traveling waves. The system (2.11) has a class of traveling wave solutions,

$$\tilde{\theta}_0 = \mathbf{F} \cdot \mathbf{X} - \omega T, \qquad \tilde{Q}_0 = \mathbf{G} \cdot \mathbf{X} - cT, \tag{2.13}$$

corresponding to planar waves ($\mathbf{F} \neq 0$) propagating on the background of an inclined front ($\mathbf{G} \neq 0$), where $\omega = \mathbf{F} \cdot \mathbf{G} - CF^2$, $c = G^2/2 - wF^2$. In this section we study the instabilities of these solutions in the framework of the system (2.11).

Consider the perturbed solution, $\tilde{\theta} = \tilde{\theta}_0 + \widehat{\theta}$, $\tilde{Q} = \tilde{Q}_0 + \widehat{Q}$, with $(\hat{\theta}, \hat{Q}) = (\hat{A}, \hat{B})\, e^{i\mathbf{k}\cdot\mathbf{X} + \sigma T}$, and linearize the system (2.11) to obtain the dispersion relation

$$\begin{vmatrix} \sigma + Ak^2 - i(2C\mathbf{k}\cdot\mathbf{F} - \mathbf{k}\cdot\mathbf{G}) & Bk^2 + i\mathbf{k}\cdot\mathbf{F} \\ Ek^2 - i2w\mathbf{k}\cdot\mathbf{F} & \sigma + Dk^2 + i\mathbf{k}\cdot\mathbf{G} \end{vmatrix} = 0. \tag{2.14}$$

For homogeneous frontal oscillations, $\mathbf{F} = \mathbf{G} = 0$, one finds $\sigma = \Sigma k^2$ [23], where Σ satisfies the quadratic equation $\Sigma^2 + \mathrm{tr}(M) \cdot \Sigma + \det(M) = 0$, with $\mathrm{tr}(M) = 1 + uv + m + ws_r$ and $\det(M) = m(1 + uv) + w(s_r + us_i)$ being the trace and the determinant of the matrix

$$M = \begin{pmatrix} A & B \\ E & D \end{pmatrix}. \tag{2.15}$$

Spatially homogeneous oscillations of the planar front, corresponding to $\mathbf{F} = \mathbf{G} = 0$, are monotonically unstable ($\mathrm{Im}\Sigma = 0, \mathrm{Re}\Sigma > 0$) if $\det(M) < 0$, and oscillatory unstable ($\mathrm{Im}\Sigma \neq 0$, $\mathrm{Re}\Sigma > 0$) if $\mathrm{tr}(M) < 0$, $\det(M) > [\mathrm{tr}(M)]^2/4$. Both the *monotonic instability boundary*,

$$1 + uv = -\frac{w}{m}(s_r + us_i), \tag{2.16}$$

and the *oscillatory instability boundary*,

$$1 + uv = -(m + ws_r), \qquad (m + ws_r) < \frac{w}{m}(s_r + us_i), \tag{2.17}$$

can be considered as modifications of the Benjamin–Feir instability boundary, $1 + uv = 0$, of the CGL equation, caused by the coupling with the Burgers equation. The homogeneous oscillations are stable with respect to long-wave perturbations if $\mathrm{tr}(M) > 0$, $\det(M) > 0$, i.e., for

$$1 + uv > \max\left\{-(m + ws_r), \qquad -\frac{w}{m}(s_r + us_i)\right\}. \tag{2.18}$$

Let us emphasize that the nonlinear development of the Benjamin–Feir instability *cannot* be studied in the framework of the system (2.11) because the latter has no short-wave cut-off and becomes ill-posed when the Benjamin–Feir instability appears. To describe the saturation of the Benjamin–Feir instability, the terms with fourth-order derivatives are necessary [7, 10]. Later on, we shall assume that there is no instability of the homogeneous oscillations, i.e., $\det(M) > 0$, $\mathrm{tr}(M) > 0$.

Consider now the stability of traveling wave solutions with $\mathbf{F} \neq 0$, which is determined by the dispersion relation (2.14). Since $\sigma(\mathbf{k}; \mathbf{F}, \mathbf{G}) = \sigma(\mathbf{k}; \mathbf{F}, 0) - i\mathbf{k}\cdot\mathbf{G}$, the stability of traveling waves is not affected by the inclination of the front characterized by the vector $\mathbf{G}$, and we shall further consider $\mathbf{G} = 0$. In this case, the dispersion relation (2.14) can be written as

$$\Sigma^2 + [\mathrm{tr}(M) - 2iC\alpha]\Sigma + \det(M) - i\alpha\zeta - 2w\alpha^2 = 0, \tag{2.19}$$

where $\Sigma \equiv \sigma/k^2$, $\alpha \equiv \mathbf{k}\cdot\mathbf{F}/k^2$, $\zeta = 2m(v-u) + w[2s_i + u(1-2s_r)]$. If $\mathbf{k} = 0$, then $\sigma = 0$ independently of $\mathbf{F}$; thus, traveling waves are neutrally stable with respect to homogeneous perturbations with $\mathbf{k} = 0$. If $\mathbf{k}$ and $\mathbf{F}$ are orthogonal, the dispersion relation coincides with that for $\mathbf{F} = 0$. Thus, if the condition (2.18) is not satisfied, traveling wave solutions with any $\mathbf{F}$ are unstable at least with respect to *transverse* perturbations, with $\mathbf{k} \perp \mathbf{F}$. Later on, we shall investigate the stability of traveling wave solutions with $\mathbf{F} \neq 0$ with respect to disturbances with $\mathbf{k} \neq 0$ for the parameter region defined by (2.18).

In the limit $\alpha \to 0$ (i.e., for $k_n F \ll k^2$) (2.19) is reduced to the quadratic equation for Σ corresponding to $\mathbf{F} = 0$; thus there is no instability for small α if $\mathrm{tr}(M) > 0$, $\det(M) > 0$. For finite values of α, the instability can occur for

$$2w - C^2 + \frac{Z^2}{[\mathrm{tr}(M)]^2} > 0, \tag{2.20}$$

when there exists a solution of the equation $\mathrm{Re}(\Sigma(\alpha_*)) = 0$,

$$\alpha_*^2 = \frac{[\mathrm{tr}(M)]^2 \det(M)}{[\mathrm{tr}(M)]^2(2w - C^2) + Z^2}, \qquad Z = \zeta - C\,\mathrm{tr}(M). \tag{2.21}$$

If (2.20) is satisfied, a traveling wave is unstable with respect to the disturbances with $|\,\alpha\,| > \alpha_*$. Thus, the solution characterized by the wavevector $\mathbf{F}$ is unstable with respect to perturbations with the wavevector $\mathbf{k}$ within the region

$$\left(\frac{\mathbf{k}}{F} \pm \frac{\mathbf{n}}{2\alpha_*}\right)^2 < \frac{1}{4\alpha_*^2}, \tag{2.22}$$

where $\mathbf{n} = \mathbf{F}/F$, i.e., within two disks of the radius $1/2\alpha_*$ with the centers in the points $\mathbf{k} = \pm\mathbf{n}F/2\alpha_*$ (see Figure 1; $k_n = \mathbf{k}\cdot\mathbf{n}$, $k_\tau = \mathbf{k}\cdot\boldsymbol{\tau}$, $\mathbf{n}\cdot\boldsymbol{\tau} = 0$). At the instability boundary, $\sigma = i\Omega k^2$, $\Omega = \alpha_*\zeta/\mathrm{tr}(M)$.

The instability inside the region (2.22) is a new type of instability which is absent in the case of the pure CGL equation. Indeed, if the CGL equation is decoupled from the Burgers equation ($w = 0$), the condition (2.20) is not satisfied for $1 + uv > 0$, $m > 0$. An unusual feature of this new kind of long-wave instability is the fact that if the conditions (2.18), (2.20) hold, then *all* traveling wave solutions with $F \neq 0$ are unstable, while homogeneous oscillations with $F = 0$ are stable; therefore the Busse balloon contains the *single* wavenumber $F = 0$. This property is not the result of the

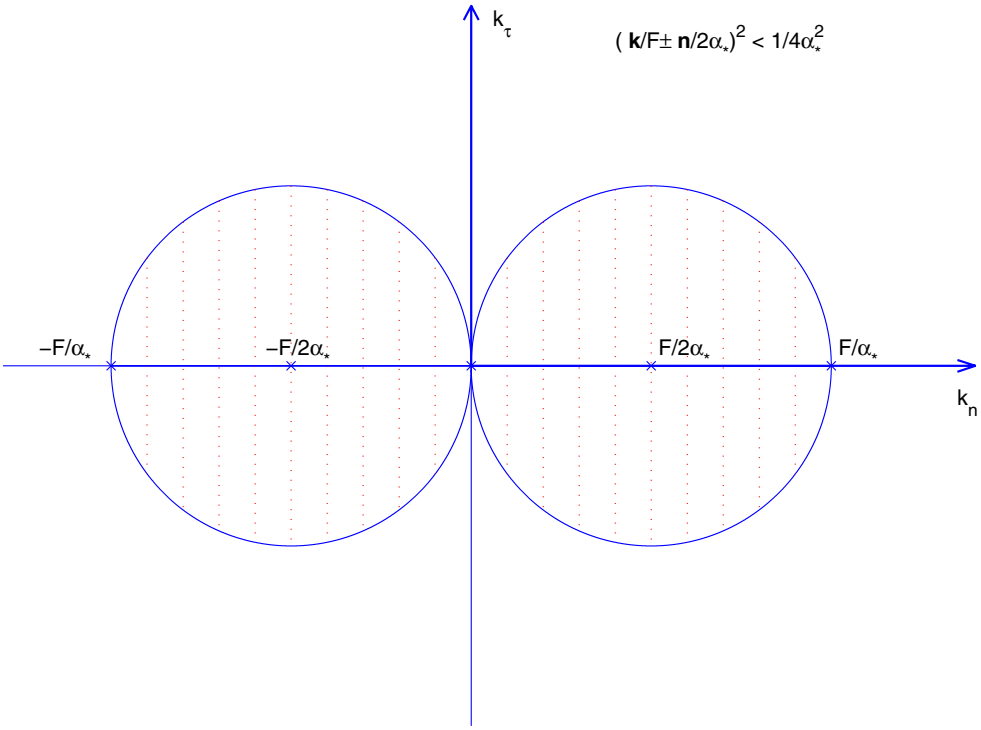


FIG. 1. *Regions of instability (dashed) of the traveling wave* (2.13) *with the wavevector* $\mathbf{F}$, $k_n = \mathbf{k}\cdot\mathbf{n}$, $k_\tau = \mathbf{k}\cdot\boldsymbol{\tau}$; $\mathbf{n}$ *and* $\boldsymbol{\tau}$ *are orthogonal unit vectors;* $\mathbf{n} = \mathbf{F}/\mathrm{F}$.

long-wave approximation of the system (1.1) by the coupled Burgers equations (2.11) but it also remains in the full system (1.1) [24]. The same property was obtained for modulated waves governed by a generalized Ginzburg–Landau equation [33] and [34]. Note that the inequality (2.22) contains only the *ratio* $\mathbf{k}/F$, which is the result of the similarity properties of the system of coupled Burgers equations.

It is instructive to analyze the asymptotics of the solution $\Sigma(\alpha)$ defined by (2.19) in the limit $\alpha \to \infty$, i.e., for $k_n F \gg k^2$ (note that in this case $(k_\tau/F)^2 \ll k_n/F \ll 1$). For $2w - C^2 \neq 0$, one expands $\Sigma(\alpha) \sim \Sigma_1\alpha + \Sigma_0 + \Sigma_{-1}\alpha^{-1} + \dots$, where

$$\Sigma_1 = iC \pm \sqrt{2w - C^2}, \qquad \Sigma_0 = -\frac{1}{2}\mathrm{tr}(M) \pm \frac{Z}{2\sqrt{C^2 - 2w}}. \tag{2.23}$$

If $2w - C^2 < 0$ (specifically, near the instability boundary, where $2w - C^2 = -Z^2/[\mathrm{tr}(M)]^2$), then Σ_1 is purely imaginary and Σ_0 is real. It is easy to show that in the case $\mathrm{tr}(M) > 0$ the instability condition $\Sigma_0 > 0$ is equivalent to (2.20). In this case, $\mathrm{Re}(\sigma) = \Sigma_0 k^2$ to the leading order; i.e., the instability is of the negative viscosity type.

If $2w - C^2 > 0$, one of the roots Σ_1 has a positive real part. In this case, the growth rate $\mathrm{Re}(\sigma) = \pm\sqrt{(2w - C^2)}\cdot\mathbf{k}\cdot\mathbf{F}$ depends *linearly* on the perturbation wavenumber. In the order $O(k)$, the frequency of the disturbances, $\omega(\mathbf{k}) = i\sigma(\mathbf{k})$, is determined by the eigenvalues of the matrix (2.14). Since the matrix is not symmetric it can have a pair of complex conjugate eigenvalues that corresponds to the instability.

This phenomenon is known to be the origin of similar instabilities in several hydrodynamic problems, e.g., the Kelvin–Helmholtz instability in ideal fluid flows [35] and kinetic instability of viscous flows with two or more interfaces ("kinetic α-effect") [36] (see also [37]). Note that on the boundary $2w - C^2 = -Z^2/[\mathrm{tr}(M)]^2$ the instability always appears as a negative-viscosity one, while the kinetic α-effect develops only in the interior of the instability region in the parameter space where $2w - C^2 > 0$ (see Figure 2).

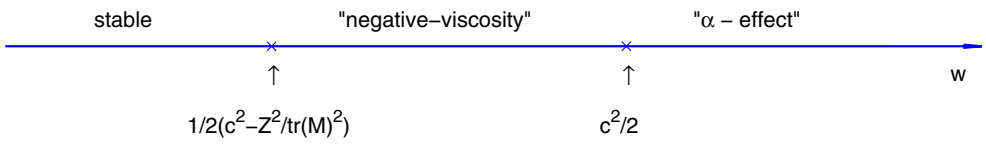


FIG. 2. *Instability types.*

On the boundary $2w - C^2 = 0$, the growth rate of the unstable mode is described by the relation $\mathrm{Re}(\Sigma(\alpha)) \sim \sqrt{\frac{|\alpha|Z}{2}}$; thus $\mathrm{Re}(\sigma(k^2)) \sim k(\mathbf{k}\cdot\mathbf{F})^{1/2}$.

2.3. Weakly nonlinear amplitude equations for long-wave disturbances. In this section we analyze the nonlinear evolution of the new type of instability of the traveling wave solutions described in the previous section, in the case when the size of the instability region in the plane (k_n, k_τ) is small. Consider finite-amplitude perturbations of the traveling wave solution (2.13), $\hat\theta = \theta - \theta_0$, $\hat Q = Q - Q_0$, and choose a new reference frame, $\mathbf{X}' = \mathbf{X} - \mathbf{G}T$. In the new frame of reference, the evolution of $\hat\theta$ and $\hat Q$ is described by the following system of equations:

$$\widehat{\theta_T} = A\nabla^2\widehat\theta + B\nabla^2\widehat Q + 2C\mathbf{F}\cdot\nabla\widehat\theta - \mathbf{F}\cdot\nabla\widehat Q + C\left(\nabla\widehat\theta\right)^2 - \nabla\widehat\theta\cdot\nabla\widehat Q,$$
$$\widehat Q_T = D\nabla^2\widehat Q + E\nabla^2\widehat\theta + 2w\mathbf{F}\cdot\nabla\widehat\theta + w\left(\nabla\widehat\theta\right)^2 - \frac{1}{2}\left(\nabla\widehat Q\right)^2. \tag{2.24}$$

In the present section, we derive the weakly nonlinear amplitude equations valid near the instability boundary determined by (2.20), where the instability takes place only with respect to disturbances with $k/F \ll 1$. In order to obtain the asymptotic equation equally valid for the negative viscosity instability and for the α-effect instability, we will consider in the present section the case when the boundaries of both types of the instabilities mentioned above are close to each other. According to the results of the linear stability analysis (see Figure 2) the two boundaries are close to each other if $[Z/\mathrm{tr}(M)]^2 \ll 1$. Below we consider the case $|Z| \ll 1$, $\mathrm{tr}(M) = O(1)$. Also, we assume that the parameter w is close to the two boundaries, so that the instability takes place with respect to perturbations with $|k_n| \ll 1$ and $|k_\tau| \ll 1$. Below we derive a long-wave asymptotic limit for the system (2.24).

Define $s_i^0 = -2m/C + u(s_r - 1/2)$ and assume $s_i = s_i^0 + R\varepsilon$, $w = C^2/2 + W\varepsilon^2$. Under these assumptions, the width of the instability interval $\frac{F}{\alpha_*}$ (see Figure 1) is $O(\varepsilon)$. Introduce slow time variables, $T_1 = \varepsilon T$, $T_2 = \varepsilon^2 T$, and a new, long-scale spatial variable, $\mathbf{z} = \varepsilon\mathbf{X}' + \mathbf{F}CT_1$; expand $\widehat\theta = \varepsilon^{-1}\widehat\theta_{-1} + \widehat\theta_0 + \varepsilon\widehat\theta_1 + \ldots$, $\widehat Q = \varepsilon^{-1}\widehat Q_{-1} + \widehat Q_0 + \varepsilon\widehat Q_1 + \ldots$; and substitute this expansion in the system (2.24) to obtain the series of problems for $\hat\theta_i$, $\hat Q_i$, $i = -1, 0, 1, \ldots$, in the successive orders of the small

parameter ε. Using solvability conditions for $i = -1, 0$ one obtains the following system of equations for renormalized functions θ and ϕ (see Appendix A for details):

$$\begin{aligned} \frac{\partial\ \theta}{\partial T_2} &= D_1 {\nabla_z}^2 \theta - \nabla_z \theta \cdot \nabla_z \phi, \\ \frac{\partial \phi}{\partial T_2} &= D_2 {\nabla_z}^2 \phi - \frac{1}{2}(\nabla_z \phi)^2 + {\nabla_z}^2 \theta + \mu (\nabla_z \theta)^2. \end{aligned} \tag{2.25}$$

The function θ has a nonzero average gradient and therefore is unbounded for $|\mathbf{z}| \to \infty$; the function ϕ has a nonzero average time derivative and thus is unbounded for $T_2 \to \infty$. It is convenient to introduce new functions, $\widetilde{\theta} = \theta - \mathfrak{z} \cdot \mathbf{n}$ and $\widetilde{\phi} = \phi - \mu \mathfrak{t}$, where $\mathfrak{z} = C^2 RF\mathbf{z}$, $T_2 = \mathfrak{t}/(C^2 R\mathbf{F})^2$, $\mathbf{n} = \mathbf{F}/F$, that will be bounded for $|\mathfrak{z}| \to \infty$ and $\mathfrak{t} \to \infty$. The system (2.25) will then become

$$\begin{aligned} \widetilde{\theta}_{\mathfrak{t}} &= D_1 {\nabla_{\mathfrak{z}}}^2 \widetilde{\theta} - (\mathbf{n} + \nabla_{\mathfrak{z}} \widetilde{\theta}) \cdot \nabla_{\mathfrak{z}} \widetilde{\phi}, \\ \widetilde{\phi}_{\mathfrak{t}} &= D_2 {\nabla_{\mathfrak{z}}}^2 \widetilde{\phi} - \frac{1}{2}(\nabla_{\mathfrak{z}} \widetilde{\phi})^2 + {\nabla_{\mathfrak{z}}}^2 \widetilde{\theta} + \mu (\nabla_{\mathfrak{z}} \widetilde{\theta})^2 + 2\mu \mathbf{n} \cdot \nabla_{\mathfrak{z}} \widetilde{\theta}. \end{aligned} \tag{2.26}$$

The system (2.26) can be considered as a simplified version of the system (2.24) which is still capable of describing both types of the long-wave instabilities. Indeed, the linear dispersion relation for the system (2.26),

$$\begin{vmatrix} \sigma + D_1 k^2 & ik_n \\ -2\mu i k_n + k^2 & \sigma + D_2 k^2 \end{vmatrix} = 0, \tag{2.27}$$

keeps all characteristic features of the general dispersion relation (2.14). The oscillatory instability of the negative viscosity type occurs in the region

$$\mu > \mu_* = -\frac{1}{2(D_1 + D_2)^2}, \tag{2.28}$$

with the boundary of the instability region in $\mathbf{k}$-space described by

$$(\mathbf{k} \pm k_m \cdot \mathbf{n})^2 = {k_m}^2 \equiv \frac{\mu - \mu_*}{2D_1 D_2}. \tag{2.29}$$

The α-effect occurs for $\mu > 0$.

2.4. Absolute and convective instability. At this point one needs to distinguish between the absolute instability and the convective instability [38, 39]. Consider the one-dimensional version of the dispersion relation for the system (2.26),

$$\sigma^2 + D_1 D_2 k^4 + (D_1 + D_2) k^2 \sigma - 2\mu k^2 - ik^3 = 0. \tag{2.30}$$

In order to find the boundary of the absolute instability, consider the roots $k(\sigma)$ for $\mathrm{Re}\sigma \gg 1$, separate them into two classes: $\mathrm{Im}k(\sigma) > 0$ and $\mathrm{Im}k(\sigma) < 0$, and follow the appearance of double roots $k_i = k_i(\sigma)$, $i = 1, 2, \dots$, with the decrease of $\mathrm{Re}\sigma$. The double roots correspond to $d\sigma/dk = 0$, but only those that appear by merging of the roots corresponding to *different* classes are relevant to the development of the absolute instability. The latter takes place for $\mathrm{Re}\sigma(k_i) > 0$.

For $|\mu| \ll 1$ dispersion relation (2.30) is simplified under the assumption $k = O(|\mu|)$, $\sigma = O(|\mu|^{3/2})$, and in the leading order becomes

$$-ik^3 - 2\mu k^2 + \sigma^2 = 0. \tag{2.31}$$

Similar to [38], we shall use the complex frequency $\omega(k) = \omega'(k) + i\omega''(k) = i\sigma(k)$. Let us analyze the "motion" of roots in the complex k-plane with the decrease of ω''. In the limit $\omega'' \gg 1$, the equation (2.31) has three roots $k(\omega)$: one pure imaginary root with $\mathrm{Im}k > 0$ and two roots of the type $\pm a + ib$ in the lower half-plane ($b < 0$).

For $\mu < 0$, one obtains that the two roots in the lower half plane can merge on the imaginary axis, but this is not a sign of the appearance of an absolute instability, because both merging roots belong to the same class.

For $\mu > 0$, one gets that neither of the two roots $\pm a + ib$ crosses the imaginary axis and, therefore, cannot merge. The double root can occur in the upper half-plane, $\mathrm{Im}k > 0$, due to merging of one of these roots and the pure imaginary root. Positive values of $\mathrm{Re}\sigma$ for these double roots show that an absolute instability appears.

Finally, one comes to the conclusion that $\mu = 0$ is the boundary of the absolute instability which coincides with the transition from the negative viscosity instability type to the α-effect instability type in the long-wave limit, so that for $-|\mu_*| < \mu < 0$ the instability is *convective* and of the negative viscosity type, and for $\mu > 0$ the instability is *absolute* and of the α-effect type.

2.5. Nonlinear waves near the threshold $\boldsymbol{\mu = \mu_*}$. Now we study the spontaneous generation of nonlinear waves near the threshold $\mu = \mu_*$, described by (2.26), in more detail. For small deviations of the parameter μ near the threshold, $\mu = \mu_* + \mu_2\delta^2 + \ldots$, we expand $\widetilde{\theta} = \delta\theta_1 + \delta^2\theta_2 + \delta^3\theta_3 + \ldots$, $\widetilde{\phi} = \delta\phi_1 + \delta^2\phi_2 + \delta^3\phi_3 + \ldots$ and use $\nabla_z = \delta\nabla_\mathfrak{z}$, $\partial_\mathfrak{t} = \delta\partial_{t_1} + \delta^2\partial_{t_2} + \delta^3\partial_{t_3} + \delta^4 t_4 + \ldots$.

For the lowest order, one obtains

$$\frac{\partial\theta_1}{\partial t_1} = -\mathbf{n}\cdot\nabla_\mathfrak{z}\phi_1, \qquad \frac{\partial\phi_1}{\partial t_1} = 2\mu_*\mathbf{n}\cdot\nabla_\mathfrak{z}\theta_1, \tag{2.32}$$

which can be transformed to

$$\frac{\partial^2\theta_1}{\partial {t_1}^2} + 2\mu_*(\mathbf{n}\cdot\nabla_\mathfrak{z})^2\theta_1 = 0. \tag{2.33}$$

Introduce the coordinates (x, y) as $x = \mathbf{n}\cdot\mathbf{z}$, $y = \boldsymbol{\tau}\cdot\mathbf{z}$. Then (2.33) becomes $\partial_{t_1 t_1}\theta_1 + 2\mu_*\partial_{xx}\theta_1 = 0$, whose solution is

$$\theta_1 = \theta_1^+(x + \sqrt{2|\mu_*|}t_1, y) + \theta_1^-(x - \sqrt{2|\mu_*|}t_1, y), \tag{2.34}$$

$$\phi_1 = \sqrt{2|\mu_*|}[-\theta_1^+(x + \sqrt{2|\mu_*|}t_1, y) + \theta_1^-(x - \sqrt{2|\mu_*|}t_1, y)]. \tag{2.35}$$

In the next order, $O(\delta^3)$, one obtains

$$\theta_2 = \theta_2^+(x_+, y) + \theta_2^-(x_-, y), \qquad \phi_2 = \phi_2^+(x_+, y) + \phi_2^-(x_-, y), \tag{2.36}$$

where $x_\pm = x \pm \sqrt{2|\mu_*|}t_1$. Using solvability conditions in the successive orders of δ, after appropriate rescaling, one finally obtains the following evolution equation for a single function $\Theta \propto \theta_1^+ + \delta\theta_2^+$:

$$\begin{aligned}&\frac{\partial\Theta}{\partial T} + \widetilde{C}\frac{\partial\Theta}{\partial X} + \mathcal{I}[\nabla^2\Theta] + 3(\nabla\Theta)^2 \\ &+\widetilde{\delta}\left\{\nabla^2\Theta + \mathcal{I}^2[\nabla^2\Theta] + \widetilde{D}\mathcal{I}[(\nabla\Theta)^2] + \widetilde{E}\nabla\Theta\cdot\mathcal{I}[\nabla\Theta]\right\} + O(\widetilde{\delta}^2) = 0,\end{aligned} \tag{2.37}$$

where $\mathcal{I}[f(X, Y)] \equiv \int_0^X \nabla^2 f(X', Y)\, dX'$. The coefficients and other details of the derivation are given in Appendix B.

In the one-dimensional case, $\Theta = \Theta(X,T)$, (2.37) is reduced to the *perturbed Korteweg–de Vries equation* for the function $U = \frac{\partial \Theta}{\partial X}$:

$$\frac{\partial U}{\partial T} + \widetilde{C}\frac{\partial U}{\partial X} + \frac{\partial^3 U}{\partial X^3} + 6U\frac{\partial U}{\partial X} + \widetilde{\delta}\left[\frac{\partial^2 U}{\partial X^2} + \frac{\partial^4 U}{\partial X^4}\right] = 0. \tag{2.38}$$

This equation has been derived formerly in different physical problems (see [12, 13, 14]). Several two-dimensional generalizations of (2.38) are known (see [40, 41, 42, 43, 44]). However, to our knowledge, (2.37) is obtained here for the first time.

2.6. Numerical simulations. Far from the threshold $\mu = \mu_*$, full equations (2.26) should be used. The advantage of the system (2.26) is that it governs both the negative viscosity instability and the α-effect instability.

First, we have performed numerical simulations of the system (2.26) in one dimension by means of a pseudospectral code, with periodic boundary conditions. The time-integration has been carried out in the Fourier space using the Crank–Nicolson scheme for the linear operator, and the Adams–Bashford scheme for the nonlinear one. Figure 3 shows typical snapshots of traveling wave solutions obtained numerically for different values of the parameter μ. For $\mu < 0$, as well as for very small positive values of μ, we have observed the formation of spatially periodic uniformly traveling waves; see Figure 3(a). These waves correspond to spatiotemporal periodic modulations of the original wave (2.13) coupled to the traveling wave of the zero mode (frontal deformations). With the increase of the parameter $\mu > 0$ (corresponding to the α-effect-type instability of the original wave (2.13)), this wave structure becomes unstable with respect to spatial modulations; see Figures 3(b) and 3(c). Patterns shown in Figures 3(b) and 3(c) are snapshots of waves that travel from right to left and also undergo very small modulations on a very slow time scale. Note that these modulated wave structures resulted from a very slow evolution of a spatially irregular system of typical "Burgers shocks" that forms initially from random initial data. This intermediate system of Burgers shocks is similar to that shown in Figure 3(d) for $\mu = 0.15$; it produces spatially irregular modulations of the original wave (2.13) coupled to the modulations of the zero mode. On the fast time scale, the pattern shown in Figure 3(d) travels from right to left as a whole with almost constant speed, while it also changes on a very slow time scale. For example, in the case $\mu = 0.1$, this evolution results in a spatially modulated wave structure shown in Figure 3(c). With the increase of the parameter μ, the time required for the formation of a modulated wave pattern from the spatially irregular one increases drastically. Probably, with the increase of μ, the attractor corresponding to a modulated wave pattern becomes unstable and transition to chaos occurs. However, in order to verify this assumption, one needs to perform extremely long computations which are beyond the scope of this paper. Note that the characteristic spatial scale of the wave patterns increases with the increase of μ. At the same time, once established, the characteristic spatial scale does not change, so no coarsening has been observed when the value of the parameter μ is fixed.

Our numerical simulations of the system (2.26) in two dimensions (by means of a pseudospectral code with periodic boundary conditions), both for $\mu_* < \mu < 0$ and $\mu > 0$, exhibit two-dimensional wavy patterns shown in Figure 4. Here, the pattern is rapidly moving along the x-axis in the negative direction and undergoes slow evolution in the transversal direction: the cells shown in Figure 4 slowly merge together and then split, forming new cells. The average spatial scale of the cells remains the same, so no coarsening has been observed in the two-dimensional case either.

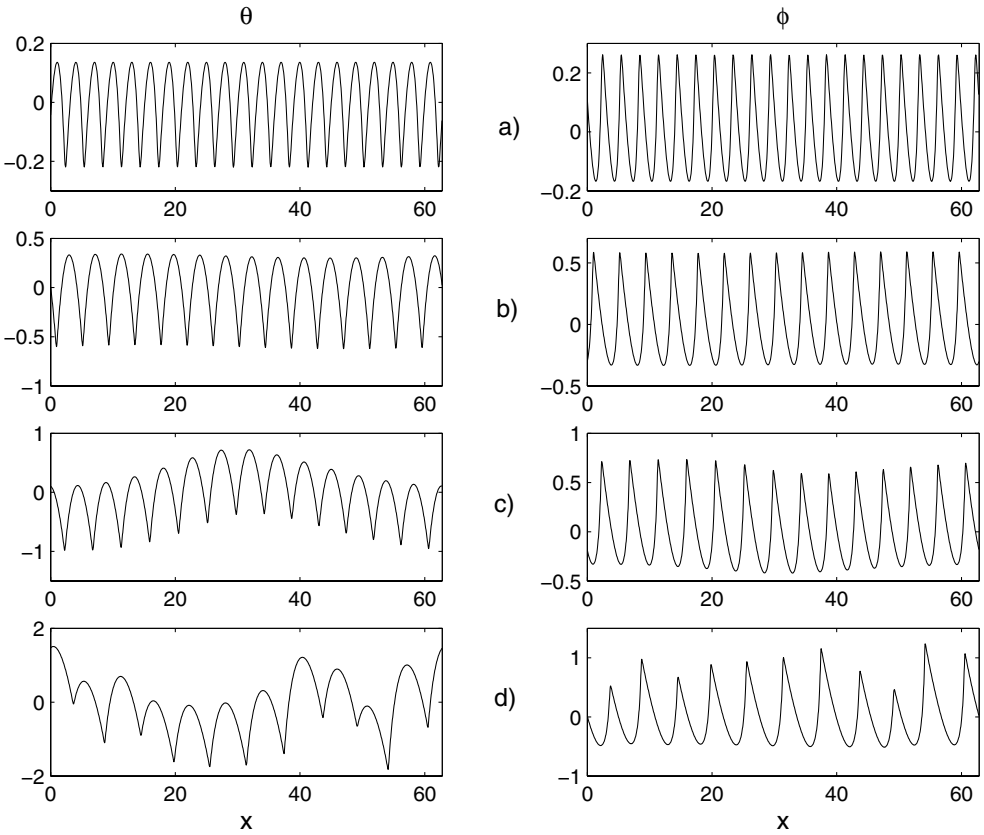


FIG. 3. *Typical traveling wave structures (travel from right to left): numerical solutions of the system* (2.26) *in one dimension for* $D_1 = 0.1$, $D_2 = 0.6$ *and various values of* μ*:* (a) $\mu = -0.3$*;* (b) $\mu = 0.051$*;* (c) $\mu = 0.1$*;* (d) $\mu = 0.15$.

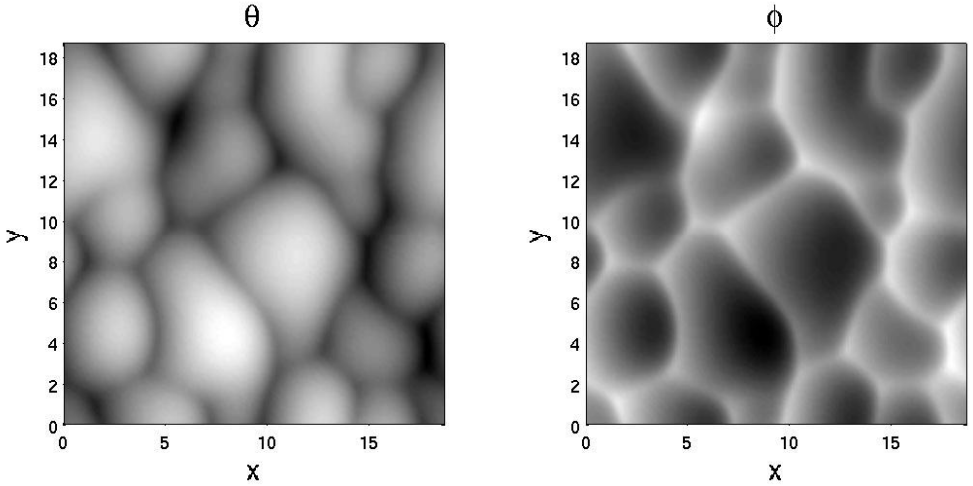


FIG. 4. *Numerical solution of the system* (2.26) *in two dimensions at a particular moment of time;* $D_1 = 0.1$, $D_2 = 0.6$*;* $\mu = -0.3$.

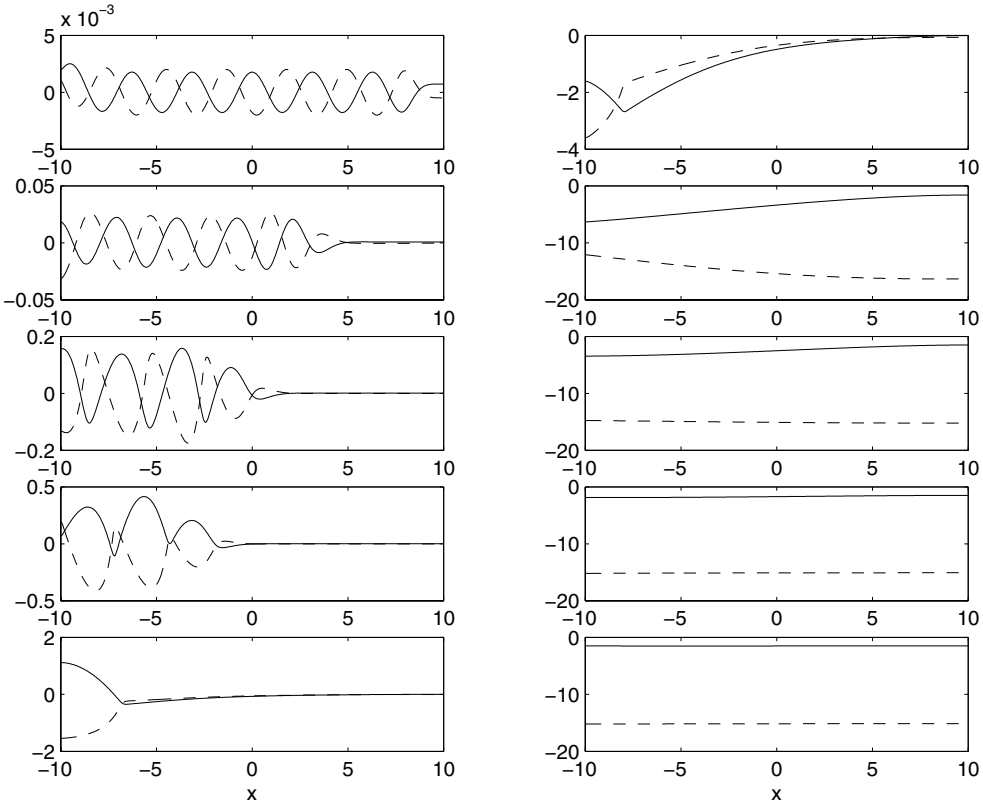


FIG. 5. *Numerical solutions of the system* (2.26) *in one dimension with Neumann boundary conditions, at different moments of time (time increases from top to bottom and from left to right), demonstrating the* convective *instability. Solid lines—*$\theta(z)$*; dashed lines—*$\phi(z)$.

Note that the formation of the modulated traveling wave structure is caused by the periodic boundary conditions. With these conditions, no difference between the absolute and convective instabilities can be observed since the wave leaving the computational domain comes back from the other side. In order to demonstrate the difference between these two instability types, we have performed numerical simulations of the system (2.26) in one dimension by means of a finite-difference code (using the semi-implicit Crank–Nicolson scheme) with Neumann boundary conditions. The results of the simulations are presented in Figures 5 and 6. Figure 5 shows the development of the *convective* instability for $\mu_* < \mu < 0$. One can see that growing perturbations are transported by the convective terms from right to left but gradually escape from the domain and leave behind a trivial steady state, $\theta = \text{const}$, $\phi = \text{const}$. On the contrary, in the case of $\mu > 0$ corresponding to the *absolute* instability, shown in Figure 6, one can see that, although the perturbations are still transported to the left by the convective terms, they are too slow to leave the domain so that finally they grow in the whole domain and form a nontrivial spatially nonuniform structure. This structure is not stationary but continues to evolve in time.

In concluding this section we present some typical results of numerical simulations of the full system (2.24) in one dimension, with periodic boundary conditions. Figure 7 shows a snapshot of a wave, traveling from right to left, resulting from the instability

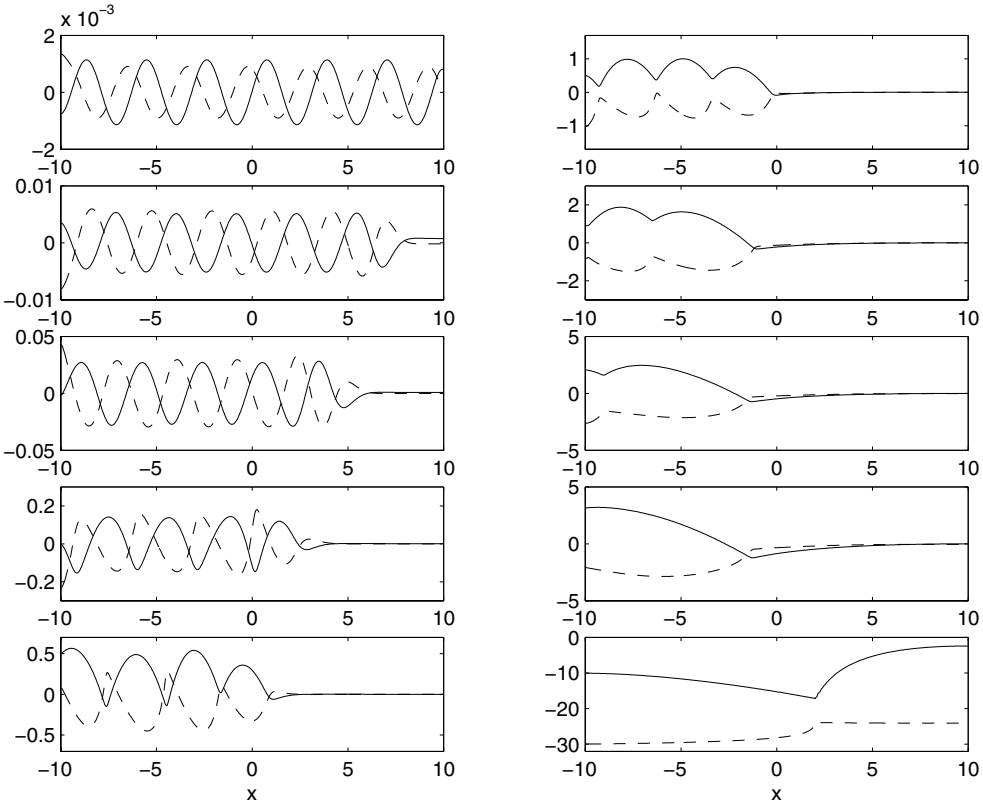


FIG. 6. *Numerical solutions of the system* (2.26) *in one dimension with Neumann boundary conditions, at different moments of time (time increases from top to bottom and from left to right), demonstrating the* absolute *instability. Solid lines—$\theta(z)$; dashed lines—$\phi(z)$.*

of the solution (2.13) with $F = 4.0$. Other parameters have been taken equal to those used in the computations of the full system of CGL–Burgers equations (1.1) performed in [24] that showed the instability of a harmonic traveling wave resulting in a modulated traveling wave. One can see that it corresponds to the formation of a modulated traveling wave that was observed in the numerical simulations of the full system (1.1) performed in [24].

3. Domain wall solutions. In the previous section, we investigated the stability of solutions (2.13) with the constant values of **F** and **G**. One can think of a situation when several waves with different wavevectors **F** and different front-slope vectors **G** develop in different parts of the domain. In this case, regions occupied by the waves with different wavevectors will be divided by domain walls.

In the present section, we investigate a particular case of a one-dimensional stationary domain wall between traveling waves with equal values of the parameters ω and c (see (2.13)). Consider the one-dimensional version of the system (2.25) and define $\theta_z = F(z)$, $\phi_z = H(z)$ and consider solutions with $F(+\infty) = -F(-\infty)$ and $H(+\infty) = -H(-\infty)$. Such solutions correspond to the *domain walls* between two traveling waves with the opposite wavenumbers, each wave traveling on the background of an inclined front.

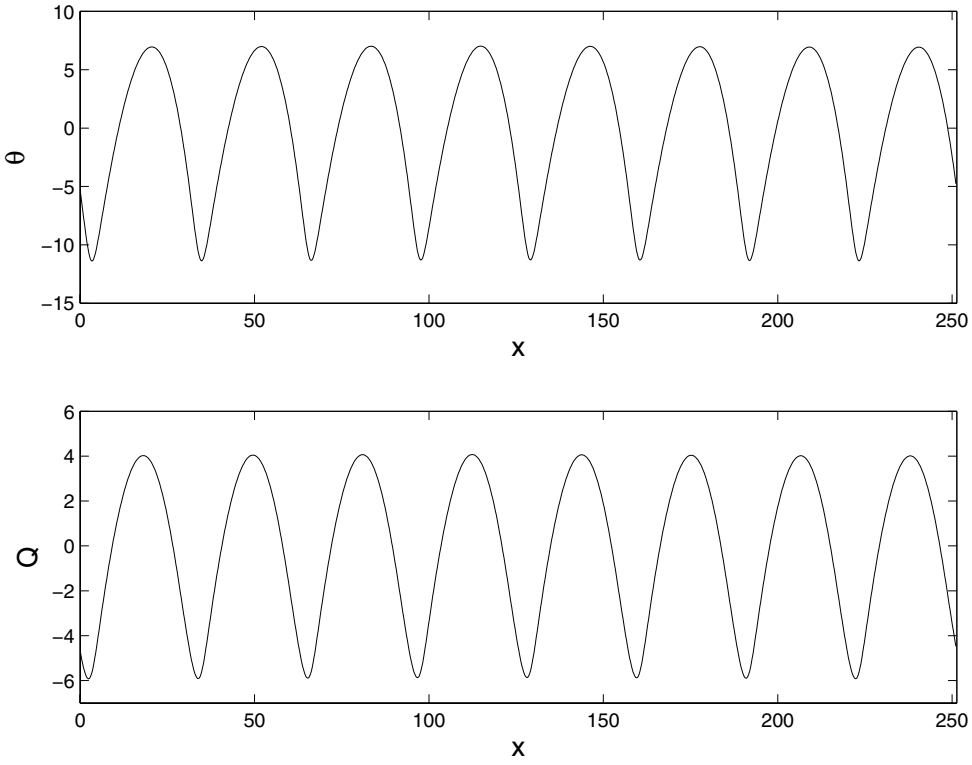


FIG. 7. *Numerical solution of the system* (2.24) *for* $A = 3.0$, $B = -3.0$, $C = 1.0$, $D = 10.45$, $E = -0.9$, $w = 0.45$*: Spatially periodic wave traveling from right to left.*

Assume that $\theta_{T_2} = \Omega$, $\phi_{T_2} = \Phi$ are constants and rewrite the system (2.25) in the form

$$D_1 F_z = \Omega + FH, \tag{3.1}$$
$$D_2 H_z + F_z = \Phi + \frac{1}{2}H^2 - \mu F^2.$$

The dynamical system (3.1) determines stationary profiles $F(z)$, $H(z)$. The critical points of this dynamical system correspond to traveling waves, while the domain walls are governed by the separatrices joining different critical points.

For $\Phi \neq 0$, $\Omega \neq 0$, the critical points of system (3.1), $F(z) = f = \text{const}$, $H(z) = h = \text{const}$, must satisfy the conditions

$$f^2 = f_\pm^2 \equiv \frac{-\Phi \pm \sqrt{\Phi^2 + 2\mu\Omega^2}}{(-2\mu)}, \qquad h = h_\pm \equiv -\frac{\Omega}{f_\pm}. \tag{3.2}$$

Thus, for $\mu < -\Phi^2/(2\Omega^2)$, the dynamical system (3.1) has no critical points. For $\mu = -\Phi^2/(2\Omega^2)$, a pair of critical points appears in the case $\Phi < 0$; otherwise, there are no critical points. For $-\Phi^2/(2\Omega^2) < \mu < 0$ and $\Phi < 0$ there exist two pairs of critical points. In the special case $\mu = 0$, there is one pair of critical points while the other pair of critical points approaches infinity. For $\mu > 0$, there is one pair of critical points for *any* Φ. Thus, the critical points always appear in pairs $(f_\pm, h_\pm)$, $(-f_\pm, -h_\pm)$. Later on, we shall define $f_\pm > 0$.

The analysis of the eigenvalues of the linearized system (3.1) near the critical points leads to the following conclusions. For $\Phi < 0$, $\Omega > 0$, $-\Phi^2/2\Omega^2 < \mu < 0$, the points $(\pm f_+, \pm h_+)$ are saddles, and the point (f_-, h_-) is a stable node, while the point $(-f_-, -h_-)$ is an unstable node. For $\Phi < 0$, $\Omega < 0$, $-\Phi^2/2\Omega^2 < \mu < 0$, the points $(\pm f_+, \pm h_+)$ are saddles, and for $\mu < -1/[2(D_1 - D_2)^2]$, the points $(\pm f_-, \pm h_-)$ are nodes. However, for $-1/[2(D_1 - D_2)^2] < \mu < 0$, there always exists such an interval of $|\Phi/\Omega|$ that the point (f_-, h_-) is a focus. This focus is unstable for $|\mu| > 1/[2(D_1 + D_2)^2]$ and stable otherwise.

Consider now solutions corresponding to the domain walls. The trajectories in the phase plane (F, H) are governed by the equation

$$\frac{D_2}{D_1}\frac{dH}{dF} = -\frac{1}{D_1} + \frac{\Phi + H^2/2 - \mu F^2}{\Omega + FH}. \tag{3.3}$$

Any trajectory that leads from one critical point to another corresponds to a domain wall. There exists a class of symmetric domain walls that are described by trajectories invariant with respect to the transformation $F \to -F$, $H \to -H$. Such a trajectory connects the critical points $(-f, -h)$ and (f, h) passing through the point $(0, 0)$.

A trajectory that leaves the point $(0, 0)$ is generically attracted to the stable critical point or leads to infinity. However, there exists a codimension-1 manifold in the parameter space (Ω, Φ, μ) where the trajectory is "nongeneric" and leads to a saddle critical point. This manifold separates the region of the attraction to the stable critical point and the region where the trajectory tends to infinity.

Generally, (3.3) is not analytically integrable. However, some classes of analytical exact solutions can be found. There exist surfaces in the parameter space (Ω, Φ, μ) where a trajectory that passes through the point $(0, 0)$ is just $H(z) = kF(z)$, $k \neq 0$. These surfaces are determined by the conditions

$$\frac{\Phi}{\Omega} = \frac{sD_2 + D_1}{D_1}, \qquad s = \frac{1 \pm \sqrt{1 - 2\mu D_1(2D_2 - D_1)}}{2D_2 - D_1}, \tag{3.4}$$

where $1 - 2\mu D_1(2D_2 - D_1) > 0$. The corresponding solutions are

$$F = f^{\pm} \tanh \beta(z - z_0), \qquad H = h \tanh \beta(z - z_0), \tag{3.5}$$

where

$$f^{\pm} = \frac{1 \pm \sqrt{1 - 2\mu D_1(2D_2 - D_1)}}{2\mu}\beta, \qquad h = -D_1\beta, \tag{3.6}$$

and $\mu < 0$, β and z_0 are arbitrary constants (β can be taken positive without loss of generality).

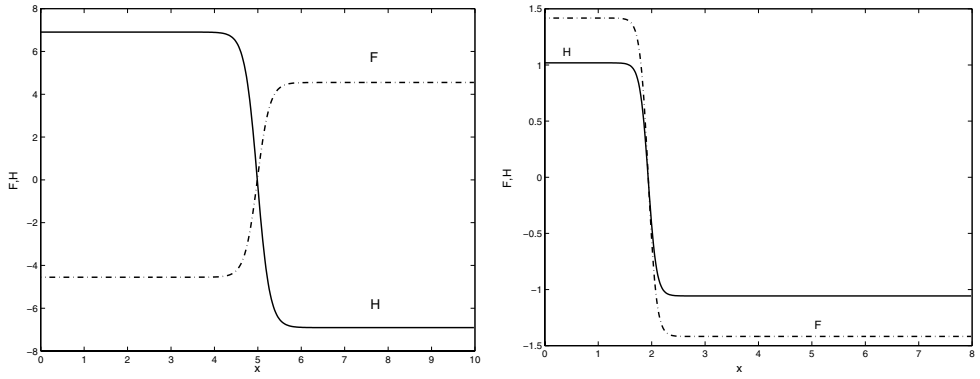
One can show that for $(2D_2 - D_1) > 0$, $\Omega < 0$, the exact solution (3.5) corresponds to a nongeneric trajectory that leads to the saddle critical point (f^-, h) if $\mu < -1/[2(D_2 - D_1)^2]$. Otherwise, the point (f^-, h) determined by (3.6) corresponds to a node. For $2D_2 - D_1 < 0$, $|\mu| > 1/[2{D_1}^2]$, $\Omega < 0$, the trajectory (3.5) leads to a saddle for $\mu > -1/[2(D_2 - D_1)^2]$, and to a node for $\mu < -1/[2(D_2 - D_1)]^2$. For $\Omega < 0$, $|\mu| > 1/[2{D_1}^2]$, $D_2 - D_1 < 0$, the point (f^+, h) is a node for $\mu < -1/[2(D_2 - D_1)^2]$. For $2D_2 - D_1 = 0$, the point (f^-, h) is a node, and the point (f^+, h) is a saddle for $\mu > -2/D_1^2$. Otherwise, the point (f^+, h) is a node. In the case $\Omega > 0$, $D_2 > D_1$ the point (f, h) is a node for $\mu > -1/[2(D_2 - D_1)^2]$. Otherwise, the trajectory (3.5) leads to a saddle.

Note that there is one more class of exact solutions, corresponding to the case $\Omega = 0$,

$$F = 0, \qquad H(z) = h \tanh \beta z, \tag{3.7}$$

where $h = -2D_2\beta, \beta > 0$. In this case, the point $(0, h)$ is a node.

We have studied the family of solutions that correspond to the domain walls between two traveling waves with the opposite wavenumbers by rewriting the system (2.25) in one dimension for $\chi(z,t) = \theta_z$ and $\psi(z,t) = \phi_z$, and solving it numerically, by means of a finite-difference, semi-implicit Crank–Nicholson scheme, with Neumann boundary conditions. We have found that, depending on the initial conditions and the values of the parameters D_1, D_2, and μ, the solutions evolve either toward stationary domain walls described above or tend to constant solutions, $\chi = \text{const}$, $\psi = \text{const}$, corresponding to a single traveling wave on a slant planar front. Examples of stable domain wall solutions are shown in Figure 8. We have checked that they compare very well with the analytical solutions (3.5), (3.6).



(a) (b)

FIG. 8. *Numerical solution of the system* (2.25) *in one dimension for* $\theta_z = F$ *and* $\phi_z = H$ *in the form of stationary domain walls:* (a) $D_1 = 2.0$, $D_2 = 6.0$, $\mu = -5.0$; (b) $D_1 = 0.2$, $D_2 = 0.6$, $\mu = -5.0$.

In concluding this section we note that there also exists a family of solutions in the form of moving domain walls: indeed, if $F_0(z) = \theta_z$, $H_0(z) = \phi_z$ is a solution of (3.1) that describes a stationary domain wall with the definite values of $\Omega_0 = \theta_{T_2}$, $\Phi_0 = \phi_{T_2}$, then a moving domain wall $F(z,T_2) = \theta_z(z,T_2) = F_0(z - cT_2)$, $H(z,T_2) = \phi_z(z,T_2)_z = c + H_0(z - cT_2)$ with $\Omega = \Omega_0$, $\Phi = \Phi_0 + c^2/2$ ($\theta(z,T_2) = \int F(z,T_2)\,dz + \Omega T_2$, $\phi(z,T_2) = \int H(z,T_2)\,dz + \Phi T_2$) is a solution of (2.25) (in one dimension).

4. Conclusions. We have investigated the stability and nonlinear dynamics of traveling waves in a system with translation symmetry that exhibits a long-wave oscillatory instability. Near the instability threshold, the nonlinear dynamics of this system is governed by a CGL equation for the unstable mode coupled to the Burgers equation for the Goldstone mode generated by the translation symmetry. We have demonstrated that, in the long-wave limit, this system of coupled nonlinear evolution equations is reduced to the system of two *coupled Burgers equations*—for the *phase* of

the unstable mode and for the Goldstone mode—and we have studied the nonlinear dynamics of waves described by this system.

We have shown that the derived system of coupled Burgers equations captures well the main effect of the coupling between the unstable oscillatory mode and the Goldstone mode, namely, the change of the Benjamin–Feir stability region as well as the unique situation when all waves, except the one with the zero wavenumber (planar oscillations), are unstable. Also, the derived system of the coupled Burgers equations describes the two main instability types: the convective instability of the negative viscosity type, and the absolute instability of the α-effect type. At the boundaries of these instabilities we have derived the weakly nonlinear evolution equations describing the weakly nonlinear dynamics of unstable waves near the instability thresholds. We have also performed numerical simulations of the coupled Burgers equations and studied the strongly nonlinear dynamics of the discovered instabilities. We have observed the formation of periodically and chaotically modulated traveling waves, in both one and two dimensions. The similar type of nonlinear behavior was also observed earlier in simulating the coupled CGL and Burgers equations. Finally, we have shown that the derived system of coupled Burgers equations also describes the domain boundaries between waves with different wavenumbers.

Thus, the derived system of coupled Burgers equations can be considered as a *generic* system that describes the *long-wave* dynamics of waves in systems that exhibit an oscillatory instability at a zero wavenumber in the presence of the Goldstone mode associated with the translation symmetry. This situation is typical of physical systems with uniformly propagating fronts, such as planar flame fronts in combustion [23], [24], solidification in a hypercooled melt [45], as well as in general reaction-diffusion systems [46]. In this paper we considered combustion fronts as an example, but the results obtained also can be applied to any other system of this class. Waves on uniformly propagating fronts were observed in experiments on premixed-flame gaseous combustion [47], solid combustion (self-propagating high-temperature synthesis) [48], as well as frontal polymerization [49]. These systems would be natural candidates for experimental verification of theoretical conclusions presented in this paper.

Appendix A. Derivation of (2.25). We introduce slow time variables, $T_1 = \varepsilon T$, $T_2 = \varepsilon^2 T$, and a new, long-scale spatial variable, $\mathbf{z} = \varepsilon \mathbf{X}' + \mathbf{F} C T_1$; we expand $\widehat{\theta} = \varepsilon^{-1}\widehat{\theta}_{-1} + \widehat{\theta}_0 + \varepsilon\widehat{\theta}_1 + \dots, \widehat{Q} = \varepsilon^{-1}\widehat{Q}_{-1} + \widehat{Q}_0 + \varepsilon\widehat{Q}_1 + \dots$; and we substitute this expansion in the system (2.24). As a leading-order problem, one obtains

$$\frac{\partial \widehat{\theta}_{-1}}{\partial T_1} = C(\nabla_z \widehat{\theta}_{-1})^2 - \nabla_z \widehat{\theta}_{-1} \cdot \nabla_z \widehat{Q}_{-1} + C\mathbf{F} \cdot \nabla_z \widehat{\theta}_{-1} - \mathbf{F} \cdot \nabla_z \widehat{Q}_{-1}, \tag{A.1}$$

$$\frac{\partial \widehat{Q}_{-1}}{\partial T_1} = \frac{C^2}{2}(\nabla_z \widehat{\theta}_{-1})^2 - \frac{1}{2}(\nabla_z \widehat{Q}_{-1})^2 + C^2 \mathbf{F} \cdot \nabla_z \widehat{\theta}_{-1} - C\mathbf{F} \cdot \nabla_z \widehat{Q}_{-1},$$

where ∇_z is the gradient with respect to the new long-scale variable $\mathbf{z}$. Taking $\widehat{Q}_{-1} = C\widehat{\theta}_{-1} + \phi_{-1}$, one obtains from (A.1)

$$\frac{\partial \phi_{-1}}{\partial T_1} = -\frac{1}{2}(\nabla_z \phi_{-1})^2. \tag{A.2}$$

If ϕ_{-1} is initially nonzero, $|\nabla_z \phi_{-1}|$ becomes infinite in a finite time.[1] Thus, in order

[1] Indeed, the equation $u_t + uu_z = 0$, $u = \frac{\partial \phi_{-1}}{\partial z}$ for any initial condition has an exact solution $z = f(u) + ut$; therefore $u_z = \frac{1}{f'(u)+t}$. Except the case when $f'(u) > 0$ everywhere, $u_z \to \infty$ in a finite time.

to describe smooth solutions we have to assume $\phi_{-1} = 0$. Therefore,

$$\widehat{Q}_{-1} = C\widehat{\theta}_{-1}. \tag{A.3}$$

From (A.3) and (A.1) one finds $\partial\widehat{\theta}_{-1}/\partial T_1 = \partial\widehat{Q}_{-1}/\partial T_1 = 0$, and hence the solution does not depend on T_1.

In the next order one obtains

$$\frac{\partial\widehat{\theta}_{-1}}{\partial T_2} = (A + B^0C)\nabla_z{}^2\widehat{\theta}_{-1} + C(\nabla_z\widehat{\theta}_{-1} + \mathbf{F})\cdot\nabla_z\widehat{\theta}_0 - (\nabla_z\widehat{\theta}_{-1} + \mathbf{F})\cdot\nabla_z\widehat{Q}_0, \tag{A.4}$$

$$C\frac{\partial\widehat{\theta}_{-1}}{\partial T_2} = (CD + E)\nabla_z{}^2\widehat{\theta}_{-1} + C^2(\nabla_z\widehat{\theta}_{-1} + \mathbf{F})\cdot\nabla_z\widehat{\theta}_0 - C(\nabla_z\widehat{\theta}_{-1} + \mathbf{F})\cdot\nabla_z\widehat{Q}_0,$$

where $B^0 = vs_r - s_i^0$, and we take $\widehat{Q}_0 = C\widehat{\theta}_0 + \phi_0$. The solvability condition for the problem (A.4) is

$$E - 2w_0B^0 - C(A - D) = 0, \tag{A.5}$$

where $C^2 = 2w_0$. One can see that the solvability condition (A.5) is automatically satisfied. From (A.4) one obtains

$$\frac{\partial\widehat{\theta}_{-1}}{\partial T_2} = (A + B^0C)\nabla_z{}^2\widehat{\theta}_{-1} - (\nabla_z\widehat{\theta}_{-1} + \mathbf{F})\cdot\nabla_z\phi_0. \tag{A.6}$$

Taking a corresponding linear combination of the equations for $\hat{\theta}_0$ and $\hat{Q}_0$, which are obtained in the next order, one gets the following problem:

$$\frac{\partial\phi_0}{\partial T_2} = (D - CB^0)\nabla_z{}^2\phi_0 - \frac{1}{2}(\nabla_z\phi_0)^2 + RC^2\nabla_z{}^2\widehat{\theta}_{-1} + W(\nabla_z\widehat{\theta}_{-1} + 2\mathbf{F})\cdot\nabla_z\widehat{\theta}_{-1}. \tag{A.7}$$

Define $D_1 = A + CB^0$, $D_2 = D - CB^0$, $\phi = \phi_0 + WF^2T_2$, $\theta = C^2R(\widehat{\theta}_{-1} + \mathbf{F}\cdot\mathbf{z})$, $\mu = W/(RC^2)^2$. Since $\mathrm{tr}(M) > 0$, $\det(M) > 0$, the coefficients D_1 and D_2 are positive. Note that if the function $\widehat{\theta}_{-1}$ is bounded for $|\mathbf{z}| \to \infty$, the asymptotics of the function θ is $\theta \sim \mathbf{F}\cdot\mathbf{z}$. Then (A.6) and (A.7) can be written in the form of the system (2.25).

Appendix B. Derivation of (2.37). The functions ${\theta_2}^{\pm}$, ${\phi_2}^{\pm}$ from (2.36) satisfy the following equations:

$$\pm\sqrt{2|\mu_*|}\frac{\partial\theta_2^{\pm}}{\partial x_{\pm}} + \frac{\partial\theta_1^{\pm}}{\partial t_2} = D_1\left(\frac{\partial^2\theta_1^{\pm}}{\partial {x_{\pm}}^2} + \frac{\partial^2\theta_1^{\pm}}{\partial y^2}\right) - 2\frac{\partial\phi_2^{\pm}}{\partial x_{\pm}}, \tag{B.1}$$

$$\pm\sqrt{2|\mu_*|}\frac{\partial\phi_2^{\pm}}{\partial x_{\pm}} \mp \sqrt{2|\mu_*|}\frac{\partial\theta_1^{\pm}}{\partial t_2} = \left(1 \mp \sqrt{2|\mu_*|}D_2\right)\left(\frac{\partial^2\theta_1^{\pm}}{\partial {x_{\pm}}^2} + \frac{\partial^2\theta_1^{\pm}}{\partial y^2}\right) - 2|\mu_*|\frac{\partial\theta_2^{\pm}}{\partial x_{\pm}}, \tag{B.2}$$

which can be transformed to

$$\mp 2\sqrt{2|\mu_*|}\frac{\partial\theta_1^{\pm}}{\partial t_2} = \left[1 \mp (D_1 + D_2)\sqrt{2|\mu_*|}\right]\left(\frac{\partial^2\theta_1^{\pm}}{\partial x_{\pm}^2} + \frac{\partial^2\theta_1^{\pm}}{\partial y^2}\right). \tag{B.3}$$

Taking into account the definition of μ_* in (2.28), one finds that θ_1^+ does not depend on t_2, while θ_1^- is governed by the equation

$$\frac{\partial \theta_1^-}{\partial t_2} = (D_1 + D_2)\left(\frac{\partial^2 \theta_1^-}{\partial x_-^2} + \frac{\partial^2 \theta_1^-}{\partial y^2}\right), \tag{B.4}$$

where $(D_1 + D_2) = (A + D) > 0$. Therefore, $\theta_1^- \to 0$, $\phi_1^- \to 0$ on the time scale t_2, and one obtains from (B.1) the following problem:

$$\frac{\partial}{\partial x_+}\left[\phi_2^+ + \sqrt{2|\mu_*|}\theta_2^+\right] = D_1 \nabla^2 \theta_1^+, \tag{B.5}$$

$$\frac{\partial}{\partial x_-}\left[\phi_2^- - \sqrt{2|\mu_*|}\theta_2^-\right] = 0, \tag{B.6}$$

where $\nabla^2 \theta_1^+ = \partial^2 \theta_1^+ / \partial x_+^2 + \partial^2 \theta_1^+ / \partial y^2$. The solution of (B.6) is

$$\phi_2^- = \sqrt{2|\mu_*|}\theta_2^- + c_-(t_3, \ldots), \tag{B.7}$$

and the solution of (B.5) can be written formally as

$$\phi_2^+ = D_1 \mathcal{I}[\theta_1^+] - \sqrt{2|\mu_*|}\theta_2^+ + c_+(t_3, \ldots), \tag{B.8}$$

where

$$\mathcal{I}[f(x_+, y)] \equiv \int_0^{x_+} dx'_+ \nabla^2 f(x'_+, y). \tag{B.9}$$

In the next order, $O(\delta^4)$, one obtains

$$\theta_3 = \theta_3^+(x_+, y) + \theta_3^-(x_-, y), \qquad \phi_3 = \phi_3^+(x_+, y) + \phi_3^-(x_-, y), \tag{B.10}$$

where the functions ${\theta_3}^{\pm}$, ${\phi_3}^{\pm}$ satisfy the following equations:

$$\pm\sqrt{2|\mu_*|}\frac{\partial \theta_3^\pm}{\partial x_\pm} + \frac{\partial \phi_3^\pm}{\partial x_\pm} = D_1(\nabla^2 \theta_2^\pm) \pm \sqrt{2|\mu_*|}(\nabla^2 \theta_1^\pm)^2 - \frac{\partial \theta_1^\pm}{\partial t_3}, \tag{B.11}$$

$$\pm\sqrt{2|\mu_*|}\frac{\partial \phi_3^\pm}{\partial x_\pm} \mp \sqrt{2|\mu_*|}\frac{\partial \theta_1^\pm}{\partial t_3} + 2|\mu_*|\frac{\partial \theta_3^\pm}{\partial x_\pm} = D_2 \nabla^2 \phi_2^\pm + \nabla^2 \theta_2^\pm + 2\mu_2 \frac{\partial \theta_1^\pm}{\partial x_\pm} + 2|\mu_*|(\nabla^2 \theta_1^\pm)^2. \tag{B.12}$$

Equations (B.11) and (B.12) can be transformed to

$$\frac{\partial \theta_1^+}{\partial t_3} = \sqrt{2|\mu_*|}(\nabla \theta_1^+)^2 - \frac{\mu_2}{\sqrt{2|\mu_*|}}\frac{\partial \theta_1^+}{\partial x_+} - \frac{D_1 D_2}{2\sqrt{2|\mu_*|}}\mathcal{I}[\nabla^2 \theta_1^+] + c_+(t_3, \ldots), \tag{B.13}$$

$$\nabla^2 \theta_2^- = 0. \tag{B.14}$$

Also, one obtains from (B.11), (B.12) the following problem for $\phi_3^\pm$:

$$\frac{\partial}{\partial x_+}\left[\phi_3^+ + \sqrt{2|\mu_*|}\theta_3^+\right] = D_1 \nabla^2 \theta_2^+ + \frac{\mu_2}{\sqrt{2|\mu_*|}}\frac{\partial \theta_1^+}{\partial x_+} + \frac{D_1 D_2}{2\sqrt{2|\mu_*|}}\mathcal{I}\left[\nabla^2 \theta_1^+\right], \tag{B.15}$$

$$\frac{\partial}{\partial x_-}\left[\phi_3^- - \sqrt{2|\mu_*|}\theta_3^-\right] = 0. \tag{B.16}$$

The solutions of (B.15), (B.16) are

$$\phi_3^+ = -\sqrt{2|\mu_*|}\theta_3^+ + D_1\mathcal{I}\left[\theta_2^+\right] + \frac{\mu_2\theta_1^+}{\sqrt{2|\mu_*|}} + \frac{D_1D_2}{2\sqrt{2|\mu_*|}}\mathcal{I}\left[\nabla^2\theta_2^+\right], \tag{B.17}$$
$$\phi_3^- = \sqrt{2|\mu_*|}\theta_3^- + \tilde{c}_-(t_3,\ldots).$$

In the next order, $O(\delta^5)$, one obtains

$$\theta_4 = \theta_4^+(x_+,y) + \theta_4^-(x_-,y), \qquad \phi_4 = \phi_4^+(x_+,y) + \phi_4^-(x_-,y), \tag{B.18}$$

where the functions ${\theta_4}^{\pm}$ and ${\phi_4}^{\pm}$ satisfy the two equations:

(B.19)
$$\pm\sqrt{2|\mu_*|}\frac{\partial\theta_4^\pm}{\partial x_\pm} + \frac{\partial\phi_4^\pm}{\partial x_\pm} + \frac{\partial\theta_1^\pm}{\partial t_4} + \frac{\partial\theta_2^\pm}{\partial t_3} = D_1\left(\nabla^2\theta_3^\pm\right) - \nabla\theta_1^\pm\cdot\nabla\phi_2^\pm - \nabla\theta_2^\pm\cdot\nabla\phi_1^\pm,$$
$$\pm\sqrt{2|\mu_*|}\left(\frac{\partial\phi_4^\pm}{\partial x_\pm} - \frac{\partial\theta_1^\pm}{\partial t_4}\right) + \frac{\partial\phi_2^\pm}{\partial t_3} = D_2\nabla^2\phi_3^\pm + \nabla^2\theta_3^\pm + 2\mu_2\frac{\partial\theta_2^\pm}{\partial x_\pm}$$

(B.20)
$$-2|\mu_*|\left(\frac{\partial\theta_4^\pm}{\partial x_\pm} + \nabla\theta_1^\pm\cdot\nabla\theta_2^\pm\right) - \nabla\phi_1^\pm\cdot\nabla\phi_2^\pm\,.$$

Using (2.35), (B.7), (B.8), (B.17) one gets

$$\frac{\partial\theta_1^+}{\partial t_4} + \frac{\partial\theta_2^+}{\partial t_3} = \frac{-1}{2\sqrt{2|\mu_*|}}\left(2\mu_2\frac{\partial\theta_2^+}{\partial x_+} + D_1D_2\mathcal{I}[\nabla^2\theta_2^+] + \frac{D_1D_2}{4|\mu_*|}\mathcal{I}^2[\nabla^2\theta_1^+] + \frac{\mu_2}{2|\mu_*|}\nabla^2\theta_1^+\right)$$
$$+2\sqrt{2|\mu_*|}\nabla\theta_1^+\cdot\nabla\theta_2^+ - D_1\nabla\theta_1^+\mathcal{I}[\nabla\theta_1^+] + \frac{D_1}{2}\mathcal{I}[(\nabla\theta_1^+)^2]\,, \tag{B.21}$$

$$\frac{\partial\theta_2^-}{\partial t_3} = \frac{1}{\sqrt{2|\mu_*|}}\left(\nabla^2\theta_3^- + \mu_2\frac{\partial\theta_2^-}{\partial x_-}\right). \tag{B.22}$$

Equations (B.13) and (B.21) can be combined as the following single equation:

$$\frac{\partial\vartheta}{\partial\tau} = -\frac{\mu_2}{\sqrt{2|\mu_*|}}\frac{\partial\vartheta}{\partial x_+} - \frac{D_1D_2}{2\sqrt{2|\mu_*|}}\mathcal{I}[\nabla^2\vartheta] + \sqrt{2|\mu_*|}(\nabla\vartheta)^2$$
$$+\delta\left(-\frac{\mu_2}{4|\mu_*|\sqrt{2|\mu_*|}}\nabla^2\vartheta + \Psi) + O(\delta^2\right), \tag{B.23}$$

where

$$\Psi = -\frac{D_1D_2}{8|\mu_*|\sqrt{2|\mu_*|}}\mathcal{I}^2[\nabla^2\vartheta] + \frac{D_1}{2}\mathcal{I}[(\nabla\vartheta)^2] - D_1\nabla\vartheta\cdot\mathcal{I}[\nabla\vartheta]\,,$$
$$\mathcal{I}[f(x_+,y)] \equiv \int_0^{x_+} dx'_+\nabla^2 f(x'_+,y)\,,$$

and $\vartheta = \theta_1^+ + \delta\theta_2^+ + \ldots\,,\ \partial_\tau = \partial_{\tau_3} + \delta\partial_{\tau_4} + \ldots.$

By means of the scaling transformation $\vartheta = C_1\Theta,\ x_+ = C_2\tilde{X},\ y = C_2\tilde{Y},\ \tau = C_3\tilde{T}$, where

$$C_1 = -\frac{3\sqrt{2\mu_2D_1D_2}}{4|\mu_*|}, \qquad C_2 = \left(\frac{D_1D_2}{2\mu_2}\right)^{1/2}, \qquad C_3 = \frac{2\sqrt{2|\mu_*|D_1D_2}}{(2\mu_2)^{3/2}},$$

(B.23) can be written in the form (2.37), where $\nabla = \vec{e}_X \partial_X + \vec{e}_Y \partial_Y$, the coefficients are

$$\widetilde{C} = 1,\ \widetilde{\delta} = \delta \frac{(2|\mu_*|)^{3/2}}{6\mu_2^2},\ \widetilde{D} = -\frac{D_1}{2}\left(\frac{3\mu_2}{2|\mu_*|}\right)^2 \sqrt{\frac{2\mu_2}{D_1 D_2}},\ \widetilde{E} = D_1 \left(\frac{3\mu_2}{2|\mu_*|}\right)^2 \sqrt{\frac{2\mu_2}{D_1 D_2}},$$

and tildes in $\tilde{X}$, $\tilde{Y}$, $\tilde{\nabla}$, $\tilde{T}$ have been dropped.

Acknowledgments. E. A. Glasman acknowledges the support of the Sherman Fellowship. The authors are thankful to B. J. Matkowsky, L. M. Pismen, and A. Oron for useful discussion.

REFERENCES

[1] M. C. Cross and P. C. Hohenberg, *Pattern formation outside of equilibrium*, Rev. Modern Phys., 65 (1993), pp. 851–1112.
[2] K. Stewartson and J. T. Stuart, *Non-linear instability theory for a wave system in plane Poiseuille flow*, J. Fluid Mech., 48 (1971), pp. 529–545.
[3] A. C. Newell, *Envelope equations*, in Nonlinear Wave Motion, Lectures in Appl. Math. 15, AMS, Providence, RI, 1974, pp. 157–163.
[4] A. A. Nepomnyashchy, *Modulated wave motions arising due to the instability of spatially periodic secondary motions*, Proc. Perm State Univ., 316 (1974), pp. 105–113 (in Russian).
[5] I. S. Aranson and L. Kramer, *The world of the complex Ginzburg–Landau equation*, Rev. Modern Phys., 74 (2002), pp. 99–143.
[6] A. A. Nepomnyashchy, *Order-parameter equations for long-wavelength instabilities*, Phys. D, 86 (1995), pp. 90–95.
[7] Y. Kuramoto and T. Tsuzuki, *Persistent propagation of concentration waves in dissipative media far from thermal equilibrium*, Progr. Theoret. Phys., 55 (1976), pp. 356–369.
[8] A. A. Nepomnyashchy, *Stability of wavy regimes in a film flowing down an inclined plane*, Fluid Dynam., 9 (1974), pp. 354–359.
[9] G. M. Homsy, *Model equations for wavy viscous film flow*, in Nonlinear Wave Motion, Lectures in Appl. Math. 15, AMS, Providence, RI, 1974, pp. 191–194.
[10] T. Yamada and Y. Kuramoto, *Reduced model showing chemical turbulence*, Progr. Theoret. Phys., 56 (1976), pp. 681–683.
[11] G. I. Sivashinsky, *Non-linear analysis of hydrodynamic instability in laminar flames.* 1. *Derivation of basic equations*, Acta Astronaut., 4 (1977), pp. 1177–1206.
[12] A. A. Nepomnyashchy, *Wavy motions in a layer of viscous fluid flowing down the inclined plane*, Proc. Perm State Univ., 362 (1976), pp. 114–124 (in Russian).
[13] A. N. Garazo and M. G. Velarde, *Dissipative Korteweg–deVries description of Marangoni–Benard oscillatory convection*, Phys. Fluids A, 3 (1991), pp. 2295–2300.
[14] B. Janiaud, A. Pumir, D. Bensimon, V. Croquette, H. Richter, and L. Kramer, *The Eckhaus instability for traveling waves*, Phys. D, 55 (1992), pp. 269–286.
[15] A. C. Newell and J. A. Whitehead, *Finite bandwidth, finite amplitude convection*, J. Fluid Mech., 38 (1969), pp. 279–303.
[16] B. I. Cohen, J. A. Krommes, W. M. Tang, and M. N. Rosenbluth, *Nonlinear saturation of dissipative trapped-ion mode by mode-coupling*, Nuclear Fusion, 16 (1976), pp. 971–992.
[17] U. Frisch, Z.-S. She, and O. Thual, *Viscoelastic behavior of cellular solutions to the Kuramoto–Sivashinsky model*, J. Fluid Mech., 168 (1986), pp. 221–240.
[18] H.-C. Chang, E. A. Demekhin, and D. I. Kopelevich, *Laminarizing effects of dispersion in an active-dissipative nonlinear medium*, Phys. D, 63 (1993), pp. 299–320.
[19] D. E. Bar and A. A. Nepomnyashchy, *Stability of periodic-waves governed by the modified Kawahara equation*, Phys. D, 86 (1995), pp. 586–602.
[20] A. A. Nepomnyashchy, *Nonlinear modulational instability of periodic-waves governed by the Kuramoto–Sivashinsky equation*, Europhys. Lett., 31 (1995) pp. 437–441.
[21] M. I. Tribelskii, *Short-wavelength instability and transition to chaos in distributed systems with additional symmetry*, Uspekhi Fiz. Nauk, 167 (1997), pp. 167–190.
[22] P. C. Matthews and S. M. Cox, *Pattern formation with a conservation law*, Nonlinearity, 13 (2000), pp. 1293–1320.

[23] A. A. Golovin, B. J. Matkowsky, A. Bayliss, and A. A. Nepomnyashchy, *Coupled KS–CGL and coupled Burgers–CGL equations for flames governed by a sequential reaction*, Phys. D, 129 (1999), pp. 253–298.
[24] A. A. Golovin, A. A. Nepomnyashchy, and B. J. Matkowsky, *Traveling and spiral waves for sequential flames with translation symmetry: Coupled CGL–Burgers equations*, Phys. D, 160 (2001), pp. 1–28.
[25] B. A. Malomed, *Patterns produced by a short-wave instability in the presence of a zero mode*, Phys. Rev. A, 45 (1992), pp. 1009–1017.
[26] F. Daviaud, J. Lega, P. Berge, P. Coullet, and M. Dubois, *Spatiotemporal intermittency in a 1D convective pattern—theoretical-model and experiments*, Phys. D, 55 (1992), pp. 287–308.
[27] S. E. Esipov, *Coupled Burgers equations - A model of polydispersive sedimentation*, Phys. Rev. E, 52 (1995), pp. 3711–3718.
[28] G. M. Webb, M. Brio, G. P. Zank, and T. Story, *Wave-wave interactions in two-fluid cosmic-ray hydrodynamics*, J. Plasma Phys., 57 (1997), pp. 631–676.
[29] J. P. Nee and J. Duan, *Limit set of trajectories of the coupled viscous Burgers' equations*, Appl. Math. Lett., 11 (1998), pp. 57–61.
[30] M. V. Foursov, *On integrable coupled Burgers-type equations*, Phys. Lett. A, 272 (2000), pp. 57–64.
[31] J. Fleischer and P. H. Diamond, *Burgers' turbulence with self-consistently evolved pressure*, Phys. Rev. E, 61 (2000), pp. 3912–3925.
[32] M. L. Frankel, P. V. Gordon, and G. I. Sivashinsky, *On disintegration of near-limit cellular flames*, Phys. Lett. A, 310 (2003), pp. 389–392.
[33] B. A. Malomed, *Nonlinear waves in nonequilibrium systems of the oscillatory type, Part* I, Z. Phys. B, 55 (1984), pp. 241–248.
[34] B. A. Malomed, *Nonlinear waves in nonequilibrium systems of the oscillatory type, Part* II, Z. Phys. B, 55 (1984), pp. 249–256.
[35] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics*, Pergamon Press, Oxford, UK, 1987.
[36] I. Kliakhandler and G. Sivashinsky, *Kinetic alpha effect in viscosity stratified creeping flows*, Phys. Fluids, 7 (1995), pp. 1866–1871.
[37] U. Frisch, Z.-S. She, and P. L. Sulem, *Large-scale flow driven by the anisotropic kinetic alpha-effect*, Phys. D, 28 (1987), pp. 382–392.
[38] E. M. Lifshitz and L. P. Pitaevskii, *Physical Kinetics*, Pergamon, Oxford, UK, 1981.
[39] C. Godrèche and P. Manneville, *Hydrodynamics and Nonlinear Instabilities*, Cambridge University Press, Cambridge, UK, 1998.
[40] A. Garazo and M. G. Velarde, *Marangoni-driven solitary waves*, in Proceedings of the 8th European Symposium on Materials and Fluid Sciences in Microgravity, European Space Agency, Paris, 1999, pp. 711–715.
[41] A. A. Nepomnyashchy and M. G. Velarde, *A 3-dimensional description of solitary waves and their interaction in Marangoni–Benard layers*, Phys. Fluids, 6 (1994), pp. 187–198.
[42] A. L. Frenkel and K. Indireshkumar, *Wavy film flows down an inclined plane: Perturbation theory and general evolution equation for the film thickness*, Phys. Rev. E, 60 (1999), pp. 4143–4157.
[43] A. A. Golovin and S. H. Davis, *Effect of anisotropy on morphological instability in the freezing of a hypercooled melt*, Phys. D, 116 (1998), pp. 363–391.
[44] D. E. Bar and A. A. Nepomnyashchy, *Stability of periodic waves generated by long-wavelength instabilities in isotropic and anisotropic systems*, Phys. D, 132 (1999), pp. 411–427.
[45] S. H. Davis, *Theory of Solidification*, Cambridge University Press, Cambridge, UK, 2001.
[46] A. J. Bernoff, R. Kuske, B. J. Matkowsky, and V. A. Volpert, *Mean field effects for counterpropagating traveling wave solutions of reaction-diffusion systems*, SIAM J. Appl. Math., 55 (1995), pp. 485–519.
[47] H. Pearlman and P. Ronney, *Self organized spiral and circular waves in premixed gas flames*, J. Chem. Phys., 101 (1994), pp. 2632–2633.
[48] A. G. Merzhanov and E. N. Rumanov, *Physics of reaction waves*, Rev. Modern Phys., 71 (1999), pp. 1173–1211.
[49] I. R. Epstein and J. A. Pojman, *Overview: Nonlinear dynamics related to polymeric systems*, Chaos, 9 (1999), pp. 255–259.

SIAM J. APPL. MATH.
Vol. 65, No. 1, pp. 252–266

T-SCAN ELECTRICAL IMPEDANCE IMAGING SYSTEM FOR ANOMALY DETECTION*

HABIB AMMARI†, OHIN KWON‡, JIN KEUN SEO§, AND EUNG JE WOO¶

Abstract. We consider an inverse conductivity problem arising in anomaly detections with its mathematical model based on the T-Scan system (breast cancer detection system). In this model, we try to detect an anomaly D from one or two sets of measured data that are available only on a small portion Γ of the boundary of the subject Ω. In practice, Ω differs in each subject, so our detection algorithm should not depend much on the global geometry of Ω. The purpose of this work is to provide a mathematical ground for the reconstruction of a rough feature of D which is stable against any measurement noise and any change of geometry $\partial\Omega$. Based on rigorous estimates with a simplified model, we found an approximation that gives a noniterative detection algorithm of finding a useful feature of anomaly. We also present a multifrequency approach to handling the case where the complex conductivity of the background is not homogeneous and is not known a priori.

Key words. breast cancer detection, electrical conductivity, T-Scan, anomaly estimation algorithm

AMS subject classifications. 35R30, 34A45, 65N21, 78A30, 78A70

DOI. 10.1137/S003613990343375X

1. Introduction.

Bioimpedance techniques such as electrical impedance tomography (EIT) have been used as a diagnostic tool for breast cancer detection due to the high contrast of the complex conductivity between cancerous and healthy tissues. Among them, the T-Scan is one of the most successful systems that has received FDA approval for adjunctive clinical uses with X-ray mammography [3]. Use of the T-Scan is to decrease equivocal findings and thereby reduce unnecessary biopsies. However, diagnostic information from the currently available T-Scan system lacks a sophisticated reconstruction method of finding lesions. Increasing the ability to accurately detect breast cancer requires improvements in the sensitivity and accuracy of the T-Scan system. Although some observations on processing data from the T-Scan system have been published [3, 21], rigorous mathematical theory for supporting their results has not been presented. Systematic studies are essential to achieve higher performance in breast cancer detection, and it is necessary to derive agreements between experimental results and mathematical theory that provide the relationship between lesions and measured data acquired by a scanning probe through the breast skin.

The mathematical model of the T-Scan can be viewed essentially as a realistic or practical version of a general EIT system, so any developed theory from this model can be applied to other areas in EIT, especially in anomaly detection problems. In the T-Scan system, a patient holds in one hand a metallic cylindrical reference electrode through which a constant voltage of 1 to 2.5 V is applied with frequencies spanning

*Received by the editors August 21, 2003; accepted for publication (in revised form) February 19, 2004; published electronically October 8, 2004. This work was supported by grant R11-2002-103 from the Korea Science and Engineering Foundation.

http://www.siam.org/journals/siap/65-1/43375.html

†Centre de Mathématiques Appliquées, CNRS UMR 7641, Ecole Polytechnique, 91128 Palaiseau Cedex, France (ammari@cmapx.polytechnique.fr). The work of this author was supported by ACI Jeunes Chercheurs (0693) from the Ministry of Education and Scientific Research, France.

‡Department of Mathematics, Konkuk University, Seoul 143-701, Korea (oikwon@konkuk.ac.kr).

§Department of Mathematics, Yonsei University, Seoul 120-749, Korea (seoj@yonsei.ac.kr).

¶College of Electronics and Information, Kyung Hee University, Kyungkido, Korea (ejwoo@khu.ac.kr).

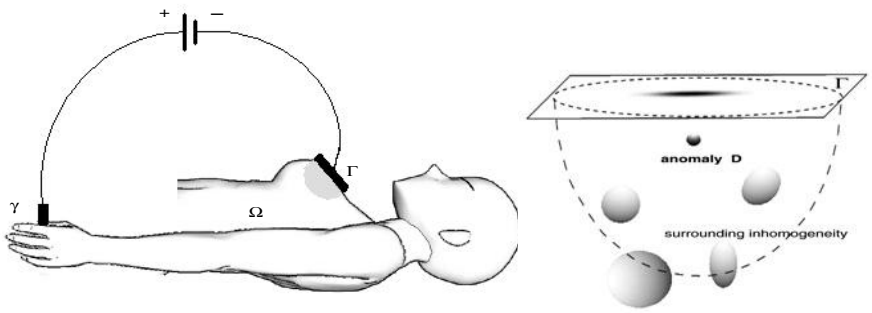


FIG. 1.1. *T-Scan configuration (left) and the breast region of interest (right).*

100 Hz to 100 KHz. A scanning probe with a planar array of electrodes kept at the ground potential is placed on the breast as shown in Figure 1.1. The voltage difference between the hand and the breast skin produces current flows from the reference electrode to each electrode of the probe through the breast. We can extract some information of the complex conductivity distribution within a breast region under the probe by measuring the exit current through each electrode of the probe.

Let the human body occupy a three-dimensional region Ω bounded by a smooth surface $\partial\Omega$. Let Γ and γ be portions of $\partial\Omega$, denoting the probe plane placed on the breast and the surface of the metallic reference electrode contacting the hand, respectively. Since Γ is a planar domain, without loss of generality, we let x_3 be the label of the axis normal to Γ so that Γ is in the plane $\{\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3 : x_3 = 0\}$. Suppose the diameter of Γ is between 4ρ and 6ρ; $\{x_3 = 0\} \cap B_{2\rho} \subset \Gamma \subset \{x_3 = 0\} \cap B_{3\rho}$, where $B_\rho := \{\mathbf{x} \in \mathbb{R}^3 \mid |\mathbf{x}| < \rho\}$ and denote $\Gamma_\rho = \Gamma \cap B_\rho$. Then the breast region of interest is the half ball $\Omega_\rho := \Omega \cap B_\rho$. If we apply a boundary voltage f with a frequency ω on $\Gamma \cup \gamma$, the resulting internal complex voltage $u(\mathbf{x})$ at the position $\mathbf{x}$ in Ω satisfies the following mixed boundary value problem:

$$\begin{cases} \nabla \cdot ((\sigma + i\omega\epsilon)\nabla u(\mathbf{x})) = 0, & \mathbf{x} \in \Omega, \\ u(\mathbf{x}) = f, & \mathbf{x} \in \Gamma \cup \gamma, \\ (\sigma + i\omega\epsilon)\nabla u(\mathbf{x}) \cdot \nu(\mathbf{x}) = 0, & \mathbf{x} \in \partial\Omega \setminus \Gamma \cup \gamma. \end{cases} \tag{1.1}$$

Here, ν is the unit outward normal vector to the boundary, $\sigma = \sigma(\mathbf{x}, \omega)$ is the conductivity, and $\epsilon = \epsilon(\mathbf{x}, \omega)$ is the permittivity. As in the T-Scan system, we use $f = 0$ on Γ and $f = 1$ to 2.5 V on γ with $\omega = 2\pi \times 10^2$ to $2\pi \times 10^5$ rad/s. The goal is to extract some core information of the complex conductivity $(\sigma + i\omega\epsilon)$ in the breast region Ω_ρ from the Cauchy data given only on the small portion Γ.

Let us denote a breast tumor by D and suppose $D \subset \Omega_\rho$. Since there is an abrupt change in the complex conductivity across ∂D, it is convenient to write

$$\sigma + i\omega\epsilon = \begin{cases} \sigma_1 + i\omega\epsilon_1 := \tau_1 & \text{in } \Omega \setminus \overline{D}, \\ \sigma_2 + i\omega\epsilon_2 := \tau_2 & \text{in } D. \end{cases} \tag{1.2}$$

Our anomaly detection problem is to identify D near Γ from the exit current data $g := (\sigma + i\omega\epsilon)\nabla u(\mathbf{x}) \cdot \nu(\mathbf{x})|_\Gamma$ under the following limitations:

- We measure the data only in a small surface Γ instead of the whole surface $\partial\Omega$.
- Since Ω differs for each subject, our detection algorithm should not depend much on the global geometry of Ω.
- Electrical safety regulations limit the amount of the total current flowing through the human subject and therefore the range of the applied voltage is also limited.

Although this type of anomaly detection problem has been studied in many papers [1, 2, 4, 5, 6, 7, 11, 12, 16, 17, 18], these limitations are indispensable in practice and raise serious difficulties in applying previous techniques. The challenge of this problem is to develop a proper analysis for a quantitative information of D in the breast region with some measured data in such a way that a reconstruction formula for D is reasonably stable to any change of the conductivity distributions outside the breast region. In this work we fix the voltage $f = 0$ on Γ as in the T-Scan system instead of applying various voltages f. Keeping $f = 0$ on Γ has a great advantage because it forces the level surface of the voltage in the breast region to be approximately parallel to the probe plane Γ and its electric field $-\nabla u$ will be in the direction perpendicular to the level surface, so more current will flow along D of which the conductivity σ_2 is much higher than the surrounding. Although one could apply many different f on Γ to acquire additional information of D, technical difficulties related with measurement noise make us hesitate to use various patterns of voltages f.

The purpose of this work is to provide a mathematical ground for reconstruction of a rough feature of D which is stable against any measurement noise and any change of geometry $\partial\Omega$. We carry out some quantitative analysis for a simplified model with a single applied voltage $f = 0$ on Γ. We relate D to $(g - g_0)$, where g_0 is the corresponding Neumann data of (1.1) in the absence of D. This analysis provides the reconstruction method of extracting a rough feature of D, although we do not know the overall structure of the complex conductivity distribution in Ω.

To end this section, let us review some previous results toward the T-Scan model. Assenheimer et al. [3] and Scholtz [21] studied the model, and their results were based on the physical insight that an anomaly can be viewed as a distribution of aligned dipoles. However, their expressions lack generality and flexibility without a rigorous mathematical theory. To perform a quantitative analysis of the T-Scan model and increase the accuracy of lesion detection, we need to develop a rigorous mathematical theory counting on the effects of the anomaly. In the recent paper by Seo et al. [22], a framework for analyzing the mathematical model of the T-Scan system was presented, providing a stable reconstruction algorithm for locating D, and numerical simulations showed that their methods can extract key features of D. However, rigorous mathematical theory has not yet been provided to support all these results and so their accuracy has not been confirmed.

We should also note that there have been different approaches to detecting breast cancers based on the current-injection EIT system. Kerner et al. placed a circular array of electrodes around the breast and produced cross-sectional conductivity images [15]. Larson-Wiseman [19], Mueller, Isaacson, and Newell [20], and Kao et al. [14] investigated the usefulness of planar electrode arrays instead of conventional circular electrode arrays. They studied the optimal injection current patterns, distinguishability, and image reconstruction algorithms using the measured voltage data on the electrode arrays. Cherepenin et al. used a planar array of 256 electrodes to reconstruct so-called electrical impedance mammograms by sequentially injecting currents through chosen electrodes and measuring voltage data on other electrodes in the array [8, 9].

Although there could be various approaches in breast cancer detection using EIT techniques, we focus on the T-Scan model in this paper. Although we deal only with the problem in three-dimensional space, all results in this paper work for general dimension $n \geq 2$ with minor modifications.

2. Anomaly detection in half space. The problem we consider in this section is a simplified version of the T-Scan model in (1.1). Let Ω be the lower half space $\Omega = \mathbb{R}^3_- := \{\mathbf{x} = (x_1, x_2, x_3) \mid x_3 < 0\}$. Let $B_r(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^3 : |\mathbf{x} - \mathbf{y}| < r\}$ and simply $B_r = B_r(0)$. Let ρ be a fixed positive number and let D be a domain contained in $\Omega_\rho = \Omega \cap B_\rho$ with connected C^2-boundary. Let Γ be a portion of the plane $\partial\Omega$ such that $\{x_3 = 0\} \cap B_{2\rho} \subset \Gamma \subset \{x_3 = 0\} \cap B_{3\rho}$ and let $\Gamma_\rho = \Gamma \cap B_\rho$.

We consider the following mixed boundary value problem:

$$\text{(2.1)} \qquad \begin{cases} \nabla \cdot ((1 + \mu\chi_D)\nabla v(\mathbf{x})) = 0, \quad \mathbf{x} \in \Omega = \mathbb{R}^3_-, \\ v(\mathbf{x}) = 1, \quad \mathbf{x} \in \Gamma, \\ \frac{\partial v}{\partial x_3}(\mathbf{x}) = 0, \quad \mathbf{x} \in \Gamma^{ext} := \partial\Omega \setminus \Gamma, \end{cases}$$

where μ is a positive constant. The $H^1(\Omega)$-solution u of (1.1) is related to the solution v of (2.1) in such a way that $u = 1 - v$ when $\tau_1 = 1$, $\tau_2 = 1 + \mu$, and $\gamma = \infty$. For related works of this simplified model, see [3] and [21].

The inverse problem is to determine D from the Neumann data $g = \frac{\partial v}{\partial x_3}|_\Gamma$. We assume that

$$\xi \in D \subset \Omega_\rho, \qquad \operatorname{diam}(D) \leq r_0, \qquad \text{and} \qquad r_0 \leq \operatorname{dist}(D, \Gamma) \leq A r_0,$$

where A is a fixed positive constant. We assume that the ratio r_0/ρ is small. Here, we exclude two possible cases; (i) $\operatorname{dist}(D, \Gamma) < r_0$ and (ii) D is far away from Γ. For the first case where D is very close to Γ, the Neumann data g manifest the presence of D that was used in the original T-Scan system, so any aid of mathematical analysis is not necessary. For the second case where D is far away from Γ, the change of the Neumann data g due to D will be negligibly small so that any analysis in this case may not have practical meaning if we consider inevitable measurement noise and ill-posedness of this inverse problem.

Let v_0 be the solution of (2.1) in the absence of D and $g_0 = \frac{\partial v_0}{\partial x_3}|_\Gamma$. We try to find a relation between D and the difference $(g - g_0)$. Recall that v and v_0 satisfy the following integral representations [13]: for $\mathbf{x} \in \Omega$,

(2.2)

$$v_0(\mathbf{x}) = -\int_\Gamma \Phi(\mathbf{x}, \mathbf{y}) g_0(\mathbf{y}) ds_{\mathbf{y}} + \int_{\Gamma \cup \Gamma^{ext}} \partial_{x_3} \Phi(\mathbf{x}, \mathbf{y})\, v_0(\mathbf{y}) ds_{\mathbf{y}},$$

$$v(\mathbf{x}) = -\int_\Gamma \Phi(\mathbf{x}, \mathbf{y}) g(\mathbf{y}) ds_{\mathbf{y}} + \int_{\Gamma \cup \Gamma^{ext}} \partial_{x_3} \Phi(\mathbf{x}, \mathbf{y}) v(\mathbf{y}) ds_{\mathbf{y}} + \int_D \mu \nabla_{\mathbf{y}} \Phi(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{y}) d\mathbf{y}.$$

The next theorem enables us to approximate

$$\frac{1}{2\mu}[g(\mathbf{x}) - g_0(\mathbf{x})] \approx \int_D \nabla_{\mathbf{y}} \frac{\partial}{\partial x_3} \Phi(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{y}) d\mathbf{y}, \quad \mathbf{x} \in \Gamma_\rho.$$

Here, $\Phi(\mathbf{x}, \mathbf{y})$ denotes the fundamental solution to Laplace's equation: $\Phi(\mathbf{x}, \mathbf{y}) = \frac{-1}{4\pi|\mathbf{x}-\mathbf{y}|}$. Throughout this section, the constant C will be different in each occurrence, but all Cs are independent of r_0, ρ, and Γ.

THEOREM 2.1. *The difference* $(g - g_0)$ *on* Γ *can be expressed as*

$$\frac{1}{2\mu}[g(\mathbf{x}) - g_0(\mathbf{x})] = \int_D \nabla_{\mathbf{y}} \left[\frac{\partial}{\partial x_3} \Phi(\mathbf{x}, \mathbf{y}) + T(\mathbf{x}, \mathbf{y}) \right] \cdot \nabla v(\mathbf{y}) d\mathbf{y}, \quad \mathbf{x} \in \Gamma_\rho, \tag{2.3}$$

where $T(\mathbf{x}, \mathbf{y})$ *satisfies the following estimate:*

$$|\nabla_{\mathbf{y}} T(\mathbf{x}, \mathbf{y})| \leq C \, \rho^{-3}, \qquad \mathbf{y} \in D, \ \mathbf{x} \in \Gamma_\rho. \tag{2.4}$$

Proof. Using (2.2) and $v = v_0$ on Γ, we have

$$\begin{aligned} v(\mathbf{x}) - v_0(\mathbf{x}) = &- \int_\Gamma \Phi(\mathbf{x}, \mathbf{y})[g - g_0](\mathbf{y}) ds_{\mathbf{y}} + \int_D \mu \nabla_{\mathbf{y}} \Phi(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{y}) d\mathbf{y} \\ &+ \int_{\Gamma^{ext}} \partial_{x_3} \Phi(\mathbf{x}, \mathbf{y})[v(\mathbf{y}) - v_0(\mathbf{y})] ds_{\mathbf{y}} \end{aligned} \tag{2.5}$$

for $\mathbf{x} \in \Omega$. Applying $\frac{\partial}{\partial x_3}$ over both sides of the identity (2.5) for $\mathbf{x} \in \Gamma$ yields

$$\begin{aligned} g(\mathbf{x}) - g_0(\mathbf{x}) = \frac{\partial}{\partial x_3}[v(\mathbf{x}) - v_0(\mathbf{x})] &= - \lim_{x_3 \to 0^-} \frac{\partial}{\partial x_3} \int_\Gamma \Phi(\mathbf{x}, \mathbf{y})[g - g_0](\mathbf{y}) ds_{\mathbf{y}} \\ &+ \frac{\partial}{\partial x_3} \int_D \mu \nabla_{\mathbf{y}} \Phi(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{y}) d\mathbf{y} + \int_{\Gamma^{ext}} (v(\mathbf{y}) - v_0(\mathbf{y})) \partial_{y_3}^2 \Phi(\mathbf{x}, \mathbf{y}) ds_{\mathbf{y}}. \end{aligned} \tag{2.6}$$

Since $\lim_{x_3 \to 0^-} \frac{\partial}{\partial x_3} \int_\Gamma \Phi(\mathbf{x}, \mathbf{y})[g - g_0](\mathbf{y}) ds_{\mathbf{y}} = -\frac{1}{2}[g - g_0](\mathbf{x})$ for $\mathbf{x} \in \Gamma$ from the trace formula for the single layer potential in [10], the identity (2.6) becomes

$$\begin{aligned} \frac{1}{2}[g(\mathbf{x}) - g_0(\mathbf{x})] = \mu \frac{\partial}{\partial x_3} &\int_D \nabla_{\mathbf{y}} \Phi(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{y}) d\mathbf{y} \\ &+ \int_{\Gamma^{ext}} (v(\mathbf{y}) - v_0(\mathbf{y})) \partial_{y_3}^2 \Phi(\mathbf{x}, \mathbf{y}) ds_{\mathbf{y}}, \quad \mathbf{x} \in \Gamma. \end{aligned} \tag{2.7}$$

Now, we investigate $[v - v_0]|_{\Gamma^{ext}}$. For a fixed $\mathbf{x} \in \mathbb{R}^3 \setminus \Gamma$, let $h(\mathbf{x}, \cdot)$ be the $H^1(\mathbb{R}^3 \setminus \Gamma)$-solution of the following:

$$\begin{cases} \Delta_{\mathbf{y}} h(\mathbf{x}, \mathbf{y}) = 0, \quad \mathbf{y} \in \mathbb{R}^3 \setminus \Gamma, \\ h(\mathbf{x}, \mathbf{y}) = \frac{1}{2\pi |\mathbf{x} - \mathbf{y}|}, \quad \mathbf{y} \in \Gamma, \\ h(\mathbf{x}, \mathbf{y}) = 0, \quad \text{as } |\mathbf{y}| \to \infty. \end{cases}$$

Let $N(\mathbf{x}, \mathbf{y})$ denote the Neumann function for the half space $\mathbb{R}^3_-$; $N(\mathbf{x}, \mathbf{y}) := \Phi(\mathbf{x}, \mathbf{y}) + \Phi(\mathbf{x}, \mathbf{y}^+)$, where $\mathbf{y}^+ = (y_1, y_2, -y_3)$ is the reflection point of $\mathbf{y}$ by the plane $\{y_3 = 0\}$. Then $\Psi(\mathbf{x}, \mathbf{y}) := N(\mathbf{x}, \mathbf{y}) + h(\mathbf{x}, \mathbf{y})$ satisfies

$$\begin{cases} \Delta_{\mathbf{x}} \Psi(\mathbf{x}, \mathbf{y}) = \delta(\mathbf{x} - \mathbf{y}), \quad \mathbf{x}, \mathbf{y} \in \Omega, \\ \Psi(\mathbf{x}, \mathbf{y}) = 0, \quad \mathbf{x} \in \Gamma, \ \mathbf{y} \in \Omega, \\ \frac{\partial}{\partial x_3} \Psi(\mathbf{x}, \mathbf{y}) = 0, \quad \mathbf{x} \in \Gamma^{ext}, \ \mathbf{y} \in \mathbb{R}^3_-, \\ \Psi(\mathbf{x}, \mathbf{y}) \to 0 \quad \text{as } |\mathbf{x} - \mathbf{y}| \to \infty. \end{cases}$$

It is easy to see from the standard argument in PDE that $\Psi(\mathbf{x}, \mathbf{y}) = \Psi(\mathbf{y}, \mathbf{x})$, $\mathbf{x}, \mathbf{y} \in \Omega$. Since $-N(\mathbf{x}, \mathbf{y}) = h(\mathbf{x}, \mathbf{y})$ for $\mathbf{y} \in \Gamma$ and $\mathbf{x} \in \Omega$, in view of the maximum principle, we have $|h(\mathbf{x}, \mathbf{y})| \leq |N(\mathbf{x}, \mathbf{y})|$ for $\mathbf{x}, \mathbf{y} \in \Omega$, and so

$$|\Psi(\mathbf{x}, \mathbf{y})| \leq 2|N(\mathbf{x}, \mathbf{y})|, \qquad \mathbf{x}, \mathbf{y} \in \Omega. \tag{2.8}$$

Using Ψ and following the process in (2.5), the difference $v - v_0$ can be expressed as

$$v(\mathbf{y}) - v_0(\mathbf{y}) = \mu \int_D \nabla_{\mathbf{z}} \Psi(\mathbf{y}, \mathbf{z}) \cdot \nabla_{\mathbf{z}} v(\mathbf{z}) d\mathbf{z}, \qquad \mathbf{y} \in \Omega.$$

Hence, the second term in (2.7) can be written as

$$\int_{\Gamma^{ext}} (v(\mathbf{y}) - v_0(\mathbf{y})) \partial_{y_3}^2 \Phi(\mathbf{x}, \mathbf{y}) ds_{\mathbf{y}} = \mu \int_D \nabla_{\mathbf{z}} T(\mathbf{x}, \mathbf{z}) \cdot \nabla v(\mathbf{z}) d\mathbf{z},$$

where

$$T(\mathbf{x}, \mathbf{z}) = \int_{\Gamma^{ext}} \Psi(\mathbf{y}, \mathbf{z}) \partial_{y_3}^2 \Phi(\mathbf{x}, \mathbf{y}) ds_{\mathbf{y}}.$$

The term $\nabla_{\mathbf{z}} T(\mathbf{x}, \mathbf{z})$ can be estimated as

$$|\nabla_{\mathbf{z}} T(\mathbf{x}, \mathbf{z})| \le \int_{\Gamma^{ext}} |\nabla_{\mathbf{z}} \Psi(\mathbf{y}, \mathbf{z})| \frac{1}{|\mathbf{x} - \mathbf{y}|^3} ds_{\mathbf{y}} \qquad \text{for } \mathbf{z} \in D, \mathbf{x} \in \Gamma^{ext}.$$

For a fixed $\mathbf{y} \in \Gamma^{ext}$, $\Psi(\mathbf{y}, \cdot)$ is harmonic in the lower half ball $\Omega_{2\rho}$ and zero on Γ, so $\Psi(\mathbf{y}, \cdot)$ has the harmonic extension $\bar{\Psi}(\mathbf{y}, \cdot)$ to the entire ball $B_{2\rho}$. Using (2.8) and the interior estimate, we have

$$\begin{aligned} |\nabla_{\mathbf{z}} \Psi(\mathbf{y}, \mathbf{z})| &= \frac{1}{|B_\rho(\mathbf{z})|} \left| \int_{B_\rho(\mathbf{z})} \nabla_{\mathbf{z}} \Psi(\mathbf{y}, \tilde{\mathbf{z}}) d\tilde{\mathbf{z}} \right| = \frac{1}{|B_\rho(\mathbf{z})|} \left| \int_{\partial B_\rho(\mathbf{z})} \nu(\tilde{\mathbf{z}}) \Psi(\mathbf{y}, \tilde{\mathbf{z}}) ds_{\tilde{\mathbf{z}}} \right| \\ &\le C \frac{1}{\rho |\mathbf{y} - \mathbf{z}|}, \qquad \mathbf{z} \in D,\ \mathbf{y} \in \Gamma^{ext}. \end{aligned}$$

Since $\Gamma^{ext} \subset \partial\Omega \setminus B_{2\rho}$, we obtain

$$|\nabla_{\mathbf{z}} T(\mathbf{x}, \mathbf{z})| \le \int_{\Gamma^{ext}} \frac{1}{\rho |\mathbf{y} - \mathbf{z}|} \frac{1}{|\mathbf{x} - \mathbf{y}|^3} ds_{\mathbf{y}} \le C \rho^{-3}, \qquad \mathbf{z} \in D,\ \mathbf{x} \in \Gamma_\rho.$$

This completes the proof. □

In the next theorem, we provide a more precise estimate by investigating $\nabla v|_D$.

THEOREM 2.2. *Let $\xi^* = (\xi_1, \xi_2, 0)$, the projection of ξ to Γ. Then*

(2.9)
$$\frac{1}{2\mu}[g(\mathbf{x}) - g_0(\mathbf{x})] = g_0(\xi^*) \int_D \frac{\partial}{\partial x_3} \frac{(\mathbf{x} - \mathbf{y}) \cdot [\mathbf{e}_3 + \mu \nabla V(\mathbf{y})]}{4\pi |\mathbf{x} - \mathbf{y}|^3} d\mathbf{y} + Err(\mathbf{x}), \quad \mathbf{x} \in \Gamma_\rho,$$

where the error term $Err(\mathbf{x})$ satisfies the estimate

$$|Err(\mathbf{x})| \le C g_0(\xi^*) |D| \left(\frac{r_0}{\rho |\mathbf{x} - \xi|^3} + \frac{1}{\rho^3} \right), \quad \mathbf{x} \in \Gamma_\rho. \tag{2.10}$$

Here, V is the $H^1(\Omega)$-solution of

$$\begin{cases} \Delta V = 0 \quad \text{in } \Omega \setminus \partial D, \\ V^+ = V^- \quad \text{on } \partial D, \\ (1 + \mu) \dfrac{\partial V^+}{\partial \nu} - \dfrac{\partial V^-}{\partial \nu} = -\nu \cdot \mathbf{e}_3 \quad \text{on } \partial D, \\ \chi_\Gamma V + (1 - \chi_\Gamma) \dfrac{\partial V}{\partial \nu} = 0 \quad \text{on } \partial\Omega, \end{cases}$$

where $V^+ = V|_{\Omega \setminus \bar{D}}$ and $V^- = V|_D$.

Proof. The v can be decomposed into

$$v(\mathbf{x}) = v_0(\mathbf{x}) + \mu g_0(\xi^*)V(\mathbf{x}) + \mu w(\mathbf{x}), \quad \mathbf{x} \in \Omega, \tag{2.11}$$

where w is the $H^1(\Omega)$-solution of

$$\begin{cases} \Delta w = 0 \quad \text{in } \Omega \setminus \bar{D}, \\ (1+\mu)\dfrac{\partial w^+}{\partial \nu} - \dfrac{\partial w^-}{\partial \nu} = [-\nabla v_0 + g_0(\xi^*)] \cdot \nu \quad \text{on } \partial D, \\ w\chi_\Gamma + \dfrac{\partial w}{\partial \nu}(1-\chi_\Gamma) = 0 \quad \text{on } \partial\Omega. \end{cases}$$

Substituting $v = v_0 + \mu g_0(\xi^*)V + \mu w$ into (2.3), the error term in (2.9) is

$$\begin{aligned} Err(\mathbf{x}) = &\int_D \nabla_{\mathbf{x}} \frac{\partial}{\partial x_3}\Phi(\mathbf{x},\mathbf{y}) \cdot [\nabla v_0(\mathbf{y}) - g_0(\xi^*)\mathbf{e}_3]d\mathbf{y} + \mu \int_D \nabla_{\mathbf{x}} \frac{\partial}{\partial x_3}\Phi(\mathbf{x},\mathbf{y}) \cdot \nabla w(\mathbf{y}) d\mathbf{y} \\ &+ \int_D \nabla_{\mathbf{y}} T(\mathbf{x},\mathbf{y}) \cdot \nabla v(\mathbf{y}) d\mathbf{y} := I_1(\mathbf{x}) + I_2(\mathbf{x}) + I_3(\mathbf{x}). \end{aligned}$$

To estimate I_1, we begin with proving the estimate

$$|\nabla v_0(\mathbf{y}) - g_0(\xi^*)\mathbf{e}_3| \le C\frac{r_0}{\rho} g(\xi^*). \tag{2.12}$$

Since $\frac{\partial v_0}{\partial x_3}$ is harmonic in Ω and $\frac{\partial v_0}{\partial x_3}|_{\partial\Omega} \ge 0$, from the maximum principle $\frac{\partial v_0}{\partial x_3} > 0$ in Ω. Let $\bar{v}_0$ be the harmonic extension of v_0 from $\Omega_{2\rho}$ to $B_{2\rho}$ across Γ in such a way that $\bar{v}_0(x_1, x_2, -x_3) = 2 - v_0(\mathbf{x})$ for $\mathbf{x} \in \Omega_{2\rho}$. It is easy to see $\frac{\partial \bar{v}_0}{\partial x_3} > 0$ in $\Omega_{2\rho}$. Now, let $\mathbf{y} \in D$ be fixed and $\mathbf{y}^* = (y_1, y_2, 0)$. Note that $\nabla v_0(\mathbf{y}^*) = \frac{\partial \bar{v}_0}{\partial y_3}(\mathbf{y}^*)\mathbf{e}_3 = g(y^*)\mathbf{e}_3$ because $v_0 = 1$ on Γ. We have

$$\begin{aligned} |\nabla v_0(\mathbf{y}) - g(\xi^*)\mathbf{e}_3| &= \left| \nabla v_0(\mathbf{y}) - \nabla v_0(\mathbf{y}^*) + \mathbf{e}_3 \left[\frac{\partial v_0}{\partial y_3}(\mathbf{y}^*) - \frac{\partial v_0}{\partial y_3}(\xi^*) \right] \right| \\ &\le C|y_3| \int_0^1 \left| \frac{\partial}{\partial y_3} \nabla v_0(t\mathbf{y} + (1-t)\mathbf{y}^*) \right| dt \\ &\quad + |\mathbf{y}^* - \xi^*| \int_0^1 \left| \frac{\partial}{\partial y_3} \nabla v_0(t\mathbf{y}^* + (1-t)\xi^*) \right| dt \\ &\le Cr_0 \left\| \nabla \frac{\partial}{\partial y_3} v_0 \right\|_{L^\infty(\Omega_\rho)} \le C\frac{r_0}{\rho} \left\| \frac{\partial}{\partial y_3} \bar{v}_0 \right\|_{L^\infty(B_{3\rho/2})}. \end{aligned} \tag{2.13}$$

In the above inequalities, we use that $|y_3|, |\mathbf{y}^* - \xi^*| < r_0$ and the standard interior estimate for the harmonic function $\nabla \frac{\partial}{\partial y_3}\bar{v}_0$. Since $\frac{\partial}{\partial y_3}\bar{v}_0 > 0$ in $B_{2\rho}$, from the Harnack inequality we have

$$\left\| \frac{\partial}{\partial y_3} \bar{v}_0 \right\|_{L^\infty(B_{3\rho/2})} \le C g_0(\xi^*),$$

and so (2.12) follows from (2.13) and the above estimate. From (2.12), we have

$$|I_1(\mathbf{x})| \le C\frac{1}{|\mathbf{x}-\xi|^3} \int_D |\nabla v_0(\mathbf{y}) - g(\xi^*)\mathbf{e}_3| d\mathbf{y} \le C g_0(\xi^*) \frac{r_0|D|}{\rho|\mathbf{x}-\xi|^3}, \quad \mathbf{x} \in \Gamma_\rho.$$

Next, we will estimate I_2. Using (2.12) and the definition of w, we have

$$(1+\mu)\int_D |\nabla w|^2 d\mathbf{y} + \int_{\Omega\setminus\bar{D}} |\nabla w|^2 d\mathbf{y} = \int_{\partial D} w[-\nabla v_0 + g_0(\xi^*)\mathbf{e}_3]\cdot \nu ds_{\mathbf{y}}$$
$$= \int_D \nabla w\cdot[\nabla v_0 - g_0(\xi^*)\mathbf{e}_3]d\mathbf{y} \le C g_0(\xi^*)\frac{r_0\sqrt{|D|}}{\rho}\left(\int_D |\nabla w|^2\right)^{1/2}$$

and so

$$\left(\int_D |\nabla w|^2\right)^{1/2} \le C g_0(\xi^*)\frac{r_0\sqrt{|D|}}{(1+\mu)\rho}. \tag{2.14}$$

Hence, we have

$$|I_2(\mathbf{x})| \le C\frac{1}{|\mathbf{x}-\xi|^3}\int_D |\nabla w| d\mathbf{y} \le C\frac{\sqrt{|D|}}{|\mathbf{x}-\xi|^3}\left(\int_D |\nabla w|^2\right)^{1/2}$$
$$\le C g(\xi^*)\frac{r_0|D|}{\rho|\mathbf{x}-\xi|^3}, \quad \mathbf{x}\in\Gamma_\rho.$$

Finally, I_3 can be estimated using (2.4) and the previous estimates (2.12) and (2.14):

$$|I_3| \le C\frac{1}{\rho^3}\int_D |\nabla v| d\mathbf{y} \le C\frac{1}{\rho^3}\int_D |\nabla v_0| + |g(\xi^*)||\nabla V| + |\nabla w| d\mathbf{y} \le C g(\xi^*)\frac{|D|}{\rho^3}.$$

This completes the proof. □

The theorem suggests an idea of reconstructing D approximately. Since D is small and $\xi\in D$, (2.9) can be expressed roughly as follows. For $\mathbf{x}\in\Gamma_\rho$,

$$\frac{1}{2\mu}[g(\mathbf{x}) - g_0(\mathbf{x})] \approx \frac{|D| g_0(\xi^*)}{4\pi|\mathbf{x}-\xi|^3}\begin{pmatrix} \dfrac{-2\xi_3^2 + (x_1-\xi_1)^2 + (x_2-\xi_2)^2}{|\mathbf{x}-\xi|^2}\left[1+\mu\dfrac{\partial V}{\partial y_3}(\xi)\right] \\ + 3\mu\dfrac{\xi_3(x_1-\xi_1)\frac{\partial V}{\partial y_1} + \xi_3(x_2-\xi_2)\frac{\partial V}{\partial y_2}}{|\mathbf{x}-\xi|^2} \end{pmatrix} \tag{2.15}$$

because, according to the estimate (2.10), the term $Err(\mathbf{x})$ does not contribute to the distribution of $\frac{1}{2\mu}[g(\mathbf{x})-g_0(\mathbf{x})]$ significantly compared with the major term described in the above approximation; $\frac{r_0}{\rho}$ is small and $\frac{1}{|\mathbf{x}-\xi|^3} \gtrsim \{\frac{r_0}{\rho|\mathbf{x}-\xi|^3} + \frac{1}{\rho^3}\}$ for $\mathbf{x}\in\Gamma_\rho$ and $|\mathbf{x}-\xi| < \rho/2$.

Next, we will roughly estimate V. Suppose that $D = B_r(\xi)$ and $r < r_0$. Define

$$\tilde{V}(\mathbf{y}) = \frac{1}{3+\mu}\int_D \frac{z_3 - y_3}{4\pi|\mathbf{z}-\mathbf{y}|^3}\, d\mathbf{z},$$

which satisfies the following properties:

$$\tilde{V}(\mathbf{y}) = \frac{1}{3+\mu}(\xi_3 - y_3) \quad \text{for } \mathbf{y}\in D, \qquad \tilde{V}(\mathbf{y}) = \frac{r^3}{3+\mu}\frac{\xi_3 - y_3}{|\xi-\mathbf{y}|^3} \quad \text{for } \mathbf{y}\in\mathbb{R}^3\setminus\bar{D},$$
$$(1+\mu)\frac{\partial\tilde{V}^+}{\partial\nu}(\mathbf{y}) - \frac{\partial\tilde{V}^-}{\partial\nu}(\mathbf{y}) = -\nu\cdot\mathbf{e}_3 \quad \text{for } \mathbf{y}\in\partial D,$$

where $\tilde{V}^+ = V|_D$ and $\tilde{V}^- = V|_{\mathbb{R}^3\setminus\bar{D}}$ (see [18]). Hence, V can be decomposed into $V(\mathbf{y}) = \tilde{V}(\mathbf{y}) + U(\mathbf{y})$, where U satisfies

$$\nabla((1+\mu\chi_D)\nabla U) = 0 \qquad \text{in } \Omega, \qquad U|_\Gamma = \frac{r^3}{3+\mu}\frac{-\xi_3}{|\mathbf{y}-\xi|^3}.$$

Then U can be expressed as

$$U(\mathbf{y}) = U_0(\mathbf{y}) + \mu\int_D \nabla_{\mathbf{z}}[\Phi(\mathbf{y},\mathbf{z}) - \Phi(\mathbf{y}^+,\mathbf{z})]\cdot\nabla U(\mathbf{z})d\mathbf{z}, \quad \mathbf{y}\in\Omega, \tag{2.16}$$

where U_0 is the $H^1(\Omega)$-solution of the Dirichlet problem:

$$\Delta U_0 = 0 \quad \text{in } \Omega, \qquad U_0|_{\partial\Omega} = U|_{\partial\Omega}.$$

From the boundary condition $U_0|_\Gamma$, U_0 can be approximated as

$$U_0(\mathbf{y}) \approx \frac{r^3}{3+\mu}\frac{-y_3-\xi_3}{|\mathbf{y}^+-\xi|^3}, \quad \mathbf{y}\in\Omega_\rho,$$

and so $\nabla U_0(\xi) \approx \frac{r^3}{(3+\mu)16\pi|\xi_3|^3}\mathbf{e}_3$. Therefore, in view of (2.16) we may approximate

$$\nabla V(\xi) \approx \frac{\mathbf{e}_3}{3+\mu}\left(-1+\frac{r^3}{16\pi|\xi_3|^3}\right). \tag{2.17}$$

This leads to the approximation

$$\frac{1}{2\mu}[g(\mathbf{x})-g_0(\mathbf{x})] \approx \frac{|D|g_0(\xi^*)}{3+\mu}\left(1-\frac{r^3}{16\pi|\xi_3|^3}\right)\frac{2\xi_3^2-(x_1-\xi_1)^2-(x_2-\xi_2)^2}{4\pi|\mathbf{x}-\xi|^5}, \quad \mathbf{x}\in\Gamma_\rho,$$

or simply

$$\frac{1}{2}[g(\mathbf{x})-g_0(\mathbf{x})] \approx \frac{\mu|D|g(\xi^*)}{3+\mu}\,\frac{2\xi_3^2-(x_1-\xi_1)^2-(x_2-\xi_2)^2}{4\pi|\mathbf{x}-\xi|^5}, \quad \mathbf{x}\in\Gamma_\rho, \tag{2.18}$$

since $\frac{r^3}{16\pi|\xi_3|^3} < \frac{1}{16\pi}$ is small. This approximation gives an accurate reconstruction algorithm for finding the location ξ and the size $|D|$. This will be discussed in the following sections.

3. Anomaly detection algorithm in the T-Scan model. In this section, based on the mathematical analysis for the simplified model given in the previous section, we will derive a rough detection algorithm for the T-Scan model. With a realistic model, the background complex conductivity τ_1 may not be homogeneous, and we met many technical difficulties in carrying out rigorous mathematical analysis that requires various approximations. We will not do all the analysis with the complicated model but will suggest several desired estimates to the reader.

Let Ω be a bounded domain in $\mathbb{R}^3$ with a smooth boundary $\partial\Omega$. The real and imaginary parts of complex conductivities $\tau_1 = \sigma_1 + i\omega\epsilon_1$ and $\tau_2 = \sigma_2 + i\omega\epsilon_2$ are positive and bounded. As in the previous section, let $\Gamma \subset \partial\Omega$ be a smooth planar domain lying on the plane $\partial\mathbb{R}^3_-$ such that $B_{2\rho}(0)\cap\{x_3=0\} \subset \Gamma \subset B_{3\rho}(0)$, where $B_\rho(\mathbf{z}) = \{\mathbf{x}\in\mathbb{R}^3 : |\mathbf{x}-\mathbf{z}| < \rho\}$. We denote $\Omega_\rho = \Omega\cap B_\rho(0)$ and $\Gamma_\rho = \Gamma\cap B_\rho(0)$. As before, assume that D is a simply connected smooth domain and

$$\xi\in D \subset \Omega_\rho, \qquad \operatorname{diam}(D)\le r_0, \qquad \text{and} \qquad r_0 \le \operatorname{dist}(D,\Gamma) \le Ar_0.$$

The mathematical model of the T-Scan system is the following:

$$\text{(3.1)}\qquad \begin{cases} \nabla\cdot(\tau_1+(\tau_2-\tau_1)\chi_D)\nabla u(\mathbf{r}))=0, & \mathbf{r}\in\Omega,\\ u(\mathbf{r})=0, & \mathbf{r}\in\Gamma,\\ u(\mathbf{r})=1, & \mathbf{r}\in\gamma,\\ \tau_1\nabla u(\mathbf{r})\cdot\nu(\mathbf{r})=0, & \mathbf{r}\in\partial\Omega\setminus\Gamma\cup\gamma. \end{cases}$$

For a fixed frequency ω, let u be the $H^1(\Omega)$-solution of (3.1) and $g=\frac{\partial u}{\partial z}|_\Gamma$, $f=u|_{\partial\Omega}$. Our goal is to reconstruct D from g.

To derive a reconstruction algorithm, we need to introduce layer potentials over a smooth surface Υ. The single and double layer potentials are defined by

$$\mathcal{S}_\Upsilon[\varphi](\mathbf{x})=\int_\Upsilon \Phi(\mathbf{x},\mathbf{y})\varphi(\mathbf{y})ds_{\mathbf{y}},\quad \mathbf{x}\in\mathbb{R}^3,$$
$$\mathcal{D}_\Upsilon[\varphi](\mathbf{x}):=\int_\Upsilon \nu(\mathbf{y})\cdot\nabla_{\mathbf{y}}\Phi(\mathbf{x},\mathbf{y})\varphi(\mathbf{y})\,ds_{\mathbf{y}},\quad \mathbf{y}\in\mathbb{R}^3\setminus\Upsilon.$$

Suppose that u_0 is the solution of (3.1) in the absence of D and $g_0=\frac{\partial u_0}{\partial x_3}|_\Gamma$. It must be noted that we cannot compute u_0 explicitly because the background complex conductivity τ_1 is unknown. However, let us begin by finding a relation between D and $(g-g_0)$.

THEOREM 3.1. *Let $\Lambda_{2\rho}=\partial\Omega_{2\rho}\setminus\Gamma$. Then*

$$\text{(3.2)}\qquad \frac{1}{2}[g(\mathbf{x})-g_0(\mathbf{x})]=\frac{\partial}{\partial x_3}\int_D(\tau_2-\tau_1)\nabla_{\mathbf{y}}\Phi(\mathbf{x},\mathbf{y})\cdot\nabla u(\mathbf{y})d\mathbf{y}+E_1(\mathbf{x})+E_2(\mathbf{x}),\quad \mathbf{x}\in\Gamma_{2\rho},$$

where

$$E_1(\mathbf{x})=\frac{\partial}{\partial x_3}\int_{\Omega_{2\rho}}[u(\mathbf{y})-u_0(\mathbf{y})]\nabla_{\mathbf{y}}\Phi(\mathbf{x},\mathbf{y})\cdot\nabla\tau_1(\mathbf{y})d\mathbf{y},$$
$$\text{and } E_2(x)=\frac{\partial}{\partial x_3}\mathcal{S}_{\Lambda_{2\rho}}\left[\frac{\partial(u-u_0)}{\partial\nu}\right](\mathbf{x})-\frac{\partial}{\partial x_3}\mathcal{D}_{\Lambda_{2\rho}}[\tau_1(u-u_0)](\mathbf{x}).$$

Moreover,

$$E_1(\mathbf{x})\le C\|\nabla\tau_1\|_{L^p(\Omega_{2\rho})}\quad \text{for } \mathbf{x}\in\Gamma_{2\rho},$$

where $p>3$.

Proof. For $\mathbf{x}\in\Omega_{2\rho}$, we have

$$\begin{aligned}\tau_1u(\mathbf{x})&=-\int_{\Omega_{2\rho}}\nabla_{\mathbf{y}}\Phi(\mathbf{x},\mathbf{y})\cdot\nabla[\tau_1u(\mathbf{y})]d\mathbf{y}+\mathcal{D}_{\Lambda_{2\rho}}[\tau_1\,u](\mathbf{x})\\ &=-\int_{\Omega_{2\rho}}\tau_1(\mathbf{y})\nabla_{\mathbf{y}}\Phi(\mathbf{x},\mathbf{y})\cdot\nabla u(\mathbf{y})d\mathbf{y}+\mathcal{D}_{\partial\Omega_{2\rho}}[\tau_1\,u](\mathbf{x})\\ &\quad+\int_{\Omega_{2\rho}}\nabla_{\mathbf{y}}\Phi(\mathbf{x},\mathbf{y})\cdot\nabla\tau_1(\mathbf{y})\,u(\mathbf{y})d\mathbf{y}\\ &=-\int_\Omega[\tau_1+(\tau_2-\tau_1)\chi_D]\nabla_{\mathbf{y}}\Phi(\mathbf{x},\mathbf{y})\cdot\nabla u(\mathbf{y})d\mathbf{y}+\int_D[\tau_2-\tau_1]\nabla_{\mathbf{y}}\Phi(\mathbf{x},\mathbf{y})\cdot\nabla u(\mathbf{y})d\mathbf{y}\\ &\quad+\mathcal{D}_{\partial\Omega_{2\rho}}[\tau_1\,u](\mathbf{x})+\int_{\Omega_{2\rho}}\nabla_{\mathbf{y}}\Phi(\mathbf{x},\mathbf{y})\cdot\nabla\tau_1(\mathbf{y})\,u(\mathbf{y})d\mathbf{y}\end{aligned}$$

$$= -\mathcal{S}_\Gamma g(\mathbf{x}) - \mathcal{S}_{\Lambda_{2\rho}}\left[\tau_1 \frac{\partial u}{\partial \nu}\right](\mathbf{x}) + \int_D [\tau_2 - \tau_1]\nabla_{\mathbf{y}}\Phi(\mathbf{x},\mathbf{y})\cdot\nabla u(\mathbf{y})d\mathbf{y}$$
$$+ \mathcal{D}_{\partial\Omega_{2\rho}}[\tau_1\, u](\mathbf{x}) + \int_{\Omega_{2\rho}} \nabla_{\mathbf{y}}\Phi(\mathbf{x},\mathbf{y})\cdot\nabla\tau_1(\mathbf{y})\, u(\mathbf{y})d\mathbf{y}.$$

Similarly, u_0 can be expressed as

$$\tau_1 u_0(\mathbf{x}) = -\mathcal{S}_\Gamma g_0(\mathbf{x}) - \mathcal{S}_{\Lambda_{2\rho}}\left[\tau_1 \frac{\partial u_0}{\partial \nu}\right](\mathbf{x}) + \mathcal{D}_{\partial\Omega_{2\rho}}[\tau_1\, u_0](\mathbf{x}) + \int_{\Omega_{2\rho}} \nabla_{\mathbf{y}}\Phi(\mathbf{x},\mathbf{y})\cdot\nabla\tau_1(\mathbf{y})\, u_0(\mathbf{y})d\mathbf{y}, \qquad \mathbf{x}\in\Omega_{2\rho}.$$

Since $(u-u_0)=0$ on Γ, for $\mathbf{x}\in\Omega_{2\rho}$

$$\tau_1[u(\mathbf{x})-u_0(\mathbf{x})] = \mathcal{S}_\Gamma[g_0-g](\mathbf{x}) + \int_D [\tau_2-\tau_1]\nabla_{\mathbf{y}}\Phi(\mathbf{x},\mathbf{y})\cdot\nabla u(\mathbf{y})d\mathbf{y}$$
$$+ \mathcal{S}_{\Lambda_{2\rho}}\left[\tau_1\frac{\partial(u_0-u)}{\partial\nu}\right](\mathbf{x}) + \mathcal{D}_{\Lambda_{2\rho}}[\tau_1(u-u_0)](\mathbf{x})$$
$$+ \int_{\Omega_{2\rho}} \nabla_{\mathbf{y}}\Phi(\mathbf{x},\mathbf{y})\cdot\nabla\tau_1(\mathbf{y})\,[u(\mathbf{y})-u_0(\mathbf{y})]d\mathbf{y}.$$

Hence, by applying $\frac{\partial}{\partial x_3}$ through the above identity at $\mathbf{x}\in\Gamma_{2\rho}$ and using the standard trace formula for single layer potentials, we have

$$\frac{\partial}{\partial x_3}[\tau_1(u(\mathbf{x})-u_0(\mathbf{x})] = \frac{1}{2}[g(\mathbf{x})-g_0(\mathbf{x})] + \frac{\partial}{\partial x_3}\int_D [\tau_2-\tau_1]\nabla_{\mathbf{y}}\Phi(\mathbf{x},\mathbf{y})\cdot\nabla u(\mathbf{y})d\mathbf{y}$$
$$+ E_1(\mathbf{x}) + E_2(\mathbf{x}), \qquad \mathbf{x}\in\Gamma_{2\rho}.$$

Moreover, for $\mathbf{x}\in\Gamma_{2\rho}$,

$$\frac{\partial}{\partial x_3}[\tau_1(u(\mathbf{x})-u_0(\mathbf{x})] = \frac{\partial\tau_1}{\partial x_3}(u(\mathbf{x})-u_0(\mathbf{x})) + [g(\mathbf{x})-g_0(\mathbf{x})] = [g(\mathbf{x})-g_0(\mathbf{x})].$$

The E_1 can be estimated by

$$|E_1(\mathbf{x})| \le \int_{\Omega_{2\rho}} \frac{|u(\mathbf{y})-u_0(\mathbf{y})|}{|\mathbf{x}-\mathbf{y}|^3}|\nabla\tau_1(\mathbf{y})|d\mathbf{y}, \qquad \mathbf{x}\in\Gamma_{2\rho}.$$

Since $\frac{u(\mathbf{y})}{y_3}$ and $\frac{u_0(\mathbf{y})}{y_3}$ are bounded in $\Omega_{2\rho}$ due to $u=0=u_0$ on Γ,

$$|E_1(\mathbf{x})| \le C\|\nabla\tau_1\|_{L^p(\Omega_{2\rho})},$$

where $p>3$ and C depends only on p and $\|\frac{u-u_0}{y_3}\|_{L^\infty(\Omega_{2\rho})}$. This completes the proof. □

In our reconstruction algorithm, we will neglect the terms $E_1(\mathbf{x})$ and $E_2(\mathbf{x})$ in the identity (3.2). Since the background complex conductivity τ_1 is a small perturbation of a constant inside the region $\Omega_{2\rho}$, $\|\nabla\tau_1\|_{L^2(\Omega_{2\rho})}$ is small and so is $E_1(\mathbf{x})$ for $\mathbf{x}\in\Gamma_\rho$. The term $E_2(x) = \frac{\partial}{\partial x_3}\mathcal{S}_{\Lambda_{2\rho}}[\frac{\partial(u-u_0)}{\partial\nu}](\mathbf{x}) - \frac{\partial}{\partial x_3}\mathcal{D}_{\Lambda_{2\rho}}[\tau_1(u-u_0)](\mathbf{x})$ is a sum of integral over $\Lambda_{2\rho}$ that sustains a distance larger than ρ from the observation point $\mathbf{x}\in\Gamma_\rho$. Furthermore, the difference $(u-u_0)|_{\Lambda_{2\rho}}$ is small, and so $E_2(\mathbf{x}), \mathbf{x}\in\Gamma_\rho$ can be viewed as a negligibly small term. Hence, the identity (3.2) can be approximated as

$$\frac{1}{2}[g(\mathbf{x})-g_0(\mathbf{x})] \approx \frac{\partial}{\partial x_3}\int_D (\tau_2-\tau_1)\nabla_{\mathbf{y}}\Phi(\mathbf{x},\mathbf{y})\cdot\nabla u(\mathbf{y})d\mathbf{y}, \quad \mathbf{x}\in\Gamma_\rho, \tag{3.3}$$

which is essentially the same as (2.15).

Under the assumption that τ_1 and τ_2 are small perturbations of two constants, we have the following approximation that corresponds to the approximation (2.18):

$$\frac{1}{2}[g(\mathbf{x}) - g_0(\mathbf{x})] \approx \frac{3\alpha(\tau_1 - \tau_2)}{2\tau_1 + \tau_2}|D|\frac{2\xi_3^2 - (x_1 - \xi_1)^2 - (x_2 - \xi_2)^2}{4\pi|\mathbf{x} - \xi|^5}, \quad \mathbf{x} \in \Gamma_\rho,$$

where $\alpha = \frac{1}{|\Gamma_\rho|}\int_{\Gamma_\rho} g$ denotes the average of g over Γ_ρ.

This rough estimate leads us to derive the following anomaly detection algorithm via simple elementary algebra:

- *Transversal position.* The anomaly D lies below the point ξ^* at which the absolute value $|g(\mathbf{x}^*) - g_0(\mathbf{x}^*)|$ has the greatest quantity:

$$|g(\xi^*) - g_0(\xi^*)| = \max_{\mathbf{x}\in\Gamma_\rho} |g(\mathbf{x}) - g_0(\mathbf{x})|. \tag{3.4}$$

- *Depth.* Let $\mathbf{x}_0$ be any chosen point in Γ_ρ near ξ^* and let l be the distance between ξ^* and $\mathbf{x}_0$, that is, $l = |\xi^* - \mathbf{x}_0|$. The depth d is determined by the identity

$$\left| \frac{g(\xi^*) - g_0(\xi^*)}{g(\mathbf{x}_0) - g_0(\mathbf{x}_0)} \right| = \frac{|2 - \frac{l^2}{d^2}|}{2(\frac{l^2}{d^2} + 1)^{5/2}}. \tag{3.5}$$

Remark 3.2 (multifrequency). We often do not have a priori knowledge of the background complex conductivity τ_1, and so we cannot compute the data g_0. In this case, we may use more than two different frequencies for the detection algorithm provided that the frequency dependencies of conductivity and permittivity values for background and anomaly are significantly different. Suppose $\tilde{\omega}$ is a frequency such that the corresponding $\tilde{\tau}_2$ is quite different from τ_2, while $\tilde{\tau}_1$ is close to τ_1. Let $\tilde{u}$ be the solution of (3.1) for the frequency $\tilde{\omega}$ and let $\tilde{g} = \frac{\partial u}{\partial \nu}|_\Gamma$. Suppose now that the difference $\tau_2 - \tilde{\tau}_2$ in cancerous region D is much larger than the difference $\tau_1 - \tilde{\tau}_1$ in the normal region. Then the expression corresponding to (2.18) is

$$\frac{1}{2}[\tilde{g}(\mathbf{x}) - g(\mathbf{x})] \approx \frac{9\alpha\tau_1(\tau_2 - \tilde{\tau}_2)}{(2\tau_1 + \tau_2)(2\tau_1 + \tilde{\tau}_2)}|D|\frac{2\xi_3^2 - (x_1 - \xi_1)^2 - (x_2 - \xi_2)^2}{4\pi|\mathbf{x} - \xi|^5}, \quad \mathbf{x} \in \Gamma_\rho.$$

This rough estimate leads to the similar anomaly detection formula as (3.4) and (3.5).

4. Numerical experiments. In this section, we present results of numerical simulations using the anomaly estimation algorithm derived in the previous section. We set the scanning probe region as $\Gamma = [-25, 25] \times [-25, 25] \times \{0\}$ mm^2 including 16×16 electrodes. The total amount of current through all 256 electrodes is assumed to be 0.2 mA. The inhomogeneous background complex conductivity τ_1 is set as

$$\tau_1 = \begin{cases} 0.005 + 100w\epsilon_0 i & \text{if } \mathbf{x} \in E\ [(-12.5, -12.5, -35); (7.5, 5, 5)] \text{ in mm}, \\ 0.01 + 500w\epsilon_0 i & \text{if } \mathbf{x} \in E\ [(12.5, 12.5, -30); (5, 7.5, 5)] \text{ in mm}, \\ 0.3 + 40000w\epsilon_0 i & \text{if } \mathbf{x} \in E\ [(-12.5, 12.5, -40); (7.5, 7.5, 7.5)] \text{ in mm}, \\ 0.05 + 30000w\epsilon_0 i & \text{if } \mathbf{x} \in E\ [(0, -2.5, -32.5); (7.5, 2.5, 2.5)] \text{ in mm}, \\ 0.003 + 800w\epsilon_0 i & \text{otherwise}, \end{cases}$$

where $\omega = 2\pi \times 10^3$ rad/s, $\epsilon_0 = 8.854 \times 10^{-12}$ F/m is the permittivity of the free space, and $E\ [\mathbf{z}; (\lambda_1, \lambda_2, \lambda_3)] := \left\{\mathbf{x} = (x_1, x_2, x_3) \mid \sum_{i=1}^3 \left(\frac{x_j - z_j}{\lambda_j}\right)^2 \le 1\right\}$. We set the anomaly D to be reconstructed as

$$D = E[\mathbf{z}_0; (2.5, 2.5, 2.5)] \text{ in mm}$$

with its center at $\mathbf{z}_0$ and complex conductivity $\tau_2 = 0.2 + 2000\omega\epsilon_0 i$.

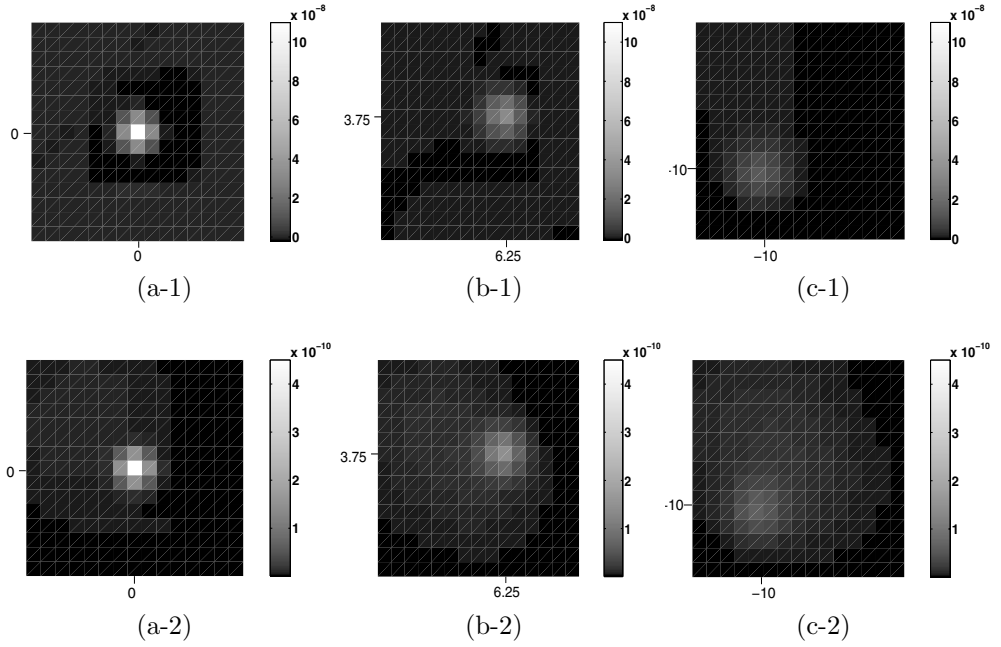


FIG. 4.1. *Plots* (a-1,2), (b-1,2), *and* (c-1,2) *are real and imaginary parts of* $(g - g_0)$ *in mA corresponding to* D *having three different centers* $\mathbf{z} = (0, 0, -5)$, $(6.25, 3.75, -7.5)$, $(-10, -10, -10)$ *in mm, respectively.*

TABLE 4.1
Performance of the proposed detection algorithm.

True center of $D(:= \mathbf{z})$	(0,0,-5)	(6.25,3.75,-7.5)	(-10,-10,-10)
Detected center of $D(:= \mathbf{z}^*)$	(0,0,-4.73)	(6.65,3.36,-7.05)	(-10,-10,-9.50)
Error$(:= \| \mathbf{z} - \mathbf{z}^* \|)$	0.27	0.72	0.5

Since we do not know the exact background complex conductivity τ_1 in practice, we compute g_0 by solving (3.1) in the absence of D with the homogeneous background $\hat{\tau}_1 = 0.003 + 800w\epsilon_0 i$ instead of the inhomogeneous τ_1. We compute g from (3.1) in the presence of D with the inhomogeneous τ_1. Notice that $(g - g_0)$ contains some unavoidable noise due to the unknown inhomogeneity $(\tau_1 - \hat{\tau}_1)$. Now we try to detect D from the difference $(g - g_0)$.

Figure 4.1 shows the real and imaginary parts of $(g - g_0)$ with D having three different centers $\mathbf{z} = (0, 0, -5), (6.25, 3.75, -7.5)$, and $(-10, -10, -10)$ in mm. First we determine the transversal position (ξ^*) using (3.4). For the depth estimate of the anomaly, we can choose any point $\mathbf{x}_0$ near ξ^*. However, if we choose a point far away from ξ^*, the difference $|g(\xi^*) - g_0(\xi^*)|$ may not be distinguished from noise. Therefore, we select four different nearest electrodes around ξ^* for the choice of the point $\mathbf{x}_0$ in (3.5) and take the average value to determine the depth using (3.5). Table 4.1 shows the results of the proposed detection algorithm.

We now test the noise tolerance of the algorithm by adding a random noise on g. We generate noised data g_n by

$$g_n = g + \| g \|_2 * NL * RN,$$

where $\| g \|_2$ is the L^2 norm of g on Γ_ρ, NL is a noise level, and RN is a ran-

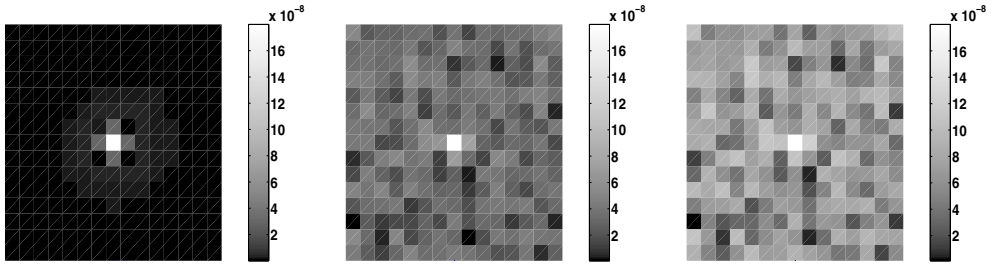


FIG. 4.2. *Noised data* $(g-g_n)$*:* (a) *noise* 0%, (b) *noise* 0.5%, *and* (c) *noise* 1% *for the anomaly* $D_1 = E[(0,0,-3.75);(2.5,2.5,2.5)]$ *in mm.*

TABLE 4.2
Detection of centers of anomalies with noised data.

Noise level	0%	0.5%	1%
Detected center of D_1	(0,0, -3.62)	(0, 0, -4.40)	(0, 0, -5.18)
Detected center of D_2	(0,0, -4.73)	(0, 0, -6.34)	(0, 0, -7.07)
Detected center of D_3	(0,0, -5.88)	(0, 0, -7.74)	(0, 0, -9.85)

dom number uniformly distributed on the interval $(-1,1)$. For the anomaly $D_1 = E[(0,0,-3.75);(2.5,2.5,2.5)]$ in mm, noised data $(g-g_n)$ with 0.5% and 1% noise are shown in Figure 4.2(b) and (c), respectively. We also apply the algorithm for anomalies $D_2 = E[(0,0,-5);(2.5,2.5,2.5)]$ and $D_3 = E[(0,0,-6.25);(2.5,2.5,2.5)]$ in mm with different depths from D_1. The detected centers of D_1, D_2, and D_3 are summarized in Table 4.2. From the results in Tables 4.1 and 4.2, we can see that the performance of the algorithm is quite good in terms of both transversal position and depth estimates. Experimental verification of these results will be a part of our future studies.

REFERENCES

[1] H. AMMARI, S. MOSKOW, AND M. VOGELIUS, *Boundary integral formulas for the reconstruction of electromagnetic imperfections of small diameter*, ESAIM Control Optim. Calc. Var., 9 (2003), pp. 49–66.

[2] H. AMMARI AND J. K. SEO, *An accurate formula for the reconstruction of conductivity inhomogeneity*, Adv. in Appl. Math., 30 (2003), pp. 679–705.

[3] M. ASSENHEIMER, O. LAVER-MOSKOVITZ, D. MALONEK, D. MANOR, U. NAHLIEL, R. NITZAN, AND A. SAAD, *The T-Scan technology: Electrical impedance as a diagnostic tool for breast cancer detection*, Physiol. Meas., 22 (2001), pp. 1–8.

[4] M. BRÜHL AND M. HANKE, *Numerical implementation of two non-iterative methods for locating inclusions by impedance tomography*, Inverse Problems, 16 (2000), pp. 1029–1042.

[5] K. BRYAN, *Numerical recovery of certain discontinuous electrical conductivities*, Inverse Problems, 7 (1991), pp. 827–840.

[6] D. J. CEDIO-FENGYA, S. MOSKOW, AND M. VOGELIUS, *Identification of conductivity imperfections of small parameter by boundary measurements. Continuous dependence and computational reconstruction*, Inverse Problems, 14 (1998), pp. 553–595.

[7] M. CHENEY, D. ISAACSON, AND J. C. NEWELL, *Electrical impedance tomography*, SIAM Rev., 41 (1999), pp. 85–101.

[8] V. CHEREPENIN, A. KARPOV, A. KORJENEVSKY, V. KORNIENKO, A. MAZALETSKAYA, D. MAZOUROV, AND D. MEISTER, *A 3D electrical impedance tomography (EIT) system for breast cancer detection*, Physiol. Meas., 22 (2001), pp. 9–18.

[9] V. CHEREPENIN, A. KARPOV, A. KORJENEVSKY, V. KORNIENKO, Y. KULTIASOV, M. OCHAPKIN, O. TROCHANOVA, AND J. MEISTER, *Three-dimensional EIT imaging of breast tissues:*

System design and clinical testing, IEEE Trans. Med. Imag., 21 (2002), pp. 662–667.

[10] G. B. Folland, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1976.

[11] A. Friedman and M. Vogelius, *Identification of small inhomogeneities of extreme conductivity by boundary measurements: A theorem on continuous dependence*, Arch. Rational Mech. Anal., 105 (1989), pp. 299–326.

[12] M. Ikehata, *On reconstruction in the inverse conductivity problem with one measurement*, Inverse Problems, 16 (2000), pp. 785–793.

[13] H. Kang and J. K. Seo, *Layer potential technique for the inverse conductivity problem*, Inverse Problems, 12 (1996), pp. 267–278.

[14] T. Kao, J. C. Newell, G. J. Saulinier, and D. Isaacson, *Distinguishability of inhomogeneities using planar electrode arrays and different patterns of applied excitation*, Physiol. Meas., 24 (2003), pp. 403–411.

[15] T. E. Kerner, K. D. Paulsen, A. Hartov, S. K. Soho, and S. P. Poplack, *Electrical impedance spectroscopy of the breast: Clinical imaging results in* 26 *subjects*, IEEE Trans. Med. Imag., 21 (2002), pp. 638–645.

[16] O. Kwon and J. K. Seo, *Total size estimation and identification of multiple anomalies in the inverse conductivity problem*, Inverse Problems, 17 (2001), pp. 59–75.

[17] O. Kwon, J. R. Yoon, J. K. Seo, E. J. Woo, and Y. G. Cho, *Estimation of anomaly location and size using electrical impedance tomography*, IEEE Trans. Biomed. Engrg., 50 (2003), pp. 89–96.

[18] O. Kwon, J. K. Seo, and J. R. Yoon, *A real-time algorithm for the location search of discontinuous conductivites with one measurement*, Comm. Pure Appl. Math., 55 (2002), pp. 1–29.

[19] J. L. Larson-Wiseman, *Early Breast Cancer Detection Utilizing Clustered Electrode Arrays in Impedance Imaging*, Ph.D. Thesis, Rensselaer Polytechnic Institute, Troy, NY, 1998.

[20] J. L. Mueller, D. Isaacson, and J. C. Newell, *A reconstruction algorithm for electrical impedance tomograpy data collected on rectangular electrode arrays*, IEEE Trans. Biomed. Engrg., 46 (1999), pp. 1379–1386.

[21] B. Scholz, *Towards virtual electrical breast biopsy: Space-frequency MUSIC for trans-admittance data*, IEEE Trans. Med. Imag., 21 (2002), pp. 588–595.

[22] J. K. Seo, O. Kwon, H. Ammari, and E. J. Woo, *Mathematical framework and anomaly estimation algorithm for breast cancer detection: Electrical impedance technique using TS*2000 *configuration*, IEEE Trans. Biomed. Eng., to appear.

SIAM J. APPL. MATH.
Vol. 65, No. 1, pp. 267–298

RESONANCE TONGUE INTERACTION IN THE PARAMETRICALLY EXCITED COLUMN*

A. R. CHAMPNEYS† AND W. B. FRASER‡

Abstract. This paper concerns a codimension-two analysis of the interaction between various resonances that occur in an upright flexible rod subject to sinusoidal parametric excitation. Particular attention is paid to rods that are just longer than their critical length for self-weight buckling, and their possible stabilization by the excitation. Previous work has identified three small dimensionless parameters in this problem: the closeness of the length (divided by the cube root of bending stiffness) to the critical one, the amplitude of excitation, and the reciprocal of the frequency of excitation. Multiple timescale analysis is used to show how the asymptotics of resonance tongues in the amplitude-versus-bending-stiffness plane becomes of lower order at certain special values of the frequency ratio where two resonances interact. In particular, an $O(1)$ change in the shape of the parameter region of the stabilized supercritical rod occurs through interaction with the pure harmonic resonance of some other mode of vibration of the rod. It is also shown how to include material damping within the analysis. The results help explain why earlier theories failed to qualitatively explain experimental observation, and are also likely to be of relevance in other three-parameter parametric resonance problems for continuous structures.

Key words. parametric resonance, codimension-two, multiple scale asymptotics, rod theory

AMS subject classifications. 70J40, 34E13, 74K10

DOI. 10.1137/S0036139902418274

1. Introduction.

It is well known that if a column exceeds a certain critical length it will, when placed upright, buckle under its own weight. A recent experiment by Mullin, reported in [14] (see also [1, 2]), has demonstrated that a piece of "curtain wire" that is longer than its critical length can be stabilized by subjecting its bottom support point to a vertical vibration of appropriate amplitude and frequency. In two previous papers, [5, 8], henceforth referred to as Part I and Part II, respectively, we made a numerical and asymptotic study of the linear and nonlinear equations that govern the behavior of such a column.

In Part I we noted that there are three key dimensionless parameters in this problem: B, the ratio of the column's bending stiffness to the cube of its length; η (called $1/\delta$ in that paper), the square of the ratio of the frequency of excitation to that of the pendulum of the same length ($\sqrt{g/\ell}$); and ε, the amplitude of excitation. By looking in the (B, ε) plane for fixed η, one can then analyze the stability of the upright position using Floquet analysis and double scale asymptotics, in much the same way as one does for the single-degree-of-freedom Mathieu equation (e.g., [11, 16]). The result for the fundamental instability curve is to show that for "most" η-values it is possible to stabilize, for small ε, a column that is just longer ($B < B_c$) than its critical length for self-weight buckling. In fact, this criterion gives a lower bound on the frequency of excitation that is required in order to stabilize a column

*Received by the editors November 22, 2002; accepted for publication (in revised form) August 18, 2003; published electronically October 8, 2004. This work was supported in part by the UK Engineering and Physical Sciences Research Council and by the National Textile Center through Clemson University, Clemson, SC.

http://www.siam.org/journals/siap/65-1/41827.html

†Department of Engineering Mathematics, University of Bristol, Bristol BS8 1TR, UK (a.r.champneys@bristol.ac.uk).

‡School of Mathematics and Statistics, The University of Sydney, NSW 2006, Australia (barrief@math.usyd.edu.au).

of a given length. However, this lower bound is too conservative when compared with experiments [14], nor does this criterion give the observed upper bound.

Part II concerns an attempt to answer these deficiencies by considering, in addition, harmonic and subharmonic instabilities. In that paper the analysis was also extended to include a fully geometrically nonlinear formulation, but the ensuing weakly nonlinear asymptotic analysis in essence held no surprises. However, it was observed that the quadratic-in-ε coefficient of the B-value bounding the fundamental stability region can undergo singularities as η varies. These occur precisely when a first harmonic resonance has the same B-value as the self-weight buckling $B = B_c$. In fact, in the analyses of both the fundamental and harmonic instabilities, the asymptotic expansions became invalid at these special values of the excitation frequency. This breakdown appears to be an artifact of the particular expansions used there. In this paper we offer an alternative expansion procedure that avoids this difficulty.

The key is to think of this as a genuinely three-parameter problem, and expand both B and η in powers of ε in a neighborhood of these codimension-two *resonance tongue interaction* points. In what follows we shall consider only the linear equations of motion of the column, since the tongue interaction we wish to describe involves the (loss of) linear stability of an upright column. As in Part II, the results can be extended to include geometrically nonlinear terms to give the amplitude and stability of the bifurcating motion. But since our prime objective is to describe the shape of stability regions in a neighborhood of codimension-two points, we shall omit such complications here.

The rest of the paper is outlined as follows. Section 2 contains a brief review of the mathematical formulation and some new insight into the behavior of eigenmodes and resonances as parameters are varied. Section 3 then considers the possibility of two instabilities occurring at once and introduces a general asymptotic expansion procedure for unfolding such a situation. The details of all the asymptotics are relegated to various appendices. It is argued how only a few of these interactions lead to a qualitatively significant change in the stability regions, including the case where there is a singularity in the coefficient of the subharmonic resonance boundary (with dimensionless frequency 1/2), due to its interaction with the tongue corresponding to frequency 3/2. Section 4 then studies the important special case where the fundamental buckling instability interacts with a first harmonic resonance of any one of the higher-order eigenmodes. Section 5 shows how the results are modified in the presence of damping. Section 6 compares the analysis of this paper to the results of the curtain wire experiments, and, finally, section 7 draws conclusions and discusses the results in a wider context.

2. Mathematical formulation. Consider an initially straight column of length ℓ, with a uniform circular cross-section of radius $a \ll \ell$ and mass linear density m per unit length (see Figure 2.1). The column is assumed to be inextensible, unshearable, and linearly elastic with bending stiffness $\bar{B}$. It is further assumed that the lower end is clamped upright and is displaced by a vertical harmonic oscillation equal to $\Delta \cos \omega_0 \bar{t}$. The upper end is assumed to be free. The derivation of the geometrically nonlinear equations that govern the motion of the centerline $\bar{\mathbf{R}}(\bar{s}, \bar{t})$ and tension $\bar{T}(\bar{s})$ of such a column was detailed in Parts I and II. There it was shown that the effects of rotary inertia and torsional waves in the angular momentum equation can be neglected. Also, in [9] it is shown, via homogenization of a system of links with stiff, damped joints, how to include a dimensionless coefficient of material damping Γ into the linearized equations. This leads to material damping that is proportional to the time derivative

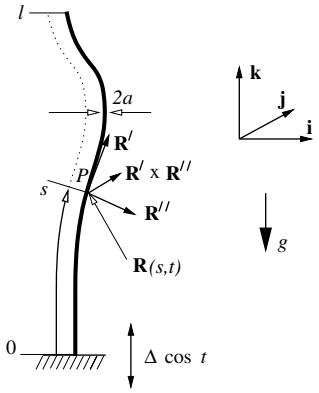


FIG. 2.1. *Definition sketch in which* $\mathbf{R}(s,t) = \ell\mathbf{r}(s,t) + \ell s\mathbf{k} + \Delta\cos(t)\mathbf{k}$, *where* ℓs *is the arclength parameter for the inextensible centerline of the circular column.*

of the fourth-derivative bending stiffness term. Note that this form of damping also occurs in models used in structural geology [13].

2.1. Linearized equations. The dimensionless form of the linearized equations and boundary conditions governing small amplitude motion $\mathbf{r}(s,t)$ of the column about the oscillating upright position are (cf. Part II, (2.5), (2.6), and (2.7))

$$\eta\{\gamma \mathrm{D}\mathbf{r}^{IV} + \mathrm{D}^2\mathbf{r} - \varepsilon\cos t[(1-s)\mathbf{r}']'\} = -\mathcal{M}\mathbf{r} + T'\mathbf{k}, \tag{2.1}$$

where $()' = \partial/\partial s$, $D = \partial/\partial t$, and

$$\mathcal{M}\mathbf{r} := B\mathbf{r}^{IV} + [(1-s)\mathbf{r}']', \quad \text{subject to} \tag{2.2}$$

$$\mathbf{r} = \mathbf{r}' = \mathbf{0} \quad \text{at } s = 0, \qquad \mathbf{r}'' = \mathbf{r}''' = \mathbf{0}, \quad T = 0 \quad \text{at } s = 1. \tag{2.3}$$

Furthermore, upon linearization, inextensibility implies

$$\mathbf{r}' \cdot \mathbf{k} = 0. \tag{2.4}$$

The dimensionless variables are defined in terms of the dimensional variables by the following relations:

$$\left.\begin{aligned} s = \frac{\bar{s}}{\ell}, \quad \mathbf{r} = \frac{\bar{\mathbf{R}}}{\ell} - (\varepsilon\cos t + s)\mathbf{k}, \quad t = \omega_0\bar{t}, \quad \gamma = \frac{\Gamma\ell\omega}{mg} \\ T = \frac{\bar{T}}{mg\ell} - (1 - \eta\varepsilon\cos t)(1-s), \quad B = \frac{\bar{B}}{mg\ell^3}, \quad \eta = \frac{\omega_0^2\ell}{g}, \quad \varepsilon = \frac{\Delta}{\ell}. \end{aligned}\right\} \tag{2.5}$$

In fact, we can eliminate T immediately by considering the $\mathbf{k}$ component of (2.1) subject to condition (2.4), which gives the result $T \equiv 0$. Moreover, without loss of generality at this linearized level, we can assume that the motion is restricted to the $(\mathbf{i},\mathbf{k})$-plane and write $\mathbf{R}(s,t) = u(s,t)\mathbf{i}$, where u satisfies the scalar inhomogeneous, parametrically forced linear PDE

$$\mathrm{M}u + \eta\{\gamma \mathrm{D}u^{IV} + \mathrm{D}^2u - \varepsilon\cos t[(1-s)u']'\} = 0, \tag{2.6}$$

where

$$\mathrm{M}u := Bu^{IV} + [(1-s)u']' \quad \text{subject to} \tag{2.7}$$

$$u = u' = 0 \quad \text{at } s = 0, \qquad u'' = u''' = 0 \quad \text{at } s = 1. \tag{2.8}$$

This completes the formulation of the linear stability problem for the vertically oscillated column. For the majority of this paper we shall consider the perfect rod without the presence of material damping.

2.2. Eigenmodes of the unforced, undamped problem. Consider $\varepsilon = \gamma = 0$. This is the unforced problem. The general linear solution for a given B-value is a superposition of eigenmodes

$$\sum_n \{A_n \cos \sqrt{\lambda_n/\eta_n}\, t + B_n g_n \sin \sqrt{\lambda_n/\eta_n}\, t\}\phi_n(s;B),$$

where ϕ_n is the eigenmode of M corresponding to λ_n. That is, (ϕ_n, λ_n) satisfy

$$\mathrm{M}\phi_n - \lambda_n \phi_n = 0 \tag{2.9}$$

together with boundary conditions (2.8). Note that for each B-value the operator M is self-adjoint and so the eigenfunctions form a complete orthonormal basis for L^2 subject to the boundary conditions (2.8), where we choose to normalize each eigenfunction such that its L^2-norm is unity. Then, upon defining

$$\langle a, b\rangle = \int_0^1 a(s)b(s)ds,$$

we have

$$\langle \phi_i, \phi_j\rangle = \delta_{i,j}, \qquad \langle \phi_i, \mathrm{M}\phi_j\rangle = \lambda_j \delta_{i,j} \tag{2.10}$$

and

$$\langle \phi_i^{IV}, v\rangle = \langle \phi_i'', v''\rangle, \qquad \langle \phi_i, \mathrm{L}v\rangle = \langle v, \mathrm{L}\phi_i\rangle = -B\langle \phi_i'', v''\rangle + \lambda_i \langle \phi_i, v\rangle \tag{2.11}$$

for any function $v(s)$ satisfying the boundary conditions (2.8). Here

$$\mathrm{L}v(s) := [(1-s)v'' - v']. \tag{2.12}$$

These identities will prove useful in what follows.

The eigenmodes $\phi_n(s;B)$ are in general not expressible in closed form except at the special values of B at which $\lambda_n(B) = 0$, in which case (see Part II, section 3) there is a solution in terms of the Bessel function $J_{-1/3}$. The same analysis shows that there are infinitely many such B-values, $B_{0,n}$, $n = 1, 2, 3, 4, \ldots$, accumulating at $B = 0$, the first few values of which are $B_{0,1} = 0.127594$, $B_{0,2} = 0.017864$, $B_{0,3} = 0.0067336$, and $B_{0,4} = 0.0003503$. These correspond, respectively, to where each eigenvalue locus $\lambda_n(B)$, for $n = 1, 2, 3, 4$, crosses the B-axis, with the corresponding eigenmode there having $n-1$ internal zeros. The lowering of B through each $B_{0,n}$-value implies that the nth mode becomes linearly unstable. Hence for $B > B_c := B_{0,1}$ the unforced rod is stable to self-weight buckling (a result known to Greenhill [10]). Numerically, in Part I it was found that each eigenmode $\phi_n(s)$ retains a qualitatively similar mode shape for $B > B_{0,n}$ and that the corresponding loci $\lambda_n(B)$ are approximately straight lines (see Part I, Figure 2 and the schematic diagram in the upper part of Figure 2.2 below). In fact, we can now identify the slopes of those lines via the following asymptotic analysis.

Consider a rod for a specific length $B = B_0$ with eigenvalue and eigenmode (λ_n, ϕ_n) for some n. Let us expand in a small parameter δ via

$$B = B_0 + \delta B_1, \qquad \phi = \phi_n + \delta f_1 + \delta^2 f_2 + \cdots,$$
$$\lambda = \lambda_n + \delta\sigma_1 + \delta^2\sigma_2 + \cdots.$$

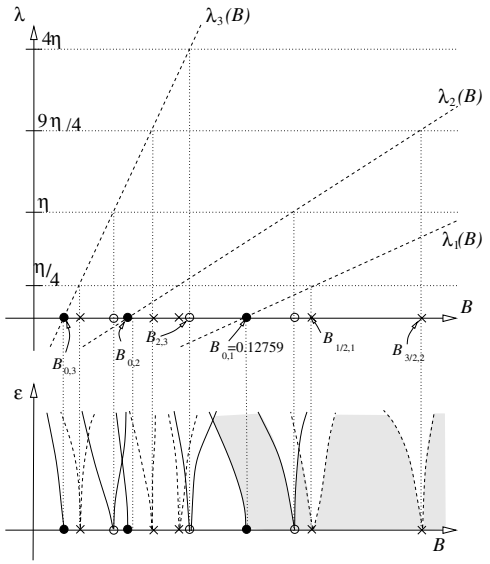


FIG. 2.2. *Summarizing schematically the results from Part* I *and the formula* (2.17). *The upper part shows eigenvalue loci* $\lambda_n(B)$ *and the definition of the points* $B_{\alpha,n}$, $\alpha = 0, 1/2, 1, 3/2, \ldots$. *In the lower part instability tongues in the* (B, ε)*-plane are shown to occur with root points* $B_{\alpha,n}$ *and to have width* $\varepsilon^{2\alpha}$. *The shaded regions correspond to where the vertical solution is stable. Solid lines represent neutral stability curves with Floquet multiplier* $+1$ *(where* α *is an integer) and dashed lines correspond to multiplier* -1 *(where* α *is half an odd integer).*

Taking these expansions and substituting them into the eigenvalue equation (2.9), and collecting powers of δ, we obtain

$$(\mathrm{M}_0 - \lambda_n) f_1 = -B_1 \phi_n^{IV} + \sigma_1 \phi_n, \tag{2.13}$$

$$(\mathrm{M}_0 - \lambda_n) f_2 = -B_1 f_1^{IV} + \sigma_1 f_1 + \sigma_2 \phi_n, \tag{2.14}$$

where M_0 is the operator M evaluated at $B = B_0$. The solvability condition for the $O(\delta)$ equation is that the right-hand side of (2.13) should be orthogonal to the eigenfunction ϕ_n of the left-hand side. Using the first identity in (2.11), this yields

$$\sigma_1 = B_1 \langle \phi_n'', \phi_n'', \rangle \quad \text{and} \quad \langle f_1, \phi_n \rangle = 0.$$

At the next order, the same solvability condition applied to (2.14) yields

$$\sigma_2 = B_1 \langle \phi_n'', f_1'' \rangle,$$

where f_1 is the solution of

$$\mathrm{M}_0 f_1 - \lambda_n f_1 = \sigma_1 \left(\phi_n - \frac{1}{\langle \phi_n'', \phi_n'' \rangle} \phi_n^{IV} \right) \tag{2.15}$$

subject to the boundary conditions (2.8). Combining these results we obtain the following asymptotic expression for dependence of any eigenvalue on B:

$$\lambda = \lambda_n + \delta B_1 \langle \phi_n'', \phi_n'' \rangle + \delta^2 B_1 \langle \phi_n'', f_1'' \rangle + O(\delta^3), \tag{2.16}$$

where we have used explicitly $\langle \phi_n, \phi_n \rangle = 1$. Note that the linear term is a positive definite quantity for all B. Hence the slopes of the loci $\lambda(B)$ are strictly positive for

all n and B. This proves a property that was observed only numerically in Part I. In fact, we found there numerically that for $\lambda_n(B) > 0$ each locus is in fact well approximated by a straight line.

2.3. Dynamic resonances. Consider now the parametrically excited problem $\varepsilon \neq 0$. It is not difficult to see that in the limit $\varepsilon \to 0$, the dimensionless drive angular frequency η will be in resonance with a natural vibration frequency λ_n whenever $\lambda_n(B) = p^2\eta$ for some $n > 0$, $p \geq 0$. Such B-values we label as $B = B_{p,n}$. Similarly, subharmonic resonances occur in the limit $\varepsilon \to 0$ at points $B = B_{p/2,n}$ defined for odd integers p such that $\lambda_n(B) = p^2\eta/2$.

Using Floquet theory (see Part I) one can deduce that for $\varepsilon > 0$, each of these resonance points $B_{\alpha,n}$ for nonnegative half-integers α is the root point of an instability tongue in the (B, ε)-plane (see Figure 2.2). One branch of the tongue corresponds to neutral modes whose leading-order term is $\phi_n(s)\cos\alpha t$ (in phase with the excitation) and the other to modes with leading-order term $\phi_n(s)\sin\alpha t$ (out of phase with the excitation). The case $\alpha = 0$ is special. This corresponds to the instability resulting from one of the pure buckling modes. From such a root point $B = B_{0,n}$ there is thus a single curve in the (B, ε)-plane corresponding to an instability whose mode shape is time independent to leading order. Figure 2.3 shows actual tongue boundaries in the (B, ε)-plane computed using the numerical Floquet theory method presented in Part I.

In fact, following the general asymptotics laid out in section 3, we can add to this result. Specifically, any instability with root point $B_{\alpha,n}$ which is a resonance whose neutral modes are like $\cos\alpha t$ or $\sin\alpha t$, leads to a tongue the boundaries of which are (under certain nondegeneracy assumptions) given by

$$B = B_{\alpha,m} + \sum_{j=2}^{2\alpha-1} B_j\varepsilon^j + B_{2\alpha}^{\pm}\varepsilon^{2\alpha} \tag{2.17}$$

for some coefficients B_j and $B_{2\alpha}^{\pm}$. Here the superscript "+" represents a neutral stability curve corresponding to motion whose time variation is $\cos\alpha t$, i.e., in phase with the drive, and the superscript "−" corresponds to the neutral stability of out-of-phase motion whose time variation is $\sin\alpha t$. See Appendix D for the details.

Thus the simplest family of subharmonic resonances, corresponding to $\alpha = 1/2$, leads to linear instability tongues, whereas the simplest harmonic instability, corresponding to $\alpha = 1$, gives quadratic tongues. Higher-order tongues have width $\varepsilon^{2\alpha}$, albeit with a generically nonzero quadratic lean in the (B, ε)-plane. Thus $\alpha = 3/2$ leads to tongues with width $O(\varepsilon^3)$ in the (B, ε)-plane; $\alpha = 2$ leads to tongues with widths varying to the fourth power of ε, etc. Therefore, for small amplitude of excitation ε, only instabilities corresponding to lower values of α are likely to lead to any significant regions of instability provided the amplitude of excitation ε is small (an observation that is further vindicated by the presence of damping; see section 5). The case $\alpha = 0$ is once again special, and it was shown in Part I that this leads to a single boundary between instability and stability that is to leading order quadratic in B.

The general result (2.17) confirms and extends what we found by detailed multiple timescale asymptotics in Parts I and II. Before proceeding to a codimension-two analysis of when the nondegeneracy conditions leading to (2.17) fail, let us extract from Part II the leading-order expressions for the coefficients in the simplest few cases.

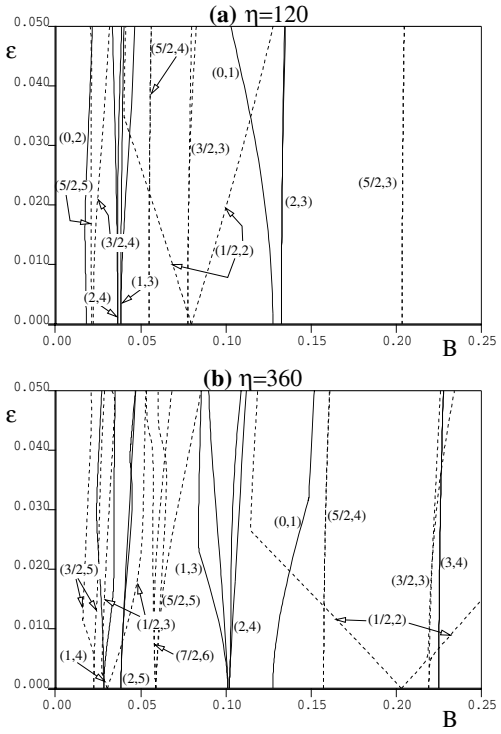


FIG. 2.3. *Resonance tongue boundaries computed using numerical Floquet analysis with* $N = 4$ *for* (a) $\eta = 120$ *and* (b) $\eta = 360$. *Depicted are all resonance tongues (labelled by* (α, n)*) with root points corresponding to* $B_{\alpha,n}$ *with* $\alpha \leq 5/2$ *and* $n \leq 4$ *and* $0.02 < B < 0.25$. *Note that the order of the resonance tongues changes between the figures (but the instability curve* $(0, 2)$ *is omitted from panel* (b) *for clarity). This is because at intermediate* η*-values there have been codimension-two resonance tongue interactions which form the subject of this paper. Note that each tongue other than the instability boundary coming out of* $B_{0,n}$ *has nonzero width which is sometimes not apparent on the scale depicted.*

Taking $\alpha = 0$, the single neutral curve is defined by

$$B = B_{0,n} + B_2\varepsilon^2 + O(\varepsilon^3)$$

where (Part II, equation (4.11), interpreted in the notation of this paper)

$$B_2 = \frac{\eta\langle\phi_n, \mathrm{L}H_1\rangle}{2\langle\phi_n'', \phi_n''\rangle}, \tag{2.18}$$

with H_1 being the solution to

$$\mathrm{M}_{0,n}H_1 - \eta H_1 = \eta\mathrm{L}\phi_n \tag{2.19}$$

subject to the usual boundary conditions (2.8).

Taking $\alpha = 1/2$, the neutral stability curves are defined by

$$B = B_{1/2,n} + B_1^{\pm}\varepsilon + O(\varepsilon^2),$$

where (Part II, equation (4.19))

$$B_1^{\pm} = \pm\frac{\eta\langle\phi_n, \mathrm{L}\phi_n\rangle}{2\langle\phi_n'', \phi_n''\rangle}. \tag{2.20}$$

In the case $\alpha = 1$, neutral stability occurs along the curves

$$B = B_{1,n} + B_2^{\pm}\varepsilon^2 + O(\varepsilon^3),$$

where (upon solving for $\alpha_1 = \pm\alpha_2$ in Part II, equation (4.26))

$$B_2^+ = \eta\frac{2\langle\phi_n, \mathrm{L}H_4\rangle + \langle\phi_n, \mathrm{L}H_3\rangle}{2\langle\phi_n'', \phi_n''\rangle}, \tag{2.21}$$

$$B_2^- = \frac{\eta\langle\phi_n, \mathrm{L}H_3\rangle}{2\langle\phi_n'', \phi_n''\rangle}, \tag{2.22}$$

with H_3 and H_4 being the solutions to

$$\mathrm{M}_{1,n}H_3 - 4\eta H_3 = \frac{\eta}{2}\mathrm{L}\phi_n, \tag{2.23}$$

$$\mathrm{M}_{1,n}H_4 = \frac{\eta}{2}\mathrm{L}\phi_n, \tag{2.24}$$

subject to the usual boundary conditions (2.8).

We now notice several anomalies from Figure 2.3. First the instability curve $(0, 1)$ representing the fundamental falling-over instability, bends back to the left for $\eta = 120$ (as indeed it does for "most" η-values, as was argued in Part I to explain the stabilization effect observed in the experiment). However, for $\eta = 360$, this curve bends to the right, thus showing that increasing ε does not lead to a region of stability of the forced rod that failed to exist for the unforced problem. This anomaly forms the subject of section 4 below. Second, the instability tongue $(1/2, 2)$ undergoes, for $\eta \approx 360$, a strange interaction process with the $(3/2, 3)$-tongue in the bottom right of Figure 2.3(b). An explanation of this and similar interactions forms the subject of section 3 below.

3. Resonance tongue interaction. The general expression (2.17) giving the leading-order expression for resonance tongues is subject to nondegeneracy conditions. These nondegeneracy conditions fail at special values of η for which

$$B_{\alpha,n} = B_{\beta,m} \quad \text{for } n \neq m, \tag{3.1}$$

and values of α and β that are related by certain conditions as we shall now explain.

It was noted in Part II that the coefficient B_2 given by (2.18) becomes singular precisely when $\lambda_m = \eta$ is another eigenvalue of M, corresponding to eigenmode ϕ_m. That is, there is the coexistence of resonances (3.1) corresponding to $\alpha = 1$ and $\beta = 0$. We can see why this is so since (2.19) becomes such that the left-hand side is solved by eigenfunction ϕ_m and hence the solution becomes unbounded unless the right-hand side function is orthogonal to ϕ_m. This will occur at a special pair of values (η_0, B_0) of η and B for which $M_0\phi_n = M_0\phi_m - \eta_0\phi_m = 0$. At such points we can see that the coefficient of the quadratic coefficient B_2^+ of the resonance tongue corresponding to $\phi_m \cos t$ also becomes unbounded, since H_4 given by (2.24) becomes unbounded. Similarly both coefficients B_2^+ and B_2^- of the $\phi_m \cos t$ and $\phi_m \sin t$ boundaries become singular when $B_{(1,n)} = B_{(2,m)}$, because the function H_3 given by (2.23) becomes unbounded. This leads to two questions: Precisely which pairs of values of frequencies α and β can lead to such singularities? and, How can one unfold these codimension-two resonance tongue interactions, allowing both η and B to vary?

3.1. Multiple timescale asymptotic expansion. In order to answer these questions, we shall develop a general multiple timescale asymptotic expansion about a given pair of values (B_0, η_0). Suppose that $B_0 = B_{\alpha,n}$ is the root point in the (B, ε)-plane of a resonance tongue corresponding to motion with angular frequency α. We shall consider the possibility that at precisely (B_0, η_0) there is the root point of a second resonance tongue so that $B_0 = B_{\alpha,n} = B_{\beta,m}$ for $m \neq n$. For the time being we do not assume any relation between the half-integers α and β, except of course the relation between the eigenvalues

$$\eta_0 = \frac{\lambda_n}{\alpha^2} = \frac{\lambda_m}{\beta^2}. \tag{3.2}$$

Since we do not know a priori which time scales will lead to a distinguished limit for this problem we will now consider all functions to be functions of a hierarchy of time scales (t, $\tau_1 = \delta t$, $\tau_2 = \delta^2 t$, ...) and introduce the following expansions:

$$\left.\begin{aligned} u(s) &= \delta u_1(s,t,\tau_1,\tau_2) + \delta^2 u_2(s,t,\tau_1,\tau_2) + \delta^3 u_3(s,t,\tau_1,\tau_2) + \cdots, \\ \eta &= \frac{\lambda_n}{\alpha^2} + \delta\eta_1 + \delta^2\eta_2 + \cdots, \\ B &= B_0 + \delta B_1 + \delta^2 B_2 + \cdots, \\ \varepsilon &= \delta\varepsilon_1, \quad \gamma = \delta\gamma_1. \end{aligned}\right\} \tag{3.3}$$

Here, for future use in section 5, we have retained the material damping γ at the same order $O(\delta)$ as the amplitude of excitation ε. By keeping the small parameter δ separate from ε we allow for the possibility of having nonzero damping at zero excitation; that is, $\varepsilon_1 = 0$, $\gamma_1 \neq 0$.

When these series are substituted into (2.6)–(2.8) and the coefficients of the terms up to δ^3 are set to zero, the result is

$$\frac{\lambda_n}{\alpha^2}\frac{\partial^2 u_1}{\partial t^2} + \mathrm{M}_0 u_1 = 0, \tag{3.4}$$

$$\frac{\lambda_n}{\alpha^2}\frac{\partial^2 u_2}{\partial t^2} + \mathrm{M}_0 u_2 = \varepsilon_1 \frac{\lambda_n}{\alpha^2}\cos t \mathrm{L} u_1 - \eta_1 \frac{\partial^2 u_1}{\partial t^2} - \frac{2\lambda_n}{\alpha^2}\frac{\partial^2 u_1}{\partial t \partial \tau_1} - B_1 u_1^{IV} - \gamma_1 \frac{\lambda_n}{\alpha^2}\frac{\partial u_1^{IV}}{\partial t}, \tag{3.5}$$

$$\begin{aligned} \frac{\lambda_n}{\alpha^2}\frac{\partial^2 u_3}{\partial t^2} + \mathrm{M}_0 u_3 = {} & \varepsilon_1 \frac{\lambda_n}{\alpha^2}\cos t \mathrm{L} u_2 - \eta_1 \frac{\partial^2 u_2}{\partial t^2} - \frac{2\lambda_n}{\alpha^2}\frac{\partial^2 u_2}{\partial t \partial \tau_1} - B_1 u_2^{IV} \\ & - \gamma_1 \frac{\lambda_n}{\alpha^2}\frac{\partial u_2^{IV}}{\partial t} + \varepsilon_1 \eta_1 \cos t \mathrm{L} u_1 - \eta_2 \frac{\partial^2 u_1}{\partial t^2} - 2\eta_1 \frac{\partial^2 u_1}{\partial t \partial \tau_1} \\ & - \frac{2\lambda_n}{\alpha^2}\frac{\partial^2 u_1}{\partial t \partial \tau_2} - \frac{\lambda_n}{\alpha^2}\frac{\partial^2 u_1}{\partial \tau_1^2} - B_2 u_1^{IV} \\ & - \gamma_1 \frac{\lambda_n}{\alpha^2}\frac{\partial u_1^{IV}}{\partial \tau_1} - \gamma_1 \eta_1 \frac{\partial u_1^{IV}}{\partial t}. \end{aligned} \tag{3.6}$$

Here the linear operator M_0 is M (defined in (2.7)) evaluated at $B = B_0$:

$$\mathrm{M}_0 u = B_0 u^{IV} + \mathrm{L} u \quad \text{with } \mathrm{L} u = [(1-s)u']',$$

and $u_1(s), u_2(s), u_3(s)$ are each subject to the boundary conditions (2.8).

Following assumptions (3.2) on coexistent resonances, we shall take the $O(\delta)$ solution to be

$$u_1 = [f(\tau_1, \tau_2, \dots)\cos\alpha t + g(\tau_1, \tau_2, \dots)\sin\alpha t]\phi_n(s) + [d(\tau_1, \tau_2, \dots)\cos\beta t + e(\tau_1, \tau_2, \dots)\sin\beta t]\phi_m(s). \tag{3.7}$$

The functions f, g, e, d will be determined at higher-order by nonresonance conditions.

Appendix A contains the general solution to the problem at $O(\delta^2)$ together with solvability conditions that can be expressed as fourth-order boundary-value problems involving the operator M_0.

In this section and the next we are interested in the zero-damping case, and so we now set

$$\gamma_1 = 0, \quad \varepsilon_1 = 1; \quad \text{hence } \delta \equiv \varepsilon,$$

and we shall use ε rather than δ as our expansion parameter.

Let us assume for now that we are not in the special case where α or $\beta = 1/2$. Appendix B then shows that the solvability condition at $O(\varepsilon^2)$ gives that the zero solution is stable except at the two isolated values

$$B_1 = \frac{\eta_1\alpha^2}{\langle\phi_n'', \phi_n''\rangle} \quad \text{or} \quad \frac{\eta_1\beta^2}{\langle\phi_n'', \phi_m''\rangle}. \tag{3.8}$$

However, note that these two conditions (3.8) can be written more simply as

$$\sigma_1 = B_1\langle\phi_p'', \phi_p''\rangle, \quad \text{where } \lambda = \lambda_p + \varepsilon\sigma_1,$$

for $p = n$ or m. We recognize immediately that this is the first-order-in-ε correction to the curve (2.16) for the eigenvalues as a function of B if we demand that either the condition $\lambda = \eta\alpha^2$ or $\lambda = \eta\beta^2$ remains true at nearby (B, η)-values. In other words, this moves the underlying B and η values to new ones that satisfy the appropriate eigenvalue condition (either of the two equalities in (3.2)). Since we are interested in expanding about η_0 and B_0, we must therefore choose

$$\eta_1 = B_1 = 0. \tag{3.9}$$

Equations (A.9) and (A.10) have a unique bounded solution unless $(\beta \pm 1)^2 = \alpha^2$. That is, unless

$$\beta = \alpha \pm 1. \tag{3.10}$$

If (3.10) is satisfied, then the above asymptotic expansion becomes invalid and we note that the function F_3 or F_4 must be combined with H_1 or H_2, respectively. Then the nonresonance conditions (B.2) and (B.3) lead to nontrivial expressions for B_1 as a function of η_1. We shall treat these on a case-by-case basis in sections 3.2, 3.3, and 4.

For the time being let us continue by assuming that (3.10) is *not* satisfied. Then, Appendix C gives the form of the general equation at $O(\varepsilon^3)$, and we now proceed to analyze its various special cases. First note that the asymptotics is now in place to derive the general description 2.17 of each codimension-one resonance tongue (where β is unrelated to α); see Appendix D.

3.2. Resonance interaction at $O(\varepsilon^2)$. Consider what happens at $O(\varepsilon^2)$ when (3.10) is satisfied, which without loss of generality we assume occurs with

$$\beta = \alpha + 1 \tag{3.11}$$

(avoiding for the time being the special cases $\alpha = 0$ or $\alpha = 1/2$). As we already remarked, the asymptotic expansion we have introduced above becomes invalid when (3.11) is satisfied. In particular we can no longer assume $B_1 = \eta_1 = 0$. Instead we find stability at this order except on a single neutral curve in the (B_1, η_1)-plane

$$(\eta_1\alpha^2 - B_1\langle\phi_n'', \phi_n''\rangle)(\eta_1(\alpha+1)^2 - B_1\langle\phi_m'', \phi_m''\rangle) = \frac{\lambda_n^2}{4\alpha^4}\langle\phi_m, \mathrm{L}\phi_n\rangle^2; \tag{3.12}$$

see appendix E.

Note that the right-hand side of (3.12) is strictly positive, whereas the left-hand side is the product of two linear functions of η_1 and B_1. This is the equation for a hyperbola in the (η_1, B_1)-plane; see Figure 3.1(a). In the limit that $|\eta_1|$ is large, the locus of solutions becomes two straight lines with slopes $\frac{1}{\alpha^2}\langle\phi_n'', \phi_n''\rangle$ and $\frac{1}{(\alpha+1)^2}\langle\phi_m'', \phi_m''\rangle$, which according to (2.16) are precisely the linear approximations to the loci $B_{\alpha,n}(\eta)$ and $B_{\alpha+1,m}(\eta)$. Hence a long way from the resonance tongue interaction there is no correction at $O(\varepsilon)$ to the root points of the resonance tongues (as expected). Note, however, as shown in Figure 3.1, the solution at finite values of the excitation ε that was attached to the (α, n) tongue for $\eta_1 \ll 0$ switches over to become associated with the $(\alpha+1, m)$ tongue for $\eta_1 \gg 0$.

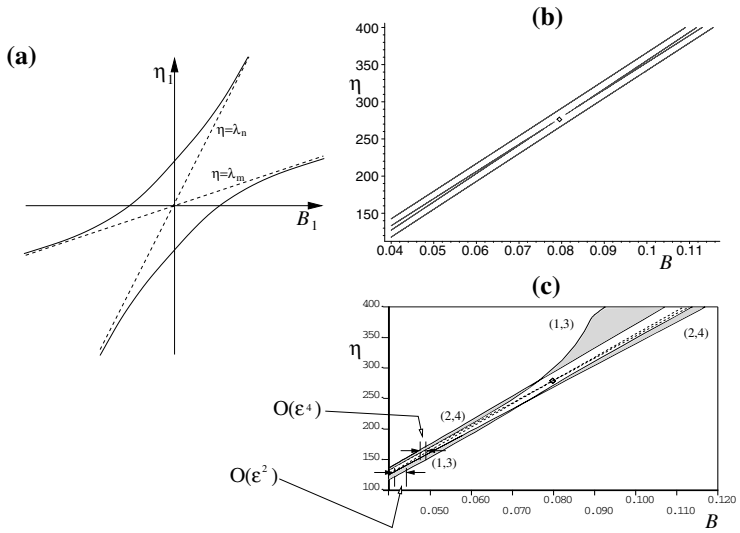


FIG. 3.1. (a) *Schematic diagram of the interaction at $O(\varepsilon)$ of the resonance tongues corresponding to α and $\beta = \alpha + 1$ according to* (3.12). *The dashed lines represent the eigenvalue loci $B_{\alpha,n}(\eta)$ and $B_{\beta,m}(\eta)$, and the solid line gives the $O(\varepsilon)$ correction. Note that at this order both tongues have zero width, provided $\alpha \geq 1$.* (b), (c) *Numerical illustration in the case $(\alpha, n) = (1, 3)$, $(\beta, m) = (2, 4)$, and $\varepsilon = 0.02$. Panel* (b) *shows the evaluation of the formula* (3.12) *(outer two curves) together with the eigenvalue loci $B = B_{1,3}(\eta)$ and $B = B_{2,4}(\eta)$ which cross at $\eta = 278.808$ (marked by a diamond). Panel* (c) *shows the numerically computed stability boundaries using Floquet theory with $N = 4$. The eigenvalue loci are dashed. Shaded regions correspond to instability inside the resonance tongue.*

Figure 3.1 numerically illustrates a particular resonance tongue interaction between the tongues $(1,3)$ and $(2,4)$. Note from Figure 2.3 that these two resonance tongues switch their order (the relative B-values for which they occur) between $\eta = 120$ and $\eta = 360$. In fact we find numerically that these two resonances interact at

$$\eta_0 = 278.809, \qquad B_0 = 0.0798242,$$

from which we have calculated that

$$\langle \phi_3'', \phi_3'' \rangle = 3807.75, \qquad \langle \phi_4'', \phi_4'' \rangle = 14618.8, \qquad \langle \phi_4'', \mathrm{L}\phi_3'' \rangle = 8.55440.$$

Using these quantities, the loci of the $O(\varepsilon)$ behavior of the tongues, (3.12) is calculated and plotted in Figure 3.1(b). Note that the slopes of the loci $\eta = B_{1,3}$ and $\eta = B_{2,4}$ are much closer than in the schematic panel (a). Finally, panel (c) of the figure shows how these results compare with a full numerical evaluation of the resonance tongues using Floquet theory for fixed amplitude $\varepsilon = 0.02$. Note that the tongues, which are $O(\varepsilon^2)$ and $O(\varepsilon^4)$ in theory away from the interaction (marked at the left-hand edge of the figure) undergo an abrupt change as they pass through a neighborhood of the resonance tongue interaction point. There are several features to this change. First, as predicted and in broad quantitative agreement with the results in panel (b), the resonance tongues do not cross, but each tongue becomes attached to the opposite instability. Second, there is a point close to the codimension-two point at which each tongue (at this value of ε) "pinches off." The result is that the cosine and sine boundaries switch sides. This pinching off of the resonance tongue is not part of the above $O(\varepsilon)$ theory, but can be seen here as a necessary consequence of the resonance tongue interaction process. Note that such pinching of resonance tongues has been described before for one-degree-of-freedom parametrically excited systems [4]. (It is also in evidence in Figure 2.3, where for $\eta = 360$, both the (1,3) and (2,4) resonance tongues become narrower for ε-values toward the top of the graph, suggesting that they pinch off for higher ε still (indeed they do); also the tongue (2,5) can be seen to undergo just such a pinching at $\varepsilon \approx 0.035$.) Finally, note that beyond the codimension-two point the $(1,3)$ tongue suddenly becomes much fatter. This is because of its interaction with the buckling mode instability at $B = B_{0,1} = 0.127594$ which was discussed in Part II, section 4(e), and will be treated further in section 4 below.

3.3. The case $\alpha = 1/2$, $\beta = 3/2$; singularity in the subharmonic resonance. This case is special because the $O(\varepsilon)$ correction due to the resonance tongue interaction is at the same order as the $O(\varepsilon)$ width of the resonance tongue itself. From Appendix E we obtain the stability boundary (B_1, η_1)-plane

$$\left(\frac{\eta_1}{4} - B_1\langle \phi_n'', \phi_n'' \rangle \pm 2\lambda_n \langle \phi_n, \mathrm{L}\phi_n \rangle\right)\left(\frac{9\eta_1}{4} - B_1\langle \phi_m'', \phi_m'' \rangle\right) = \frac{4}{9}\lambda_n^2 \langle \phi_m, \mathrm{L}\phi_n \rangle^2. \tag{3.13}$$

This describes two hyperbolae which bound the shaded region of stability shown schematically in Figure 3.2(a). For large η_1 and B_1 they asymptote to the straight lines

$$\begin{aligned} B_1 &= B_{n,1/2} \pm B_1^{\pm} = \frac{\eta_1}{4\langle \phi_n'', \phi_n'' \rangle} \pm \frac{\langle \phi_n, \mathrm{L}\phi_n \rangle}{\langle \phi_n'', \phi_n'' \rangle} \quad \text{and} \\ B_1 &= B_{m,3/2} = \frac{9\eta_1}{4\langle \phi_n'', \phi_n'' \rangle}, \end{aligned} \tag{3.14}$$

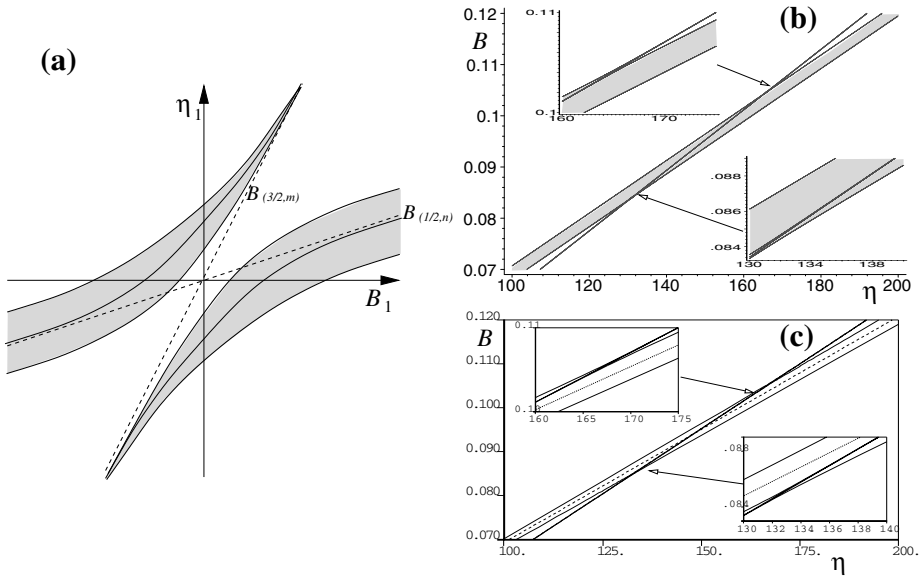


FIG. 3.2. *The interaction at $O(\varepsilon)$ of the resonance tongues corresponding to $\alpha = 1/2$ and $\beta = 3/2$.* (a) *Schematic picture; the shaded region corresponds to that of instability.* (b) *Evaluation of the formula* (3.13) *for $n = 2$, $m = 3$, using numerically calculated eigenfunctions ϕ_n and ϕ_m and $\varepsilon = 0.001$.* (c) *Numerically computed stability boundaries using Floquet theory with $N = 4$. The dashed line, as in part* (a)*, corresponds to the locus $B = B_{1/2,2}(\eta)$. The locus $B = B_{3/2,3}$ is indistinguishable in this plot from the corresponding tongue boundary with finite ε.*

which are the $O(\varepsilon)$ expressions for the two stability boundaries in the absence of their interaction. The region of instabilities for a finite ε are shaded in Figure 3.2(a), and are obtained by noting the inequalities that must be true in order for (E.11)–(E.14) to have bounded solutions. The resonance conditions $B_1 = B_{n,1/2}$ and $B_{m,3/2}$, valid when $\varepsilon = 0$, are depicted as dashed lines in the figure. Observe that the interaction between the two instabilities causes a small region of *stability* for finite ε in a neighborhood of the critical point $\eta_1 = B_1 = 0$, where the two resonance curves $B_{n,1/2}$ and $B_{m,3/2}$ cross.

By numerical computation we have found numerous examples of resonance tongue interaction of this kind. For example, taking $n = 2$ and $m = 3$ we find that $\lambda_n = \eta/4$ and $\lambda_m = 9\eta/4$ at $(B, \eta) = (0.0941064, 148.084)$. Note that this B-value is a little less than $B_c = 0.127594$, which implies that this interaction of instabilities is occurring for rods which are already marginally unstable to self-weight buckling. At those parameter values, computation of the eigenfunctions reveals

$$\begin{aligned} \langle \phi_n'', \phi_n'' \rangle &= 485.765, \qquad \langle \phi_m'', \phi_m'' \rangle = 3807.41, \\ \langle \phi_n, \mathrm{L}\phi_n \rangle &= -8.69181, \qquad \langle \phi_m, \mathrm{L}\phi_n \rangle = -1.946423. \end{aligned} \tag{3.15}$$

Using these precise values, Figure 3.2(b) shows the evaluation of the loci (3.13) with $\varepsilon = 0.001$, which are compared in Figure 3.2(c) to the computation of the same stability boundaries using the numerical Floquet theory introduced in Part I. Note from the values (3.15) that the straight lines $B_{1/2,2}$ and $B_{3/2,3}$ given by (3.14) with $n = 2$ and $m = 3$ have slopes that differ by less than 10%. Hence the stability region caused by the interaction between these two resonances is very small compared to the width of the instability tongues around $B = B_{1/2,2}$ (see insets).

3.4. Resonance interaction at higher order. The above analysis can be continued to $O(\varepsilon^3)$ to obtain the $O(\varepsilon^2)$ corrections to the four solid lines in Figure 3.2(a). Carrying out the expansion to $O(\varepsilon^4)$ would show how to couple the shapes of the shaded instability boundaries to the $O(\varepsilon^3)$-thick resonance tongue around the locus $B = B_{m,3/2}$ which itself would be a curve with nonzero coefficients of ε^2 and ε^3. There is little extra qualitative information to be obtained by carrying out these expansions explicitly. Instead, let us focus on other resonance-tongue interactions which can be captured only by going to $O(\varepsilon^3)$ or higher.

Suppose first that (3.10) is not satisfied, but instead

$$\beta = \alpha \pm 2; \quad \text{without loss of generality} \quad \beta = \alpha + 2.$$

The analysis at $O(\varepsilon^2)$ now proceeds as in Appendix A and we have $B_1 = \eta_1 = 0$. Consider the $O(\varepsilon^3)$ equation (C.1). Then the coefficients of $\cos(\beta-2)t$ and $\sin(\beta-2)t$ are also resonant and must be added to the coefficients of $\cos\alpha t$ and $\sin\alpha t$ to form solvability conditions. Also the coefficients of $\cos(\alpha+2)t$ and $\sin(\alpha+2)t$ should be added to those for $\cos\beta t$ and $\sin\beta t$. Again there will be two pairs of relations, from orthogonality to ϕ_n and ϕ_m separately. This will lead to a nontrivial equation linking B_2 and η_2, like (3.12) for η_1 and B_1 in the case of a lower-order resonance tongue interaction.

So we find in general that the coefficient of ε^3 in the asymptotics of the resonance tongues undergoes a singularity which is patched up by this correction to the $O(\varepsilon^2)$ coefficient B_2 as η passes through η_0. Similarly, by extrapolation of the above argument to higher powers of ε in our asymptotic expansion, we find that if

$$\beta = \alpha + n,$$

then there is a singularity in the $O(\varepsilon^{n+1})$ coefficient of the resonance tongues which is resolved by showing that there is a nontrivial contribution to the $O(\varepsilon^n)$ coefficient as η passes through the critical value at which the interaction occurs.

4. The case $\alpha = 1$, $\beta = 0$; first-harmonic buckling interaction. The case $\alpha = 1$, $\beta = 0$ leads to quite different results since instead of a resonance tongue, we have a single neutral stability curve corresponding to the rod "falling over" into one of its buckling modes. The bookkeeping in section 3.2 above shows that this interaction causes a singularity in the $O(\varepsilon^2)$ coefficient B_2 of this instability curve in the (B, ε)-plane. Clearly this can have a profound effect on the stability analysis of the slightly longer than critical column because it was precisely the negativity of this coefficient B_2 that gave the stabilization effect referred to colloquially as the "Indian rope trick" in Parts I and II.

Consider a neighborhood of a special η-value $\eta = \lambda_n$ at which $B_{1,n} = B_{0,m}$. Particular physical interest is in the case $m = 1$, in which case we consider the column that is only slightly longer (or shorter) than the length of column that will just stand under its own weight. That is $B \approx B_c = B_{0,1}$, the critical value of the dimensionless parameter B for self-weight buckling. We shall denote the critical eigenmode corresponding to B_c as ϕ_c and the eigenmode corresponding to the pure dynamic instability ϕ_n. In fact, the analysis below works equally well for $B_c = B_{0,m}$ for any $m \geq 2$, except that "stability" implies then "stable to vibration mode m" rather than absolute stability since such a rod is statically unstable to modes ϕ_p, $p = 1, \ldots, n-1$.

Thus the solution of the $O(\varepsilon)$ equation (3.4) we take is

$$u_1 = \{h(\tau_1,\tau_2)\phi_c(s) + [f(\tau_1,\tau_2)\cos t + g(\tau_1,\tau_2)\sin t]\phi_n(s)\}, \tag{4.1}$$

with $\phi_m(s)$ and $\phi_c(s)$ subject to the boundary conditions (2.8).

4.1. The $O(\varepsilon^2)$ interaction equation. In Appendix F the following condition for stability at $O(\varepsilon^2)$ is derived:

$$\left(\frac{\eta_1}{2\lambda_n} - \frac{B_1\langle\phi_n'',\phi_n''\rangle}{2\lambda_n}\right)\left(\frac{\eta_1}{2\lambda_n} - \frac{B_1\langle\phi_n'',\phi_n''\rangle}{2\lambda_n} + \frac{\lambda_n B_c^2\langle\phi_n'',\phi_c''\rangle^2}{4B_1\langle\phi_c'',\phi_c''\rangle}\right) > 0, \tag{4.2}$$

with the stability boundaries in (η, B, ε)-space determined by the zeros of this expression.

One boundary is determined by

$$\left(\frac{\eta_1}{2\lambda_n} - \frac{B_1\langle\phi_n'',\phi_n''\rangle}{2\lambda_n}\right) := \sigma_1 = 0; \quad \text{hence } \eta_1 = B_1\langle\phi_n'',\phi_n''\rangle. \tag{4.3}$$

In this case f is independent of τ_1 by (F.9) and, unless $f \equiv 0$, g will be a linear function of τ_1 by (F.8) and hence unbounded as $\tau_1 \to \infty$. Thus g is constant as a function of τ_1, which implies the motion is $\sin t$, out of phase with the drive.

The other boundary is determined by

$$\left(\frac{\eta_1}{2\lambda_n} - \frac{B_1\langle\phi_n'',\phi_n''\rangle}{2\lambda_n} + \frac{\lambda_n B_c^2\langle\phi_n'',\phi_c''\rangle^2}{4B_1\langle\phi_c'',\phi_c''\rangle}\right) := \sigma_2 = 0; \tag{4.4}$$

hence

$$\eta_1 = B_1\langle\phi_n'',\phi_n''\rangle - \frac{\lambda_n^2 B_c^2\langle\phi_n'',\phi_c''\rangle^2}{2B_1\langle\phi_c'',\phi_c''\rangle}.$$

In this case g is independent of τ_1 by (F.8) and therefore unless $g \equiv 0$, f will be a linear function of τ_1, and unbounded on this boundary. With this choice f and h are at most functions of τ_2, which implies a motion that is in phase with the drive, but with a nonzero lean (proportional to a constant plus $\cos t$).

To determine on which side of these boundaries solutions are stable we simply observe that for stability we must have $\sigma_1\sigma_2 > 0$ so that σ_1 and σ_2 must be either both negative or both positive. The regions of stability are shown as shaded in Figure 4.1(a).

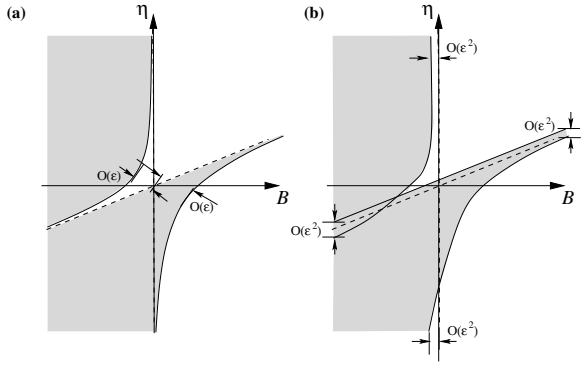


FIG. 4.1. *Schematic figure of the resonance tongue interaction process between the falling-over and first-harmonic instabilities.* (a) *at* $O(\varepsilon)$; (b) *including* $O(\varepsilon^2)$.

4.2. Correction at $O(\varepsilon^3)$. We now construct the correction $B_2\varepsilon^2$ and $\eta_2\varepsilon^2$ to these stability boundaries. This is necessary in order to see how the above $O(\varepsilon)$ correction matches the limit when the two resonances are a long way from interaction. In that case, both the falling-over and harmonic instabilities are quadratic to leading order (formulae (2.18), (2.22) and (2.21)). We shall consider a neighborhood of the boundaries $\sigma_1 = 0$ and $\sigma_2 = 0$ separately.

The case $\sigma_1 = 0$ is dealt with in Appendix G. The result is that one can combine the results for B_1 and B_2 into the single expression

$$B = B_c + \frac{(\eta - \lambda_n)}{\langle \phi_n'', \phi_n'' \rangle} - \varepsilon(\eta - \lambda_n)\frac{\langle \phi_n'', H_2'' \rangle}{\langle \phi_n'', \phi_n'' \rangle^2} + \varepsilon^2 \frac{1}{2}\lambda_n \frac{\langle \phi_n, \mathrm{L}H_3 \rangle}{\langle \phi_n'', \phi_n'' \rangle} + O(\varepsilon^3). \tag{4.5}$$

Note that the first three terms of (4.5) are precisely the expansion of the locus $B = B_{1,m}(\eta)$ up to $O(\varepsilon^2)$ that defines the condition that there is an eigenvalue $\lambda = \eta$. To see this, take $\delta = \varepsilon$, $n = m$, $\lambda = \eta$, $B_1 = B - B_c$ in (2.16) and invert the expansion up to $O(\varepsilon^2)$, noting that (2.14) satisfied by f_1 is a scalar multiple of that satisfied by H_2, which is (F.5) with $\eta_1 \equiv \sigma_1$ and $f = 0$. The fourth term of (4.5) is just the $O(\varepsilon^2)$ coefficient B_2^- given by (2.22) of the $\sin t$ boundary of the resonance tongue when it does not interact with the falling-over mode. We conclude that this boundary of the resonance tongue does not become singular as it passes through the resonance tongue interaction point.

Consider now the boundary $\sigma_2 = 0$, which contains a discontinuity. The results in Appendix H show that the correction B_2 and η_2 can be expressed in terms of B_1 and η_1:

$$B_2 = \frac{1}{2G_n\langle \phi_c'', \phi_c'' \rangle}(2B_1K_n\langle \phi_c'', H_0'' \rangle - K_n\lambda_n\langle \phi_c, \mathrm{L}H_1 \rangle - B_1\eta_1\langle \phi_c, \mathrm{L}\phi_n \rangle), \tag{4.6}$$

$$\begin{aligned}\eta_2 = {} & \frac{1}{B_1}\left[\eta_1 G_n\langle \phi_c'', \phi_n'' \rangle - K_n\lambda_n\left(\langle \mathrm{L}H_0, \phi_n \rangle + \frac{1}{2}\langle \mathrm{L}H_3, \phi_n \rangle\right)\right] \\ & + K_n\langle H_1'', \phi_n'' \rangle + B_2\langle \phi_n'', \phi_n'' \rangle.\end{aligned} \tag{4.7}$$

Now we have to decide how to interpret these results. At $O(\varepsilon)$ we have a neutral stability curve (4.4) in the (B, η)-plane that asymptotes to $\eta = \infty$ as $B \to 0$. However, the above formulae (4.6) and (4.7) provide corrections to both η and B_2 at each point on this curve. The most meaningful way of applying this asymptotic correction is to take only those components that are normal to the curve. To that end we can define a new coordinate

$$\hat{\eta} = \eta - B\langle \phi_n'', \phi_n'' \rangle$$

so that the straight line $\eta_1 = B_1\langle \phi_n'', \phi_n'' \rangle$ to which the curve (4.4) asymptotes as $\eta_1 \to \infty$ becomes the B_1-axis. Also, the second $O(\varepsilon^2)$-correction equation (4.7), which involves both B_2 and η_2, becomes just a condition for $\hat{\eta}_2$. Then taking B_1 as a single independent coordinate and, according to (4.4)

$$\hat{\eta}_1 = -\frac{\lambda_n^2 B_c^2 \langle \phi_n'', \phi_c'' \rangle^2}{2B_1\langle \phi_c'', \phi_c'' \rangle}, \tag{4.8}$$

we can write that the second-order corrections $\tilde{B}_2, \tilde{\hat{\eta}}_2$ should satisfy

$$(\tilde{B}_2, \tilde{\hat{\eta}}_2) = \frac{(B_2, \hat{\eta}_2) \cdot (-\hat{\eta}_1, B_1)}{B_1^2 + \hat{\eta}_1^2}(-\hat{\eta}_1, B_1). \tag{4.9}$$

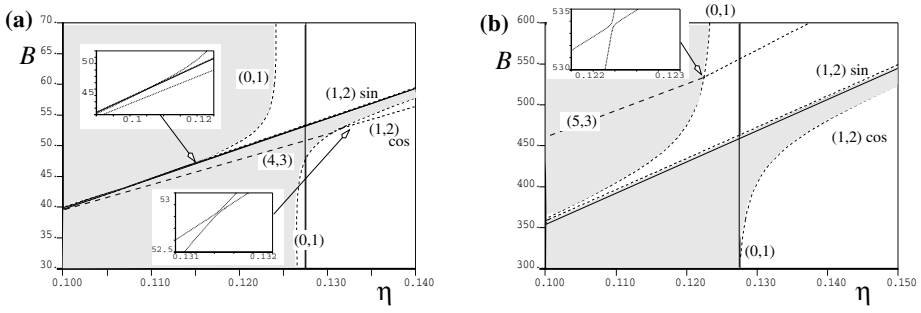


FIG. 4.2. *Numerical evaluation in the* (B, η) *parameter-plane of resonance tongue interaction. Solid lines depict the curves* $B_{(0,1)}$ *and* $B_{(1,n)}$*, and dashed lines represent the stability boundaries for finite* ε*. Shaded regions depict the areas of instability. Insets show blowups of various regions.* (a) $n = 2$ *for* $\varepsilon = 0.02$*;* (b) $n = 3$ *for* $\varepsilon = 0.005$.

The results are presented schematically in Figure 4.1(b). In order to sketch Figure 4.1(b), we have used the fact that (4.9) can be matched into the limits $B_1 \to 0$ and $B_1 \gg 1$. With the correct interpretation, the former limit yields the falling-over instability boundary, whereas the latter yields the first-harmonic resonance tongue, as shown in Appendix I.

4.3. Numerical evaluation. Figure 4.2 shows resonance tongues calculated using numerical Floquet theory with $N = 4$. Two cases of the resonance tongue interaction studied in this section are shown corresponding, respectively, to the cases $n = 2$ (for which $\eta_0 = 53.285$) and $n = 3$ ($\eta_0 = 460.68$) interacting with the fundamental falling-over instability at $B_c = 0.127594$. In both cases, the general shape of the qualitative picture in Figure 4.1 predicted by the above theory is indeed found to occur. Specifically, we find

$$\langle \phi_c'', \phi_c'' \rangle = 12.4182, \tag{4.10}$$

$$\eta = 53.285 : \langle \phi_2'', \phi_2'' \rangle = 3807.01, \quad \langle \phi_c'', \phi_2'' \rangle = 8.9822, \tag{4.11}$$

$$\eta = 460.68 : \langle \phi_3'', \phi_3'' \rangle = 14617.8, \quad \langle \phi_c'', \phi_3'' \rangle = 6.9483. \tag{4.12}$$

Moreover we also found that the mode shape corresponding to each stability boundary is also as predicted by the analysis. The mode shape of the cosine boundary of the $(1, n)$ tongue is found to pick up a large component of the $(0, 1)$ mode as it approaches B_c, whereupon the $\phi_n \cos t$ term starts to diminish in size until as $\eta \to |\infty|$ the mode becomes pure ϕ_c to leading order in ε.

Note from Figure 4.2(a) that there is an interaction with the $(4, 3)$-boundary for an η-value just greater than the critical one, such that at $\ll O(\varepsilon^2)$ these two boundaries exchange positions (see the inset to that figure). Similarly, in Figure 4.2(b), there is an interaction between the $(1, 3)$-mode and the $(5, 3)$ at higher-order in ε (again blown up in an inset). However, the existence of these remarkably thin $(4, 3)$ and $(5, 4)$ resonance tongues makes virtually no difference to the size of the (in)stability region.

5. The effect of damping. Let us now consider the effect on the above analysis of including damping. Recall the assumption made in section 3.3 that both damping and amplitude of excitation are small parameters at $O(\delta)$, but with independent coefficients ε_1 and γ_1 to allow for the possibility of allowing one of these effects to be

zero independent of the other. So we now consider the general asymptotic expansion (3.5)–(3.6) with γ_1 and ε_1 both being $O(1)$ and δ as the perturbation parameter.

5.1. Codimension-one resonances. It is well known that positive linear damping increases the size of stability regions and lifts the root points of resonance tongues off the $\varepsilon = 0$ axis (e.g., [16]). In fact, damping enters at $O(\delta)$ and so makes its first nontrivial contribution in the $O(\delta^2)$ equation derived in section A. Consider taking just a single codimension-one resonance $\eta \equiv \lambda_n/\alpha^2$. Then if $\alpha > 1/2$, we have that the $O(\delta)$ stability equation (B.6) gets replaced with the pair of equations

$$\frac{\partial g}{\partial \tau_1} + \frac{1}{2}\gamma_1 \langle \phi_n'', \phi_n'' \rangle g - K_\alpha \gamma_1 f = 0,$$
$$\frac{\partial f}{\partial \tau_1} + \frac{1}{2}\gamma_1 \langle \phi_n'', \phi_n'' \rangle f + K_\alpha \gamma_1 g = 0,$$

where now,

$$K_\alpha = -B_1 \frac{\alpha \langle \phi_n'', \phi_n'' \rangle}{2\lambda_n}.$$

Note that the eigenvalues of such a system are $-(\gamma_1/2)\langle \phi_n'' \phi_n'' \rangle \pm i K_\alpha$, so that the origin is stable for all B_1 at this order. This shows that the resonance tongue is "lifted off" from the $\varepsilon = 0$ axis by more than an $O(\delta)$ amount (see Figure 5.1(a)).

To get a nontrivial resonance tongue at $O(\delta)$ we must consider the case $\alpha = 1/2$. Then (E.11) and (E.12) become (in the absence of a codimension-two interaction so that $d = e = 0$)

$$\frac{\partial g}{\partial \tau_1} + \left(\frac{B_1 \langle \phi_n'', \phi_n'' \rangle}{4\lambda_n} - \frac{\varepsilon_1}{2} \langle \phi_n, \mathrm{L}\phi_n \rangle \right) f + \frac{\gamma_1}{2} \langle \phi_n'', \phi_n'' \rangle g = 0, \tag{5.1}$$

$$\frac{\partial f}{\partial \tau_1} - \left(\frac{B_1 \langle \phi_n'', \phi_n'' \rangle}{4\lambda_n} + \frac{\varepsilon_1}{2} \langle \phi_n, \mathrm{L}\phi_n \rangle \right) g + \frac{\gamma_1}{2} \langle \phi_n'', \phi_n'' \rangle f = 0. \tag{5.2}$$

When $\varepsilon_1 = 0$ the origin is stable. It becomes unstable along the neutral curve

$${\varepsilon_1}^2 \langle \phi_n, \mathrm{L}\phi_n \rangle^2 = \left(\gamma_1^2 + \frac{B_1^2}{4\lambda_n^2} \right) \langle \phi_n'', \phi_n'' \rangle^2, \tag{5.3}$$

which is depicted in Figure 5.1(b).

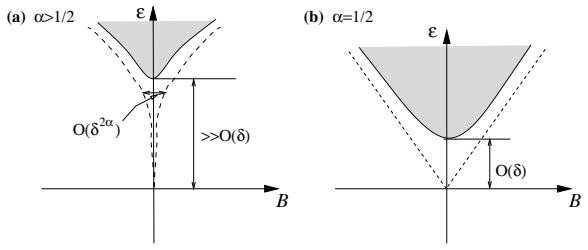


FIG. 5.1. *Sketch of resonance tongues in the presence of nonzero forcing $\varepsilon = \varepsilon_1 \delta$ and damping $\gamma_1 \delta$ (solid line) compared with that of $\gamma_1 = 0$ (dashed). Shaded regions correspond to instability.*

5.2. Codimension-two resonance tongue interaction. We can also include damping in all of the above codimension-two analyses and find its effect on resonance tongue interaction. We shall, however, present only the effect on the calculation in section 4, as this was the most involved and appears the most physically significant to explain the experimental results in [14]. To that end, we consider u_1 given by (4.1) and see the adjustment that the extra linear terms in (3.5)–(3.6) for nonzero γ_1 make on the stability boundaries.

Proceeding as in section 4, at $O(\delta^2)$, when damping is included, (F.8) and (F.9) become

$$\frac{\partial g}{\partial \tau_1} + \frac{\gamma_1}{2}\langle \phi_n'', \phi_n''\rangle g - \left(\frac{\eta_1}{2\lambda_n} - \frac{B_1\langle \phi_n'', \phi_n''\rangle}{2\lambda_n} + \varepsilon_1 \frac{\lambda_n B_c^2 \langle \phi_n'', \phi_c''\rangle^2}{4B_1\langle \phi_c'', \phi_c''\rangle} \right) f = 0,$$

$$\frac{\partial f}{\partial \tau_1} + \frac{\gamma_1}{2}\langle \phi_n'', \phi_n''\rangle f + \left(\frac{\eta_1}{2\lambda_n} - \frac{B_1\langle \phi_n'', \phi_n''\rangle}{2\lambda_n} \right) g = 0,$$

and (F.7) is unchanged.

Writing such a system as

$$\frac{\partial g}{\partial \tau_1} + \Gamma g - ((A/B_1) + C) f = 0, \qquad \frac{\partial g}{\partial \tau_1} + \Gamma f - C g = 0, \tag{5.4}$$

where

$$\Gamma = \frac{\gamma_1}{2}\langle \phi_n'', \phi_n''\rangle, \qquad A = \frac{\varepsilon_1}{4}\frac{\lambda_n B_c \langle \phi_n'', \phi_0''\rangle^2}{\langle \phi_0'', \phi_0''\rangle}, \qquad C = \frac{\eta_1 - B_1\langle \phi_n'', \phi_n''\rangle}{2\lambda_n}, \tag{5.5}$$

we note that $\Gamma > 0$ is a rescaled damping parameter and $A > 0$ is a rescaled amplitude parameter. On the other hand, C can be thought of as a shifted version of η_1 so that the B_1 and C axes, respectively, represent the $O(\varepsilon)$ loci of the bifurcation loci $B_{0,n}$ and $B_{1,m}$.

From (5.4) and (5.5), straightforward calculation shows that the region of stability is given by

$$C(C + A/B_1) + \Gamma^2 > 0, \tag{5.6}$$

which gives the shaded region in Figure 5.2. Note that in the limit $\Gamma \to 0$, we recover the undamped $O(\varepsilon)$ stability curve (4.2) sketched in Figure 4.1(a). Also if $A = 0$

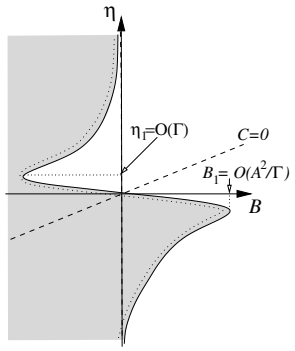


FIG. 5.2. *Schematic figure of the resonance tongue interaction process between the falling-over and first-harmonic instabilities according to* (5.6)*, in the presence of damping* $\gamma_1 > 0$*. Here the* $O(\delta)$ *curve is plotted as a solid line bounding a shaded region of instability in the* (B_1, η_1)*-plane. The dotted curve represents the* $O(\delta^2)$ *correction.*

(corresponding to $\varepsilon_1 = 0$), then we are solving the unforced problem in the presence of nonzero damping; then the only instability is that given by B_1, which is the buckling instability $B_{0,n}$. In between these two extremes we see the curve sketched. Note that the $B_{1,m}$ tongue does not lead to an instability with finite damping at $O(\delta)$ (because of the construction in Figure 5.1), but it does have a very strong influence on the shape of the falling-over instability boundary. Figure 5.2 also shows schematically the result of adding an $O(\delta^2)$ correction to this curve, which can be seen as a secondary effect. See Figure 6.1(b) for actual calculations of the $O(\delta)$ curves, based on the numerical evaluation of Γ, A, and C in (5.5).

6. Experimental comparison. In [14] the quantitative details of the experimental results using a piece of domestic curtain wire are given. A wire of critical buckling length $\ell_c = 55.3\,\text{cm}$ is held in a clamp and subjected to vertical sinusoidal oscillation with peak-to-peak amplitude of 2.2 mm and frequency between 0 and 35 Hz. In terms of the dimensionless parameters of this paper this equates to $\varepsilon = 0.02$ and $0 < \eta < 1000$. The parameter B may be varied by allowing different lengths of wire through the holding clamp. The wire is clearly damped although it is hard to estimate the true value of the dimensionless parameter γ. It is observed that for lengths of wire a little longer than ℓ_c ($B < B_c$), the upright position of the wire is unstable for lower frequencies (η-values), becomes stable at higher η, and becomes unstable again beyond a second η threshold. The nature of the instability at the lower-η stability boundary is a pure falling-over mode ($(0,1)$ in the notation of our theory). The upper boundary is a dynamic instability, at the same frequency of the drive, with a large component of the third spatial mode $(1,3)$. The shape of the stability region is shown in Figure 6.1. There is no evidence of any appreciable subharmonic instability.

Figure 6.1 also compares these results with two separate theories from this paper. First in panel (a) we compare with the results in section 4. This is essentially the same comparison that was shown in [14, Figure 2(b)], although in this paper we have developed a rational explanation for the resonance tongue interaction process that

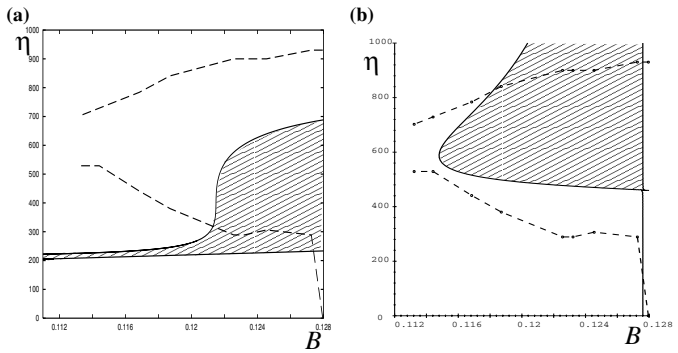


FIG. 6.1. *Comparison between theory (solid line and shading) and the experimental results of* [14] *(dashed line), for the* stability *region as a function of the dimensionless parameters B and η with $\delta = 0.02$, $\varepsilon_1 = 1$. The theory is based on two different calculations close to the resonance tongue interaction between* $(0,1)$ *and* $(1,3)$*, for which $B_{0,1} = B_{1,3}$ when $\eta = \eta_0 = 460.7$.* (a) *Using the theoretical results from Part* II*, which computes the $O(\delta^2)$ coefficients B_2 and B_2^+ given by* (2.18) *and* (2.21) *which undergo singularities in accord with the theory of section* 4 *above. In this case there is no damping.* (b) *Plotting the zeros of* (5.6)*, the $O(\delta)$ coefficient of the resonance tongue interaction in the presence of damping, evaluated using the values* (4.10), (4.12)*, with $\gamma_1 = 0.01$ being used as a representative damping value.*

underlies the shape of the stability region. Note that these results in part explain the experiments, in that the interaction between modes $(0,1)$ and $(1,3)$ leads to a wedge-shaped region in the (B,η)-plane.

Panel (b) compares the experimental data with the results of section 5, where damping is included. Here we have used $\delta = 0.02$ with $\varepsilon_1 = 1$ and have set damping to the plausible value of $\gamma_1 = 0.01$. No other fitting is employed. Also, these results do not include any $O(\varepsilon^2)$ correction to the curves. The agreement is now very good. Note that at this level of δ, the tongue corresponding to the $(1,3)$-instability has zero width. However, its presence is strongly felt in the shape of the instability curve of the $(0,1)$ instability. In particular, the large wedge-shaped region of stability for $B > B_c$ (corresponding to the unshaded wedge within the shaded instability regions in Figure 5.2) is due precisely to the resonance tongue interaction between the buckling instability and the first-harmonic resonance of the third spatial mode.

7. Discussion. The idea that parametric resonance in continuous structures can cause energy to be transferred between modes is a well-established concept, due to the pioneering work of Nayfeh (see [15]) and others (e.g., [6]). This paper has expounded a somewhat different idea, namely, that the combination of resonances corresponding to different spatial modes of a structure can have surprising effects. We have illustrated our results for the canonical model of a straight, vertically mounted elastic column subject to simple sinusoidal parametric excitation. Nevertheless many of the results are likely to have profound implications for other continuous or multi-degree-of-freedom problems subject to parametric excitation.

Let us summarize the main findings. First we studied the undamped problem. The key idea has been to study the genuinely three-parameter problem and to think of *each* degree of freedom as providing a generalized Hill or Mathieu stability diagram in two parameters (B and ε). Each diagram has a buckling (zero-harmonic) instability and a resonance tongue corresponding to every possible multiple of a half-frequency of the drive. The third parameter (η) we think of as sliding each of these diagrams over one another, the result being that instabilities or resonance tongues corresponding to different modes pass through each other; i.e., they *interact*. See Figures 2.2 and 2.3.

Via multiple-scale asymptotic methods, backed up by numerical Floquet theory, we have studied each possible codimension-two interaction in detail. The cases of most interest are when the harmonic corresponding to each of the two tongues differs by unity, that is, in the notation of this paper, when $B_{\alpha,i} = B_{\alpha+1,j}$ for different spatial modes i and j. If $\alpha > 1$, then we have shown that the interaction occurs at first order in the asymptotic parameter ε, such that there is an $O(\varepsilon)$ gap between the two instability tongues, while the width of the tongues themselves remains $\ll O(\varepsilon)$ (see Figure 3.1). In fact, as can be seen from the numerically calculated Figure 2.3(a) in the case of tongues $(5/2,4)$ and $(1/2,2)$, it appears that no two tongues that both correspond to integer (or half-integer) α and β can cross each other, other than at $\varepsilon = 0$. There is always an interaction process where the boundary of one tongue evolves into the boundary of another with the consequent gap being $\ll O(\varepsilon)$ if $\beta \neq \alpha \pm 1$. In this problem, because the mode coupling is through a $\cos t$ term only, tongues where α and β are respectively half-integer and integer (corresponding to the existence of 4π-periodic and 2π-periodic motion, respectively) are unrelated and can pass through each other without interaction (see Figure 2.3).

We then dealt with two special cases. The first was when $\alpha = 1/2$. In this case, the leading-order resonance tongue interaction with modes for which $\beta = 3/2$ occurs

at the same order as the width of the tongue. This leads to the special shape of the two-parameter bifurcation diagram calculated and computed in Figure 3.2.

The second, and most important, special case we studied, in section 4, is when $\alpha = 0$, corresponding to a static buckling instability of the column. Here surprising things happen near η-values for which a pure harmonic instability $\beta = 1$ occurs for the same B-value when $\varepsilon = 0$. In particular we showed that the cosine boundary of the harmonic tongue and the buckling instability curve, which are both ordinarily quadratic in ε, undergo a singularity of their $O(\varepsilon^2)$ coefficients which we have resolved in a new form of asymptotic expansion. The result is the shapes of the instability regions shown in Figures 4.1 and 4.2, which have a large "blob" of stability above the line $B = B_{1,j}$ for $B < B_c$ and a corresponding area of instability below $B = B_{1,j}$ for $B > B_c$. It is this blob of stability which we claim explains the qualitative shape of the stability region in the curtain wire experiment reproduced in Figure 6.1(a).

To get qualitative agreement with the experiments, though, we found it necessary to include material damping. It is well known that linear damping lifts resonance tongues off from zero amplitude in Mathieu-type stability diagrams, and this problem is no exception (Figure 5.1). The key to understanding the experimental parameter regime, though, is to include damping in the interaction between the buckling instability and the harmonic resonance. This leads to the asymptotic results in section 6, which according to Figure 6.1(b) gives good quantitative agreement with the experiments.

Now, as remarked in the conclusions to our earlier papers, we are really only scratching at the surface of the complete nonlinear dynamics of a parametrically excited vertical column. In Part II we showed how to introduce a nonlinear formulation of the problem, which leads to a differential algebraic equation formulation, with the tension in the column acting as a Lagrange multiplier. Also, as remarked in [14], the curtain wire used in the experiment is anything but linearly elastic. Even for the one-degree-of-freedom simple pendulum under parametric excitation, the dynamics of the fully nonlinear system are remarkably rich [17]. For small amplitude nonlinear motion in a neighborhood of the codimension-two instabilities we study, one could presumably infer information on the presence of (both rotating and oscillating) periodic and quasi-periodic solutions from a center manifold and normal form approach, e.g., as in [7]. Yet, even for the linear problem we have studied we have not explored the implications of many finite-amplitude effects such as the "pinching off" of resonance tongues which seemed to occur as a consequence of the interaction between (α, i) and $(\alpha + 1, j)$ resonance tongues for $\alpha > 1$, as was remarked upon in section 3. There are also questions of rigor that we have not addressed. Even for one-degree-of-freedom parametrically excited systems, whether linear stability implies nonlinear stability is a nontrivial question [3]. There are also clearly small-divisor problems, and KAM theory may be able to shed some light; see, for example, the book [12] for infinite dimensional systems.

Nevertheless we hope that this paper has provided a rational and mathematically consistent explanation for the "upside down" stability of a wire under vertical excitation, previously reported in the popular media (e.g., [2]). More seriously, we have explained what we believe to be a new universal mechanism for creation of finite regions of stability (and instability) in parametrically excited systems. Importantly this does *not* involve subharmonic resonance, but rather the interaction of first-harmonic resonance with a steady-state instability.

Appendix A. The general solution at $O(\delta^2)$. Substituting the form (3.7) into the left-hand side of the $O(\delta^2)$ equation (3.5), we obtain

(A.1)
$$\begin{aligned}
&\frac{\lambda_n}{\alpha^2}\frac{\partial^2 u_2}{\partial t^2} + \mathrm{M}_0 u_2 \\
&\quad = \varepsilon_1 \frac{\lambda_n}{2\alpha^2}\mathrm{L}\phi_n \{f[\cos(\alpha+1)t + \cos(\alpha-1)t] + g[\sin(\alpha+1)t + \sin(\alpha-1)t]\} \\
&\qquad + \left\{\eta_1\alpha^2\phi_n f - \frac{2\lambda_n}{\alpha}\frac{\partial g}{\partial \tau_1}\phi_n - B_1\phi_n^{IV} f - \gamma_1\frac{\lambda_n}{\alpha}\phi_n^{IV} g\right\}\cos\alpha t \\
&\qquad + \left\{\eta_1\alpha^2\phi_n g + \frac{2\lambda_n}{\alpha}\frac{\partial f}{\partial \tau_1}\phi_n - B_1\phi_n^{IV} g + \gamma_1\frac{\lambda_n}{\alpha}\phi_n^{IV} f\right\}\sin\alpha t \\
&\qquad + \text{similar expressions with } \alpha \mapsto \beta,\ (f,g) \mapsto (d,e),\ n \mapsto m.
\end{aligned}$$

The particular integral for u_2 can in general be expressed as

(A.2)
$$\begin{aligned}
u_2 &= H_1\cos\alpha t + H_2\sin\alpha t + F_1\cos\beta t + F_2\sin\beta t \\
&\quad + H_3[f\cos(\alpha+1)t + g\sin(\alpha+1)t] + H_4[f\cos(\alpha-1)t + g\sin(\alpha-1)t] \\
&\quad + F_3[d\cos(\beta+1)t + e\sin(\beta+1)t] + F_4[d\cos(\beta-1)t + e\sin(\beta-1)t],
\end{aligned}$$

where

$$M_0H_1 - \lambda_n H_1 = (\eta_1\alpha^2\phi_n - B_1\phi_n^{IV})f - \frac{2\lambda_n}{\alpha}\phi_n\frac{\partial g}{\partial\tau_1} - \gamma_1\frac{\lambda_n}{\alpha}\phi_n g, \tag{A.3}$$

$$M_0H_2 - \lambda_n H_2 = (\eta_1\alpha^2\phi_n - B_1\phi_n^{IV})g + \frac{2\lambda_n}{\alpha}\phi_n\frac{\partial f}{\partial\tau_1} + \gamma_1\frac{\lambda_n}{\alpha}\phi_n f, \tag{A.4}$$

$$M_0H_3 - \frac{\lambda_n(\alpha+1)^2}{\alpha^2}H_3 = \varepsilon_1\frac{\lambda_n}{2\alpha^2}\mathrm{L}\phi_n, \tag{A.5}$$

$$M_0H_4 - \frac{\lambda_n(\alpha-1)^2}{\alpha^2}H_4 = \varepsilon_1\frac{\lambda_n}{2\alpha^2}\mathrm{L}\phi_n, \tag{A.6}$$

$$M_0F_1 - \lambda_m F_1 = (\eta_1\beta^2\phi_m - B_1\phi_m^{IV})d - \frac{2\lambda_m}{\beta}\phi_m\frac{\partial e}{\partial\tau_1} - \gamma_1\frac{\lambda_m}{\beta}\phi_m e, \tag{A.7}$$

$$M_0F_2 - \lambda_m F_2 = (\eta_1\alpha^2\phi_m - B_1\phi_m^{IV})e + \frac{2\lambda_m}{\beta}\phi_m\frac{\partial d}{\partial\tau_1} + \gamma_1\frac{\lambda_m}{\beta}\phi_m d, \tag{A.8}$$

$$M_0F_3 - \frac{\lambda_n(\beta+1)^2}{\alpha^2}F_3 = \varepsilon_1\frac{\lambda_n}{2\alpha^2}\mathrm{L}\phi_m, \tag{A.9}$$

$$M_0F_4 - \frac{\lambda_n(\beta-1)^2}{\alpha^2}F_4 = \varepsilon_1\frac{\lambda_n}{2\alpha^2}\mathrm{L}\phi_m. \tag{A.10}$$

Appendix B. Trivial solution at $O(\varepsilon^2)$. We assume $\alpha, \beta \neq 1/2$. (The case $\alpha = 1/2$ leads to the fact that the $\sin(\alpha-1)t$ and $\cos(\alpha-1)t$ terms are included in the equations for the $\sin\alpha t$ and $\cos\alpha t$ coefficients so that the function H_4 is combined with functions H_1 and H_2). The solvability condition then gives the linear terms in $B_1^{\pm}$ given by (2.20). So, assuming

$$\alpha, \beta \geq 1, \tag{B.1}$$

we find that (A.5) and (A.6) have unique bounded solutions. The solvability condition comes from demanding that the right-hand sides of (A.3) and (A.4) should be orthogonal to ϕ_n, and the right-hand sides of (A.7) and (A.8) orthogonal to ϕ_m, which are eigenfunctions of the respective left-hand side operators. Hence, using the orthonormality of ϕ_n and ϕ_m, we obtain

$$\frac{\partial g}{\partial \tau_1} - (\eta_1\alpha^2 - B_1\langle\phi_n'', \phi_n''\rangle)\frac{\alpha}{2\lambda_n} f = 0, \tag{B.2}$$

$$\frac{\partial f}{\partial \tau_1} + (\eta_1\alpha^2 - B_1\langle\phi_n'', \phi_n''\rangle)\frac{\alpha}{2\lambda_n} g = 0, \tag{B.3}$$

$$\frac{\partial e}{\partial \tau_1} - (\eta_1\beta^2 - B_1\langle\phi_m'', \phi_m''\rangle)\frac{\beta}{2\lambda_m} d = 0, \tag{B.4}$$

$$\frac{\partial d}{\partial \tau_1} + (\eta_1\beta^2 - B_1\langle\phi_m'', \phi_m''\rangle)\frac{\beta}{2\lambda_m} e = 0. \tag{B.5}$$

These four equations can be simplified to read that f and g must both satisfy the equation

$$\frac{\partial^2 f}{\partial \tau_1^2} = -K_\alpha^2 f, \tag{B.6}$$

and both d and e satisfy

$$\frac{\partial^2 e}{\partial \tau_1^2} = -K_\beta^2 e, \tag{B.7}$$

where

$$K_\alpha = (\eta_1\alpha^2 - B_1\langle\phi_n'', \phi_n''\rangle)\frac{\alpha}{2\lambda_n}, \qquad K_\beta = (\eta_1\beta^2 - B_1\langle\phi_m'', \phi_m''\rangle)\frac{\beta}{2\lambda_m}. \tag{B.8}$$

We are looking for stability boundaries, that is, where the functions f, g, d, and e have neutrally stable solutions. Thinking of η as fixed, we find from (B.6)–(B.7) that this happens for an isolated B_1-value for each of (B.6) and (B.7), given by $K_\alpha = 0$ and $K_\beta = 0$, respectively.

Appendix C. The $O(\varepsilon^3)$ equation. Assuming that (B.1) holds and (3.10) is not satisfied, so that (3.9) holds, the $O(\varepsilon^2)$ solution to (3.6) is given by (A.2), where f, g, d, and e are all independent of τ_1. Substitution of this form for u_2 into (3.6) yields

$$\begin{aligned}
&\frac{\lambda_n}{\alpha^2}\frac{\partial^2 u_3}{\partial t^2} + \mathrm{M}_0 u_3 \\
&\quad = \cos(\alpha-2)t\frac{\lambda_n}{2\alpha^2}\mathrm{L}H_4 f \; + \sin(\alpha-2)t\frac{\lambda_n}{2\alpha^2}\mathrm{L}H_4 g \\
&\qquad + \cos(\alpha-1)t\left\{\frac{\lambda_n}{2\alpha^2}\mathrm{L}H_1 f + \eta_1(\alpha_1-1)^2 H_4 f - B_1 H_4^{IV} f + \frac{1}{2}\eta_2 \mathrm{L}\phi_n f\right\} \\
&\qquad + \sin(\alpha-1)t\left\{\frac{\lambda_n}{2\alpha^2}\mathrm{L}H_2 g + \eta_1(\alpha-1)^2 H_4 g - B_1 H_4^{IV} g + \frac{1}{2}\eta_2 \mathrm{L}\phi_n g\right\} \\
&\qquad + \cos\alpha t\left\{\frac{\lambda_n}{2\alpha^2}(\mathrm{L}H_3 + \mathrm{L}H_4) f + \eta_1\alpha^2 H_1 f - B_1 H_1^{IV}\right. \\
&\qquad\qquad \left. + \eta_2\alpha^2\phi_n f - 2\frac{\lambda_n}{\alpha}\phi_n\frac{\partial g}{\partial \tau_2} - B_2\phi_n^{IV} f\right\}
\end{aligned}$$

$$
\begin{aligned}
&+\sin\alpha t\left\{\frac{\lambda_n}{2\alpha^2}(\mathrm{L}H_3 g+\mathrm{L}H_4 g)+\eta_1\alpha^2 H_2 g-B_1 H_2^{IV} g+\eta_2\alpha^2\phi_n g\right.\\
&\qquad\left.+\frac{2\lambda_n}{\alpha}\phi_n\frac{\partial f}{\partial\tau_2}-B_2\phi_n^{IV} g\right\}\\
&+\cos(\alpha+1)t\left\{\frac{\lambda_n}{2\alpha^2}\mathrm{L}H_1 f+\eta_1(\alpha+1)^2 H_3 f-B_1 H_3^{IV} f+\frac{1}{2}\eta_2\mathrm{L}\phi_n f\right\}\\
&+\sin(\alpha+1)t\left\{\frac{\lambda_n}{2\alpha^2}\mathrm{L}H_2 g+\eta_1(\alpha+1)^2 H_3 g-B_1 H_3^{IV} g+\frac{1}{2}\eta_2\mathrm{L}\phi_n g\right\}\\
&+\cos(\alpha+2)t\frac{\lambda_n}{2\alpha^2}\mathrm{L}H_3 f\ +\sin(\alpha+2)t\frac{\lambda_n}{2\alpha^2}\mathrm{L}H_3 g \qquad \text{(C.1)}\\
&+\text{similar expressions with } \alpha\mapsto\beta,\quad (f,g)\mapsto(d,e),\quad n\mapsto m,\quad H_i\mapsto F_i.
\end{aligned}
$$

Appendix D. Codimension-one resonance; derivation of (2.17). In order to obtain the "linearized" results from Parts I and II, valid for undistinguished values of η, we drop all of the β terms in the above and set $\lambda_n/\alpha=\eta$. This describes the asymptotics of resonance tongues away from their codimension-two interactions. Furthermore we set $\eta=\lambda_n/\alpha^2$; i.e., $\eta_1=\eta_2=\cdots 0$. Failure to do this will result in the corrections (2.16) for the η-values that define the resonance condition as B_0 varies. From (3.9) we have, provided $\alpha\neq 1/2$, that $B_1=0$. Consider now the α-dependent terms of (C.1). The solvability condition at this level is that the coefficients of $\cos\alpha t$ and $\sin\alpha t$ must be orthogonal to the eigenfunction ϕ_n. Now, provided that we do *not* satisfy

$$
\alpha\pm 1=\pm\alpha\quad\text{or}\quad\alpha\pm 2=\pm\alpha,\quad\text{i.e., }\alpha=\frac{1}{2}\text{ or }1, \tag{D.1}
$$

this leads to the conditions (making use of (2.10), (2.11))

$$
\frac{\partial f}{\partial\tau_2}=\left[\frac{B_2\alpha}{2\lambda_n}\langle\phi_n'',\phi_n''\rangle-\frac{1}{4\alpha}(\langle\phi_n,\mathrm{L}H_3\rangle+\langle\phi_n,\mathrm{L}H_4\rangle)\right]g, \tag{D.2}
$$

$$
\frac{\partial g}{\partial\tau_2}=-\left[\frac{B_2\alpha}{2\lambda_n}\langle\phi_n'',\phi_n''\rangle-\frac{1}{4\alpha}(\langle\phi_n,\mathrm{L}H_3\rangle+\langle\phi_n,\mathrm{L}H_4\rangle)\right]f. \tag{D.3}
$$

Like (B.2) and (B.3), these equations can be expressed more simply by saying that both f and g satisfy

$$
\frac{\partial f^2}{\partial\tau^2}=-K^2 f,\quad\text{where } K=\frac{B_2\alpha}{2\lambda_n}\langle\phi_n'',\phi_n''\rangle-\frac{1}{4\alpha}(\langle\phi_n,\mathrm{L}H_3\rangle+\langle\phi_n,\mathrm{L}H_4\rangle).
$$

Hence solutions are bounded, apart from at the single neutral stability point $K=0$ given by

$$
B_2=\frac{\lambda_n}{2\alpha^2}\,\frac{\langle\phi_n,\mathrm{L}H_3\rangle+\langle\phi_n,\mathrm{L}H_4\rangle}{\langle\phi_n'',\phi_n''\rangle}. \tag{D.4}
$$

This shows that the width of the resonance tongue is not resolved at this level, but both boundaries of the tongue have the same quadratic coefficient $B_2\varepsilon^2$ given by (D.4).

Now suppose that one of (D.1) *is* satisfied. Consider first $\alpha = 1/2$, which case we have already shown leads to a nontrivial width of resonance tongue at $O(\varepsilon)$: $B = B_0 + \varepsilon B_1^{\pm} + O(\varepsilon^2)$. From the $O(\varepsilon^3)$ equation (C.1), we see that the coefficient of $\cos(\alpha-1)t$ must be added to that of $\cos\alpha t$ when seeking the orthogonality condition, and similarly the coefficient of $\sin(\alpha-1)t$ must be *subtracted* from that of $\sin\alpha t$. This will lead to different equations for B_2 and separate nontrivial corrections to the two boundaries of the resonance tongue $B = B_0 + \varepsilon B_1^{\pm} + \varepsilon^2 B_2^{\pm} + O(\varepsilon^3)$.

Consider now $\alpha = 1$. Here the coefficient of $\cos(\alpha-2)t$ in (C.1) must be added to that of $\cos\alpha t$ when seeking the orthogonality condition, and similarly the coefficient of $\sin(\alpha-2)t$ must be *subtracted* from that of $\sin\alpha t$. This will lead to different equations for B_2 and separate nontrivial corrections to the two boundaries corresponding to $\cos\alpha t$ and $\sin\alpha t$. In fact, it is easy to see that we get $B = B_0 + \varepsilon^2 B_2^{\pm}$, where $B_2^{\pm}$ are given by (2.21), (2.22) with $\eta = \lambda_n$.

Returning to the general case $\alpha \neq 1/2$ or 1, suppose we carry out the expansion to $O(\varepsilon^4)$. We would then get contributions to the particular integral u_4 from terms which come from the expansion of

$$\left\{ \begin{array}{c} \cos \\ \sin \end{array} \right\}(\alpha \pm 1)t \cdot \cos t \quad \text{and} \quad \left\{ \begin{array}{c} \cos \\ \sin \end{array} \right\}(\alpha \pm 2)t \cdot \cos t,$$

which we lead to terms proportional to

$$\left\{ \begin{array}{c} \cos \\ \sin \end{array} \right\}(\alpha \pm 3)t.$$

Therefore, provided we do *not* satisfy

$$\alpha \pm 3 = \pm\alpha, \quad \text{i.e., } \alpha = 3/2 \text{ (since } \alpha > 0),$$

then the orthogonality condition applied to the $\sin\alpha t$ and $\cos\alpha t$ equations leads to a unique condition for B_3, which is third-order correction to both resonance tongues $B = B_0 + \varepsilon^2 B_2 + \varepsilon^3 B_3$. If, however, $\alpha = 3/2$, then we get a different contribution from the $\sin(\alpha-3)t$ and $\cos(\alpha-3)t$ terms, leading to nonequal corrections $B_3^{\pm}$. Hence the first nontrivial width of the resonance tongue is $O(\varepsilon^3)$.

Extrapolating this argument we see that the first nontrivial width of any resonance tongue is always at $O(\varepsilon^{2\alpha})$, and we recover the general expression (2.17). The asymptotic procedure we have developed can in principle calculate all the coefficients B_j, $j = 2, \ldots, 2\alpha - 1$, and $B_{2\alpha}^{\pm}$ for any arbitrary half-integer α, but the expressions become rather cumbersome beyond $O(\varepsilon^3)$.

Appendix E. Neutral curves for $O(\varepsilon^2)$ interaction. The assumption $\beta = \alpha + 1$ leads to the contradiction that the $O(\varepsilon^2)$ solvability condition equation (A.10) has a solution that is an eigenfunction, but the right-hand side is not orthogonal to ϕ_m. Instead we must combine F_4 (which is the coefficient of $d\cos(\beta-1)t$ and $e\sin(\beta-1)t$) into the functions H_1 and H_2 via

$$\tilde{H}_1 = H_1 + F_4, \qquad \tilde{H}_2 = H_2 + F_4.$$

Then $\tilde{H}_1$ and $\tilde{H}_2$ satisfy

$$M_0\tilde{H}_1 - \lambda_n\tilde{H}_1 = \eta_1\alpha^2\phi_n f - B_1\phi_n^{IV} f + \frac{\lambda_n}{2\alpha^2}\mathrm{L}\phi_m d - \frac{\partial g}{\partial \tau_1}\left(\frac{2\lambda_n}{\alpha}\phi_n\right), \tag{E.1}$$

$$M_0\tilde{H}_2 - \lambda_n\tilde{H}_2 = \eta_1\alpha^2\phi_n g - B_1\phi_n^{IV} g + \frac{\lambda_n}{2\alpha^2}\mathrm{L}\phi_m e + \frac{\partial f}{\partial \tau_1}\left(\frac{2\lambda_n}{\alpha}\phi_n\right). \tag{E.2}$$

Similarly, the term H_3 proportional to $f\cos(\alpha+1)t$ and $g\sin(\alpha+1)t$ must be added to the functions F_1 and F_2 via

$$\tilde{F}_1 = F_1 + H_3, \qquad \tilde{F}_2 = F_2 + H_3,$$

where $\tilde{F}_1$ and $\tilde{F}_2$ satisfy

$$M_0\tilde{F}_1 - \lambda_m\tilde{F}_1 = \eta_1\beta^2\phi_m d - B_1\phi_m^{IV}d + \frac{\lambda_m}{2\beta^2}\mathrm{L}\phi_n f - \frac{\partial e}{\partial\tau_1}\left(\frac{2\lambda_m}{\beta}\phi_m\right), \tag{E.3}$$

$$M_0\tilde{F}_2 - \lambda_m\tilde{F}_2 = \eta_1\alpha^2\phi_m e - B_1\phi_m^{IV}e + \frac{\lambda_m}{2\beta^2}\mathrm{L}\phi_n g + \frac{\partial d}{\partial\tau_1}\left(\frac{2\lambda_m}{\beta}\phi_m\right). \tag{E.4}$$

The solvability condition is now that the right-hand sides of (E.1) and (E.2) should both be orthogonal to ϕ_n and that the right-hand sides of (E.3) and (E.4) should be orthogonal to ϕ_m. This gives

$$\frac{\partial g}{\partial\tau_1} - \frac{\alpha}{2\lambda_n}(\eta_1\alpha^2 - B_1\langle\phi_n'',\phi_n''\rangle)f - \frac{1}{4\alpha}\langle\phi_n,\mathrm{L}\phi_m\rangle d = 0, \tag{E.5}$$

$$\frac{\partial f}{\partial\tau_1} + \frac{\alpha}{2\lambda_n}(\eta_1\alpha^2 - B_1\langle\phi_n'',\phi_n''\rangle)g + \frac{1}{4\alpha}\langle\phi_n,\mathrm{L}\phi_m\rangle e = 0, \tag{E.6}$$

$$\frac{\partial e}{\partial\tau_1} - \frac{(\alpha+1)}{2\lambda_m}(\eta_1(\alpha+1)^2 - B_1\langle\phi_m'',\phi_m''\rangle)d - \frac{1}{4(\alpha+1)}\langle\phi_n,\mathrm{L}\phi_m\rangle f = 0, \tag{E.7}$$

$$\frac{\partial d}{\partial\tau_1} + \frac{(\alpha+1)}{2\lambda_m}(\eta_1(\alpha+1)^2 - B_1\langle\phi_m'',\phi_m''\rangle)e + \frac{1}{4(\alpha+1)}\langle\phi_n,\mathrm{L}\phi_m\rangle g = 0. \tag{E.8}$$

From the form of these equations we note that they are expressible as a system

$$\frac{\partial x}{\partial\tau_1} = A(x), \quad \text{where } x = \begin{bmatrix} f \\ d \\ g \\ e \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 0 & -a_1 & -a_2 \\ 0 & 0 & -a_3 & -a_4 \\ a_1 & a_2 & 0 & 0 \\ a_3 & a_4 & 0 & 0 \end{bmatrix},$$

with

$$\begin{aligned} a_1 &= \frac{\alpha}{2\lambda_n}(\eta_1\alpha^2 - B_1\langle\phi_n'',\phi_n''\rangle), \qquad a_2 = \frac{1}{4\alpha}\langle\mathrm{L}\phi_m,\phi_n\rangle, \\ a_3 &= \frac{1}{4(\alpha+1)}\langle\mathrm{L}\phi_m,\phi_n\rangle, \qquad a_4 = \frac{(\alpha+1)}{2\lambda_m}(\eta_1(\alpha+1)^2 - B_1\langle\phi_m'',\phi_m''\rangle). \end{aligned} \tag{E.9}$$

From the form of A we conclude that its eigenvalues are all double and purely imaginary unless

$$a_1a_4 - a_2a_3 = 0. \tag{E.10}$$

Hence the zero solution to the system (E.5)–(E.8) is stable unless we sit precisely on this neutral stability curve defined by (E.10). The fact that all eigenvalues of A are then zero implies that (E.10) is the neutral stability condition for both the cosine mode (corresponding to coefficients f and d) and the sine mode (with coefficients g and e). Hence at this $O(\varepsilon)$ level the resonance tongue has zero width.

It remains to express the neutral stability condition as a curve in the (B_1,η_1)-plane. Substitution of (E.9) into (E.10) results in (3.12), where we have used $\lambda_m = \lambda_n(\alpha+1)^2/\alpha^2$.

Finally, we treat the special case where $\alpha = 1/2$, $\beta = 3/2$. Here, we find that the term $H_3 f \cos(\alpha+1)$ in (A.2) must be added to $\tilde{H}_1$, and the term $H_3 g \sin(\alpha+1)$ subtracted from $\tilde{F}_1$. Hence the solvability condition becomes

$$\frac{\partial g}{\partial \tau_1} - \left(\frac{\eta_1}{16\lambda_n} - \frac{B_1\langle \phi_n'', \phi_n''\rangle}{4\lambda_n} + \frac{\langle \phi_n, \mathrm{L}\phi_n\rangle}{2} \right) f - \frac{\langle \phi_m, \mathrm{L}\phi_n\rangle}{2} d = 0, \tag{E.11}$$

$$\frac{\partial f}{\partial \tau_1} + \left(\frac{\eta_1}{16\lambda_n} - \frac{B_1\langle \phi_n'', \phi_n''\rangle}{4\lambda_n} - \frac{\langle \phi_n, \mathrm{L}\phi_n\rangle}{2} \right) g + \frac{\langle \phi_m, \mathrm{L}\phi_n\rangle}{2} e = 0, \tag{E.12}$$

$$\frac{\partial e}{\partial \tau_1} - \left(\frac{9\eta_1}{48\lambda_n} - \frac{B_1\langle \phi_m'', \phi_m''\rangle}{12\lambda_n} \right) d - \frac{\langle \phi_m, \mathrm{L}\phi_n\rangle}{6} f = 0, \tag{E.13}$$

$$\frac{\partial d}{\partial \tau_1} + \left(\frac{9\eta_1}{48\lambda_n} - \frac{B_1\langle \phi_n'', \phi_n''\rangle}{12\lambda_n} \right) e + \frac{\langle \phi_m, \mathrm{L}\phi_n\rangle}{6} g = 0. \tag{E.14}$$

The neutral mode solution of this equation can be written similarly to (E.10) in the form

$$9(a_1 \pm a_5)a_4 - a_2 a_3 = 0, \quad \text{where } a_5 = \frac{\langle \phi_n, \mathrm{L}\phi_n\rangle}{2}, \tag{E.15}$$

$a_1, \ldots, a_4$ are given by (E.9), and the sign "+" corresponds to the cosine mode and "−" to the sine mode. Expanding (E.15), we obtain the pair of loci in the (B_1, η_1)-plane given by (3.13).

Appendix F. Neutral curves for the first-harmonic buckling interaction. When the solution (4.1) is substituted into the right-hand side of the $O(\varepsilon^2)$ equation (3.5), the result is

$$\begin{aligned} \lambda_n \frac{\partial^2 u_2}{\partial t^2} + \mathrm{M}_0 u_2 = {} & \frac{1}{2}\lambda_n [f(1+\cos 2t) + g \sin 2t] \mathrm{L}\phi_n - B_1 h \phi_c^{IV} \\ & + \left[\left(\eta_1 f - 2\lambda_n \frac{\partial g}{\partial \tau_1} \right) \phi_n + \lambda_n h \mathrm{L}\phi_c - B_1 f \phi_n^{IV} \right] \cos t \\ & + \left[\left(\eta_1 g + 2\lambda_n \frac{\partial f}{\partial \tau_1} \right) \phi_n - B_1 g \phi_n^{IV} \right] \sin t. \end{aligned} \tag{F.1}$$

The particular integral of (F.1) is

$$\begin{aligned} u_2 = \{ & H_0(s, \tau_1, \tau_2) + H_1(s, \tau_1, \tau_2) \cos t + H_2(s, \tau_1, \tau_2) \sin t \\ & + H_3(s)(f \cos 2t + g \sin 2t)\}, \end{aligned} \tag{F.2}$$

where

$$\mathrm{M}_0 H_0 = \frac{1}{2}\lambda_n f \mathrm{L}\phi_n - B_1 h \phi_c^{IV}, \tag{F.3}$$

$$\mathrm{M}_0 H_1 - \lambda_n H_1 = -\left(2\lambda_n \frac{\partial g}{\partial \tau_1} - \eta_1 f \right) \phi_n + \lambda_n h \mathrm{L}\phi_c - B_1 f \phi_n^{IV}, \tag{F.4}$$

$$\mathrm{M}_0 H_2 - \lambda_n H_2 = \left(2\lambda_n \frac{\partial f}{\partial \tau_1} + \eta_1 g \right) \phi_n - B_1 g \phi_n^{IV}, \tag{F.5}$$

$$\mathrm{M}_0 H_3 - 4\lambda_n H_3 = \frac{1}{2}\lambda_n \mathrm{L}\phi_n, \tag{F.6}$$

and H_0, H_1, H_2, and H_3 must satisfy boundary conditions (2.8). Applying orthogonality with respect to ϕ_c to the right-hand side of (F.3), we obtain

$$\frac{1}{2}\lambda_n\langle\phi_c, \mathrm{L}\phi_n\rangle f - B_1\langle\phi_c'', \phi_c''\rangle h = 0,$$

which, using (2.11), we can rewrite as

$$h = -\left(\frac{\lambda_n B_c\langle\phi_n'', \phi_c''\rangle}{2B_1\langle\phi_c'', \phi_c''\rangle}\right) f := -\frac{G_n}{B_1} f. \tag{F.7}$$

Applying orthogonality with respect to ϕ_m to the right-hand sides of (F.4) and (F.5) yields

$$\frac{\partial g}{\partial \tau_1} - \left(\frac{\eta_1}{2\lambda_n} - \frac{B_1\langle\phi_n'', \phi_n''\rangle}{2\lambda_n} + \frac{\lambda_n B_c^2\langle\phi_n'', \phi_c''\rangle^2}{4B_1\langle\phi_c'', \phi_c''\rangle}\right) f = 0, \tag{F.8}$$

$$\frac{\partial f}{\partial \tau_1} + \left(\frac{\eta_1}{2\lambda_n} - \frac{B_1\langle\phi_n'', \phi_n''\rangle}{2\lambda_n}\right) g = 0. \tag{F.9}$$

Thus f and g both satisfy the differential equation

$$\frac{\partial^2 f}{\partial \tau_1^2} + \left(\frac{\eta_1}{2\lambda_n} - \frac{B_1\langle\phi_n'', \phi_n''\rangle}{2\lambda_n}\right)\left(\frac{\eta_1}{2\lambda_n} - \frac{B_1\langle\phi_n'', \phi_n''\rangle}{2\lambda_n} + \frac{\lambda_n B_c^2\langle\phi_n'', \phi_c''\rangle^2}{4B_1\langle\phi_c'', \phi_c''\rangle}\right) f = 0. \tag{F.10}$$

The condition for this equation to have bounded sinusoidal solutions on time scale τ_1 is given by (4.2).

Appendix G. The $O(\varepsilon^3)$ correction to the boundary $\sigma_1 = 0$. On this boundary $f \equiv 0$ so that $h \equiv 0$ by (F.7), and $g = g(\tau_2)$. The right-hand sides of (F.3) and (F.4) are zero in this case so that H_0 and H_1 are proportional to ϕ_c and ϕ_n, respectively. Thus, this part of the $O(\varepsilon^2)$ solution can be absorbed into the $O(\varepsilon)$ solution, and without loss of generality we can set $H_0 = H_1 \equiv 0$. The solution (F.2) can now be written as

$$\left.\begin{aligned} u_1 &= \phi_n(s)\sin t\, g(\tau_2), \\ u_2 &= \{H_2(s)\sin t + H_3(s)\sin 2t\} g(\tau_2), \end{aligned}\right\} \tag{G.1}$$

where $H_3(s)$ is the solution of (F.6) and H_2 satisfies (F.5) with $f = 0$ and the relationship (4.3) holding between B_1 and η_1 in order to ensure orthogonality of the right-hand side to the eigenfunction ϕ_n of the operator on the left.

When the form (G.1) is substituted into the right-hand side of (3.6) one obtains

$$\begin{aligned} &\lambda_n\frac{\partial^2 u_3}{\partial t^2} + \mathrm{M}_0 u_3 \\ &\quad = \left\{\frac{1}{2}\lambda_n \mathrm{L}H_3 + \eta_1\left(H_2 - \frac{1}{\langle\phi_n'', \phi_n''\rangle} H_2^{IV}\right) + \eta_2\phi_n - B_2\phi_n^{IV}\right\} g(\tau_2)\sin t \\ &\qquad -2\lambda_n\phi_n\frac{\partial g}{\partial \tau_2}\cos t + \text{terms involving } \sin 2t \text{ and } \sin 3t, \end{aligned} \tag{G.2}$$

where (4.3) has been used to eliminate B_1 in favor of η_1 from the above equations. Finally, when we require that the right-hand side of this equation be orthogonal to the ϕ_m, we see that g must be independent of τ_2 and that

$$B_2 = \frac{1}{2}\lambda_n\frac{\langle\phi_n, \mathrm{L}H_3\rangle}{\langle\phi_n'', \phi_n''\rangle} - \eta_1\frac{\langle\phi_n, \mathrm{L}H_2\rangle}{\langle\phi_n'', \phi_n''\rangle^2} + \frac{\eta_2}{\langle\phi_n'', \phi_n''\rangle}. \tag{G.3}$$

Note that $\langle \phi_n, H_2 \rangle = 0$ has been used to obtain this result.

From these results the shape of this stability boundary in the space of the original variables B, η, ε in the neighborhood of the singular point $(B_c, \lambda_n, \varepsilon = 0)$ can be constructed as follows.

First, note that from the second power series in (3.3) we may write

$$\begin{aligned} \varepsilon\eta_1 &= (\eta - \lambda_n) - \varepsilon^2\eta_2 + O(\varepsilon^3), \\ \varepsilon^2\eta_1 &= \varepsilon(\eta - \lambda_n) + O(\varepsilon^3). \end{aligned}$$

Now, when these expressions for $\varepsilon\eta_1$ and $\varepsilon^2\eta_1$ are substituted into the above expressions for B_1 and B_2 and the results are substituted into the third series in (3.3), we obtain (4.5).

Appendix H. The $O(\varepsilon^3)$ correction to the boundary $\sigma_2 = 0$. On this boundary $g \equiv 0$, $f = f(\tau_2)$, and

$$h(\tau_2) = -\frac{\lambda_n B_c \langle \phi_n'', \phi_c'' \rangle}{2B_1 \langle \phi_c'', \phi_c'' \rangle} f(\tau_2) := -\frac{G_n}{B_1} f(\tau_2). \tag{H.1}$$

We require the solution to $O(1)$ even as $B_1 \to 0$ or ∞. Hence we set

$$k(\tau_2) := \sqrt{f^2 + g^2}, \quad \text{so that } f = \frac{B_1}{K_n} k, \qquad h = \frac{-G_n}{K_n} k, \quad \text{where } K_n = \sqrt{B_1^2 + G_n^2}.$$

Also by reasoning similar to that above we may set $H_2 \equiv 0$ so that on this boundary we have

$$\left.\begin{aligned} u_1 &= (-G_n \phi_c(s) + B_1 \phi_n(s) \cos t) \frac{k(\tau_2)}{K_n}, \\ u_2 &= \{H_0(s) + H_1(s) \cos t + H_3(s) \cos 2t\} k(\tau_2), \end{aligned}\right\} \tag{H.2}$$

where in this case

$$\mathrm{M}_0 H_0 = \frac{B_1}{K_n} \left(\frac{1}{2} \lambda_n \mathrm{L}\phi_n + G_n \phi_c^{IV} \right), \tag{H.3}$$

$$\mathrm{M}_0 H_1 - \lambda_n H_1 = \frac{1}{K_n} (\eta_1 B_1 \phi_n - \lambda_n G_n \mathrm{L}\phi_c - B_1^2 \phi_n^{IV}), \tag{H.4}$$

$$\mathrm{M}_0 H_3 - 4\lambda_n H_3 = \frac{B_1 \lambda_n}{2K_n} \mathrm{L}\phi_n. \tag{H.5}$$

When the results (H.2) are substituted into the right-hand side of (3.6) we obtain

$$\begin{aligned} \lambda_n \frac{\partial^2 u_3}{\partial t^2} + \mathrm{M}_0 u_3 = &\left\{ \frac{1}{2} \lambda_n \mathrm{L} H_1 + \frac{B_1}{2K_n} \eta_1 \mathrm{L}\phi_n - B_1 H_0^{IV} + B_2 \frac{G_n}{K_n} \phi_c^{IV} \right\} k(\tau_2) \\ &+ \left\{ \lambda_n \left(\mathrm{L} H_0 + \frac{1}{2} \mathrm{L} H_3 \right) + \eta_1 \left(H_1 - \frac{G_n}{K_n} \mathrm{L}\phi_c \right) \right. \\ &\left. - B_1 H_1^{IV} + \frac{B_1}{K_n} \eta_2 \phi_n - \frac{B_1}{K_n} B_2 \phi_n^{IV} \right\} k(\tau_2) \cos t \\ &+ 2\lambda_n \phi_n \frac{B_1}{K_n} \frac{\partial k}{\partial \tau_2} \sin t + \text{ terms involving } \cos 2t \text{ and } \cos 3t. \end{aligned} \tag{H.6}$$

Applying orthogonality conditions to the various groups of terms on the right-hand side of this equation, we see that f is independent of τ_2 and hence obtain the expressions (4.6) and (4.7).

Appendix I. Matching to the codimension-one results. We consider the expression (4.9) in the two limits (i) $B_1 \to 0$ (and hence $|\hat{\eta}_1| \gg 1$, but $\ll 1/\varepsilon$) and (ii) $B_1 \gg 1$ (and hence $\hat{\eta}_1 \to 0$, but $B_1 \ll 1/\varepsilon$).

(i) Consider first $B_1 \to 0$. Taking B_1 as a small parameter, then (4.8) shows $\hat{\eta_1} = O(1/B_1)$. The leading-order term of (4.9) we then find to be in the B_2-direction and to be given by

$$\tilde{B}_2 = B_2 - \hat{\eta}_2 \frac{B_1}{\hat{\eta}_1} = -\frac{\lambda_n}{2}\langle \phi_c, \mathrm{L}H_1\rangle - (1+B_c)G_n\langle \phi_c'', \phi_n''\rangle + O(B_1), \tag{I.1}$$

where H_1 satisfies

$$\mathrm{M}_0 H_1 - \lambda_n H_1 = -2\lambda_n B_c \langle \phi_n'', \phi_c''\rangle \phi_n - \lambda_n \mathrm{L}\phi_c + O(B_1^2). \tag{I.2}$$

Now we are interested in matching to the asymptotics of the pure falling-over instability, away from the codimension-two interaction, for which ϕ_m is not a resonant mode. Setting ϕ_m to zero in (I.1) and (I.2) we obtain precisely (dropping the tilde)

$$B_2 = \frac{\lambda_n}{2}\frac{\langle \mathrm{L}H_1, \phi_c\rangle}{\langle \phi_c'', \phi_c''\rangle},$$

which is precisely the boundary B_2^+ defined by (2.21) (with η replaced by its $O(1)$ value λ_m, for the same function H_1, where now ϕ_m is called ϕ_c). Hence we recover the correct asymptotic expression well away from the resonance tongue interaction.

(ii) Now consider $B_1 \gg 1$. Then (4.8) shows $\hat{\eta}_1 = O(1/B_1)$. Then we find the leading-order term of (4.9) to be in the $\hat{\eta}$-direction and to be given by

$$\tilde{\hat{\eta}}_2 = \hat{\eta}_2 + O(1/B_1),$$

where $\hat{\eta}_2$ is given by the right-hand side of (4.7) without the B_2-term. Now, to match to the limit of just the $(1,n)$ resonance, away from its interaction with the falling-over instability, we must set $\phi_c = 0$, since this is no longer resonant, and also set $H_1 = 0$, since it is now resonant (it defines the coefficient of $\cos t$) and therefore can be subsumed into the $O(1)$ solution. We then obtain to leading order that (dropping tildes)

$$\hat{\eta}_2 = \eta_2 - B_2\langle \phi_n'', \phi_n''\rangle = -\lambda_n\left(\langle \mathrm{L}H_0, \phi_n\rangle + \frac{1}{2}\langle LH_3, \phi_n\rangle\right).$$

Hence, allowing η_2 to be zero, to model the codimension-one case, we get

$$B_2 = \frac{2\lambda_n}{\langle \phi_n'', \phi_n''\rangle}(2\langle \mathrm{L}H_0, \phi_n\rangle + \langle LH_3, \phi_n\rangle),$$

where now

$$\mathrm{M}_0 H_3 - 4\lambda_n H_3 = \frac{\lambda_n}{2}\mathrm{L}\phi_n,$$
$$\mathrm{M}_0 H_0 = \frac{\lambda_n}{2}\mathrm{L}\phi_n.$$

Hence we see that this is identical to that defined by (2.21) (with H_4 now called H_0), and we obtain perfect matching in this case also.

REFERENCES

[1] D. Acheson, *From Calculus to Chaos, An Introduction to Dynamics*, Oxford University Press, Oxford, UK, 1997.

[2] D. Acheson and T. Mullin, *Ropy magic*, New Scientist, 157 (1998), pp. 32–33.

[3] M. V. Bartuccelli, G. Gentile, and K. V. Georgiou, *On the stability of the upside-down pendulum with damping*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 458 (2002), pp. 255–269.

[4] H. Broer and M. Levi, *Geometrical aspects of stability theory for Hill's equations*, Arch. Rational Mech. Anal., 131 (1995), pp. 225–240.

[5] A. Champneys and W. Fraser, *The "Indian rope trick" for a continuously flexible rod: Linearised analysis*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 456 (2000), pp. 553–570.

[6] J. Cusumano and F. Moon, *Chaotic and non-planar vibrations of the thin elastica*, J. Sound Vibration, 179 (1995), pp. 185–226.

[7] G. Dangelmayr, B. Fiedler, K. Kirchgässner, and A. Mielke, *Dynamics of Nonlinear Waves in Dissipative Systems: Reduction, Bifurcation and Stability*, Pitman Research Notes in Mathematics Series 352, Longman, Harlow, UK, 1996.

[8] W. Fraser and A. Champneys, *The "Indian rope trick" for a parametrically excited flexible rod: Nonlinear and subharmonic analysis*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 458 (2002), pp. 1353–1373.

[9] J. Galan, W. Fraser, D. Acheson, and A. Champneys, *The parametrically excited upside-down rod: An elastic jointed pendulum model*, J. Sound Vibration, to appear.

[10] A. Greenhill, *Determination of the greatest height consistent with stability that a vertical pole or mast can be made ...*, Proceedings of the Cambridge Philosophical Society, 4 (1881), pp. 65–73.

[11] D. W. Jordan and P. Smith, *Nonlinear Ordinary Differential Equations*, 2nd ed., Oxford University Press, Oxford, UK, 1986.

[12] S. Kuksin, *Nearly Integrable Infinite-Dimensional Hamiltonian Systems*, Lecture Notes in Math. 1556, Springer-Verlag, Berlin, 1993.

[13] H.-B. Mühlhaus, H. Sakaguchi, and B. Hobbs, *Evolution of* $3d$ *folds for a non-Newtonian plate in a viscous medium*, Philos. Trans. Roy. Soc. London Ser. A, 454 (1998), pp. 3121–3143.

[14] A. Mullin, T. Champneys, W. Fraser, J. Galan, and D. Acheson, *The "Indian wire trick" via parametric excitation: A comparison between theory and experiment*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 459 (2003), pp. 539–546.

[15] A. Nayfeh, *Nonlinear Interactions: Analytical, Computational and Experimental Methods*, Wiley Interscience, New York, 2000.

[16] A. Nayfeh and D. Mook, *Nonlinear Oscillations*, Wiley Interscience, New York, 1979.

[17] M. van Noort, *The Parametrically Forced Pendulum. A Case Study in* $1\frac{1}{2}$ *Degree of Freedom*, Ph.D. thesis, RU Groningen, Groningen, The Netherlands, 2001.

SIAM J. APPL. MATH.
Vol. 65, No. 1, pp. 299–315

BOUNDS AND EXTREMAL CONFIGURATIONS FOR THE TORSIONAL RIGIDITY OF COATED FIBER REINFORCED SHAFTS*

ROBERT LIPTON† AND TUNGYANG CHEN‡

Abstract. In this paper we derive bounds on the torsional rigidity for coated fiber reinforced shafts. The bounds are used to assess the optimality or suboptimality of fiber reinforcement configurations. This investigation focuses on coated fiber reinforcements with circular cross section. It is shown how the effective antiplane shear modulus and torsional rigidity of each coated fiber are used to determine whether the configuration provides reinforcement above or below that of a homogeneous shaft containing no coated fibers. Simply connected shaft cross sections of arbitrary shape reinforced with any configuration of coated fibers are considered. Precise conditions on the effective antiplane shear modulus and torsional rigidity of each coated fiber are given under which the circular shaft reinforced with a single centered circular coated fiber is either optimal or suboptimal.

Key words. torsion, coated fibers

AMS subject classifications. 35J20, 74Q99

DOI. 10.1137/S0036139903424229

1. Introduction.

The problem of extremizing the torsional rigidity of prismatic shafts has been the focus of many investigations. For homogeneous shafts made from elastically isotropic material, de Saint-Venant [10] proposed that among all prismatic shafts with a given cross-sectional area that the greatest torsional rigidity is obtained by a shaft with a circular cross section. This proposition was proven by Polya [7]. For multiply connected cross sections of a given cross-sectional area, Polya and Weinstein [8] showed that the optimal cross section is given by the annulus. Alvino and Trombetti [1] considered composite shaft cross sections made up of perfectly bonded elastic materials. Here each phase is a cylindrical fiber of arbitrary cross section with generators parallel to the shaft. In this context they showed that circular cross sections with a radially nonincreasing arrangement of compliance delivers the maximum torsional rigidity among all cross sections with given cross-sectional area and fixed area fraction of the constituent phases.

When the materials are imperfectly bonded the elastic displacement may suffer jumps across the interface betweendifferent elastic phases. To first order one models the imperfect bonding in terms of a linear constitutive law relating tangential stress to the jump in the warping displacement. This model for imperfect bonding is well known and is referred to as the spring layer model; see Jones and Whittier [4]. In this context one considers shafts reinforced with fibers of greater shear stiffness than the matrix. One is interested in extremizing the torsional rigidity over fiber configurations and understanding how the imperfect interface compromises the benefits of the stiffer reinforcement. It is found that the degree of imperfect bonding relative to the contrast in compliance between matrix and fiber explicitly determines the type of

*Received by the editors March 7, 2003; accepted for publication (in revised form) January 29, 2004; published electronically October 28, 2004.

http://www.siam.org/journals/siap/65-1/42422.html

†Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803 (lipton@math.lsu.edu). The work of this author was supported by the Air Force Office of Scientific Research under grant F49620-02-1-0041 and by the NSF through grant DMS-0296064.

‡Department of Civil Engineering, National Cheng Kung University, Tainan 70101, Taiwan (tchen@mail.ncku.edu.tw). The work of this author was supported by the National Science Council, Taiwan under contract NSC 90-2211-E006-119.

fiber configuration that maximizes the torsional rigidity; see Lipton [5, Theorems 1.1 through 1.7]. The relative degree of imperfect bonding is given by the parameter

$$\mathrm{R}_{cr} = \frac{\alpha^{-1}}{\mathrm{G}_m^{-1} - \mathrm{G}_f^{-1}}, \tag{1.1}$$

where G_f is the shear modulus of the fiber reinforcement, G_m is the shear modulus of the matrix, and α is the interfacial shear stiffness having dimensions of shear stiffness per unit length. For a shaft with a circular cross section of radius R containing N reinforcement fibers of circular cross section with common radii given by R_{cr}, the imperfect interface balances the reinforcing effect of the fibers and the warping function outside the fibers is precisely zero. For this case the torsional rigidity is independent of the location of the fibers and is given by

$$\frac{\pi \mathrm{G_m}}{2}(R^4 - N\mathrm{R}_{cr}^4) + \frac{\pi \mathrm{G_f}}{2} N\mathrm{R}_{cr}^4 \tag{1.2}$$

and is precisely the torsional rigidity of a circular shaft of radius R reinforced with a single centered fiber of radius $N^{1/4}\mathrm{R}_{cr}$; see Lipton [5].

In many composites a third phase or inter-phase separating fiber and matrix is present. The inter-phase or coating phase often has elastic properties that are distinct from the fiber or matrix. In this context the recent work of Chen, Benveniste, and Chuang [2] treats a system of N fibers with circular cross section and radii a_i, $i = 1, \ldots, N$. The fibers are coated by a shell of uniform thickness and the outer radius of the coated fiber is b_i, $i = 1, \ldots, N$. The shear modulus of the ith fiber is denoted by G_f^i and the shear modulus of the associated coating is denoted by G_c^i. The area fraction of the fiber phase in the ith coated fiber system is denoted by ν_i and $\nu_i = a_i^2/b_i^2$. One recalls the formula for the effective antiplane shear modulus for the concentric coated cylinders assemblage of Hashin and Rosen [3] given by

$$G_{CCA}^i = \mathrm{G}_c^i \left(\frac{\mathrm{G}_c^i(1-\nu_i) + \mathrm{G}_f^i(1+\nu_i)}{\mathrm{G}_c^i(1+\nu_i) + \mathrm{G}_f^i(1-\nu_i)} \right). \tag{1.3}$$

Here G_{CCA}^i gives the effective shear stiffness of each coated fiber. Chen, Benveniste, and Chuang [2] show that when the effective shear stiffness of each coated fiber equals the matrix shear stiffness G_m, i.e.,

$$G_{CCA}^i = \mathrm{G}_m, \;\; i = 1, \ldots, N, \tag{1.4}$$

then the warping function outside the coated fibers is zero and the torsional rigidity is given by

$$\mathcal{A}^N = \frac{\pi}{2}\mathrm{G}_m R^4 + \sum_{i=1}^{N} \left(\frac{\pi}{2}(\mathrm{G}_c^i(b_i^4 - a_i^4) + \mathrm{G}_f^i a_i^4) - \frac{\pi}{2}\mathrm{G}_m\, b_i^4 \right). \tag{1.5}$$

When all fibers have the same radius and coating thickness ℓ one passes to the distinguished limit given by

$$\lim_{\ell \to 0} \lim_{\mathrm{G}_c^i \to 0} \frac{\ell}{\mathrm{G}_c^i} = \alpha^{-1} \tag{1.6}$$

in (1.4) and (1.5) to see that $\mathcal{A}^N$ is given by (1.2).

The relations given by (1.4) express the balance between the shear moduli of the matrix, fiber, coating, and coating thickness that renders the warping function zero outside the inclusions. Furthermore, under the hypotheses leading to (1.5) it is evident that if the torsional rigidity of each coated fiber given by

$$T_f^i = \frac{\pi}{2}(\mathrm{G}_c^i(b_i^4 - a_i^4) + \mathrm{G}_f^i a_i^4) \tag{1.7}$$

equals the torsional rigidity $\frac{\pi}{2}\mathrm{G}_m\, b_i^4$, obtained by replacing coating and fiber shear moduli with the matrix shear moduli, then there is complete neutrality; i.e., the torsional rigidity equals the torsional rigidity of the unreinforced shaft given by $\frac{\pi}{2}\mathrm{G}_m R^4$ (see Chen, Benveniste, and Chuang [2]). A recent summary of results involving neutral inclusions in the context of the theory of effective properties is given in Milton [6].

In this article we examine the effect of the coating phase on the torsional rigidity of coated fiber reinforced shafts. We build on the previous results and develop a variational methodology to assess the optimality or suboptimality of coated fiber configurations. Here the cross section of each coated fiber is taken to be circular, the radius of the ith fiber cross section is denoted by a_i, and the outer radius of the coating is given by b_i. The union of the coated fibers is denoted by A. The remaining part of the cross section containing matrix material is denoted by A_m. The shaft cross section is denoted by Ω and $\Omega = A \cup A_m$. The results given in this paper follow easily from a set of bounds on the torsional rigidity derived using the variational principles given by (2.1) and (2.2).

We provide a brief outline of the bounds derived in this paper. Upper and lower bounds on the torsional rigidity for shafts with circular cross section reinforced with coated fibers are given in Proposition 2.1. These bounds are given in terms of the effective shear moduli and torsional rigidity of each coated fiber. Next we consider shafts with arbitrary simply connected cross section. Here upper bounds are given in terms of the polar moment of inertia of the shaft cross section $I_0(\Omega)$ and the effective shear moduli and torsional rigidity of each coated fiber; see Proposition 3.1. If, in addition, one knows that $\mathrm{G}_c^i \le \mathrm{G}_f^i$ for $i = 1, \ldots, N$, then it is shown that one can derive a tighter upper bound given in terms of the torsional rigidity $\mathcal{T}_0(\Omega)$ of the shaft cross section and the effective shear moduli and torsional rigidity of each coated fiber; see Proposition 5.2. When $\mathrm{G}_c^i \ge \mathrm{G}_f^i$ for $i = 1, \ldots, N$, a lower bound is derived and is given in terms of $\mathcal{T}_0(\Omega)$ and the effective shear moduli and torsional rigidity of each coated fiber; see Proposition 6.2.

The bounds are used to establish the three reinforcement inequalities and three geometric inequalities presented in section 2. The reinforcement inequalities provide explicit criteria that determine when the torsional rigidity of a single coated fiber centered inside a shaft with circular cross section is either optimal or suboptimal among all coated fiber configurations for shafts with cross sections satisfying prescribed isoperimetric constraints; see Propositions 2.2, 2.3, and 2.4. The geometric inequalities provide explicit criteria that determine when the torsional rigidity of the coated fiber reinforced shaft is either greater than or less than the torsional rigidity of the same shaft in the absence of the coated fiber reinforcement; see Propositions 2.5, 2.6, and 2.7. In all cases the optimality conditions are expressed in terms of the effective shear modulus and torsional rigidity of each coated fiber.

2. Inequalities on the torsional rigidity. We begin by introducing the variational formulations for the torsional rigidity used in the subsequent analysis. The

torsional rigidity for a system of N coated fibers inside a shaft with cross section Ω is denoted by $\mathcal{T}^N(\Omega)$. Points inside Ω are denoted by $\mathbf{x} = (x_1, x_2)$, and the coordinate system is chosen such that the origin lies inside Ω. The first variational principle is given in terms of virtual stress potentials φ that vanish on the boundary of the shaft cross section that are square integrable and have square integrable gradients. It is given by

$$\mathcal{T}^N(\Omega) = -2 \min_{\varphi} \left\{ \frac{1}{2} \int_\Omega G^{-1}(\mathbf{x}) |\nabla \varphi|^2 \, d\mathbf{x} \, - \, 2 \int_\Omega \varphi \, d\mathbf{x} \right\}, \tag{2.1}$$

where the piecewise constant shear modulus $G(\mathbf{x})$ is G_m in the matrix and takes the values G_f^i and G_c^i in the ith fiber and coating, respectively. Next we define the vector $\mathbf{x}^\perp$ to be given by $(-x_2, x_1)$. The second variational principle is given in terms of virtual warping functions $\tilde{w}$ that are square integrable and have square integrable gradients. It is given by

$$\mathcal{T}^N(\Omega) = \min_{\tilde{w}} \left\{ \int_\Omega G(\mathbf{x}) |\nabla \tilde{w} + \mathbf{x}^\perp|^2 \, d\mathbf{x} \right\}. \tag{2.2}$$

Motivated by (1.4) and (1.5), we start by considering shafts with a circular cross section of radius R. For this case we denote the shaft cross section by D_R. The torsional rigidity of D_R reinforced with N coated fibers is written as $\mathcal{T}^N(D_R)$. Here the coordinates are chosen such that the center of the shaft is the origin. The method presented here is simple. The trial fields are designed so that they become the actual stress potential or warping field in the composite when $G^i_{CCA} = \mathrm{G}_m$ for $i = 1, \ldots, N$. Otherwise, these fields are admissible trials and when substituted into the variational principles give upper and lower bounds on the torsional rigidity. In this way the upper and lower bounds match when $G^i_{CCA} = \mathrm{G}_m$ for $i = 1, \ldots, N$. For a system of N coated fibers with centers located at the points $\mathbf{x}_i$, $i = 1, \ldots, N$, the bounds are given by the following proposition.

PROPOSITION 2.1.

$$\begin{aligned} \mathcal{A}^N + \pi \sum_{i=1}^N |\mathbf{x}_i|^2 b_i^2 \frac{\mathrm{G}_m}{G^i_{CCA}} (G^i_{CCA} - \mathrm{G}_m) &\le \mathcal{T}^N(D_R) \\ &\le \mathcal{A}^N + \pi \sum_{i=1}^N |\mathbf{x}_i|^2 b_i^2 (G^i_{CCA} - \mathrm{G}_m), \end{aligned} \tag{2.3}$$

where $\mathcal{A}^N$ is given by (1.5). *The upper and lower bounds agree when $G^i_{CCA} = \mathrm{G}_m$ for $i = 1, \ldots, N$.*

These upper and lower bounds are derived in sections 3 and 4, respectively.

In what follows we apply Proposition 2.1 to obtain the basic reinforcement inequality for shafts of circular cross section reinforced with a finite number N of coated fibers. Here we suppose that the shear moduli of each fiber and coating are the same, i.e., $\mathrm{G}_f^i = \mathrm{G}_f$ and $\mathrm{G}_c^i = \mathrm{G}_c$. In addition, it is supposed that the ratio of outer and inner coating radius is the same for each coated fiber, i.e., $\nu_i = \nu$, $i = 1, \ldots, N$. For this case $G^i_{CCA} = G_{CCA}$, where

$$G_{CCA} = \mathrm{G}_c \left(\frac{\mathrm{G}_c(1-\nu) + \mathrm{G}_f(1+\nu)}{\mathrm{G}_c(1+\nu) + \mathrm{G}_f(1-\nu)} \right) \tag{2.4}$$

and for $G_{CCA} = \mathrm{G}_m$ the torsional rigidity given by (1.5) becomes

$$\mathcal{A}^N = \overline{\mathcal{A}} = \frac{\pi}{2}\left(\mathrm{G}_m R^4 - \mathrm{G}_m \overline{b}^4 + \mathrm{G}_c \overline{b}^4(1-\nu^2) + \mathrm{G}_f \nu^2 \overline{b}^4\right), \tag{2.5}$$

where $\overline{b}^4 = \sum_{i=1}^N b_i^4$. Here $\overline{\mathcal{A}}$ is precisely the torsional rigidity of a single coated fiber with outer coating radius $\overline{b}$ and fiber radius $\overline{a} = \nu^{1/2}\overline{b}$ when the centers of the coated fiber cross section and shaft cross section are the same. The torsional rigidity of the concentric coated fiber shaft configuration is given by the right-hand side of (2.5) for all values of G_m, G_f, G_c, and $\overline{a} \le \overline{b} \le R$. We note here that the area of the fiber cross section is given by $\pi\overline{a}^2 = \pi\sqrt{\sum_{i=1}^N a_i^4}$.

The following reinforcement inequalities follow from Proposition 2.1 and give conditions for which the concentric coated fiber and circular shaft cross section is either optimal or suboptimal.

PROPOSITION 2.2 (reinforcement inequalities I). *If $G_{CCA} \le \mathrm{G}_m$, then the torsional rigidity associated with N coated fibers is less than or equal to the rigidity associated with a single centered circular coated fiber with fiber radius $\overline{a} = \nu^{1/2}\overline{b}$, i.e.,*

$$\mathcal{T}^N(D_R) \le \overline{\mathcal{A}}. \tag{2.6}$$

Otherwise, if $G_{CCA} \ge \mathrm{G}_m$, then the torsional rigidity associated with N coated fibers is greater than or equal to that of a single centered circular coated fiber with fiber radius $\overline{a} = \nu^{1/2}\overline{b}$, i.e.,

$$\mathcal{T}^N(D_R) \ge \overline{\mathcal{A}}. \tag{2.7}$$

These inequalities are independent of the number and location of the coated fibers.

When all fibers have the same radius a and coating thickness ℓ, one easily passes to the distinguished limit given by (1.6) in Proposition 2.2 to recover Theorem 1.3 of Lipton [5] for imperfectly bonded fiber reinforced shafts.

Next we consider the more general case where the shaft can have an arbitrary simply connected cross section Ω. Here we consider all configurations of N coated fibers with prescribed fiber radii a_i, $i = 1, \ldots, N$, and consider all cross sections Ω with prescribed polar moment of inertia. We apply the upper bound on the torsional rigidity given by Proposition 3.1 to obtain the following.

PROPOSITION 2.3 (reinforcement inequality II). *Consider any shaft with polar moment of inertia with respect to the origin equal to $\pi R^4/2$ reinforced with N circular coated fibers. If $G_{CCA} \le \mathrm{G}_m$, then the torsional rigidity $\mathcal{T}^N(\Omega)$ is less than or equal to the torsional rigidity associated with a shaft with circular cross section of radius R reinforced with a single centered circular coated fiber with fiber radius $\overline{a}$ given by*

$$\pi\overline{a}^2 = \pi\sqrt{\sum_{i=1}^N a_i^4} \tag{2.8}$$

and $\overline{b} = \nu^{-1/2}\overline{a}$.

When $\mathrm{G}_c \le \mathrm{G}_f$ we can appeal to the tighter upper bound on the torsional rigidity given by Proposition 5.2 to obtain a reinforcement inequality that holds for all shaft cross sections Ω with prescribed cross-sectional area.

PROPOSITION 2.4 (reinforcement inequality III). *Consider any shaft with cross-sectional area equal to πR^2 reinforced with N circular coated fibers. If $G_{CCA} \le \mathrm{G}_m$*

and $\mathrm{G}_c \leq \mathrm{G}_f$*, then the torsional rigidity* $\mathcal{T}^N(\Omega)$ *is less than or equal to the torsional rigidity associated with a shaft with a circular cross section of radius* R *reinforced with a single centered circular coated fiber with fiber radius* $\overline{a}$ *given by*

$$\pi\overline{a}^2 = \pi\sqrt{\sum_{i=1}^{N} a_i^4} \tag{2.9}$$

and $\overline{b} = \nu^{-1/2}\overline{a}$*.*

It is evident from the inequality $\sqrt{\sum_{i=1}^N a_i^4} \leq \sum_{i=1}^N a_i^2$ that the cross-sectional area of the single centered circular fiber appearing in Propositions 2.2, 2.3, and 2.4 is less than or equal to the joint cross-sectional area of the N fibers.

Now we consider the more general case where the shear moduli of the fiber and coating and the ratio of the inner radius and outer radius of the coating are allowed to differ between coated fibers. In this context we present explicit conditions on the effective shear modulus and torsional rigidity of each coated fiber that show when the torsional rigidity of the coated fiber reinforced shaft is either greater or less than the torsional rigidity of the shaft without reinforcement.

For shafts with circular cross sections of radius R, i.e., $\Omega = D_R$, we have the following.

PROPOSITION 2.5 (geometric inequalities I). *If* $\sum_{i=1}^N T_f^i \leq \sum_{i=1}^N \frac{\pi}{2}\mathrm{G}_m\, b_i^4$ *and* $G_{CCA}^i \leq \mathrm{G}_m$*, then*

$$\mathcal{T}^N(D_R) \leq \frac{\pi}{2}\mathrm{G}_m R^4. \tag{2.10}$$

If $\sum_{i=1}^N T_f^i \geq \sum_{i=1}^N \frac{\pi}{2}\mathrm{G}_m\, b_i^4$ *and* $G_{CCA}^i \geq \mathrm{G}_m$*, then*

$$\mathcal{T}^N(D_R) \geq \frac{\pi}{2}\mathrm{G}_m R^4. \tag{2.11}$$

The inequalities (2.10) *and* (2.11) *are independent of the number and location of the coated fibers.*

These inequalities follow immediately from Proposition 2.1.

Now we extend these results to simply connected cross sections Ω and denote the torsional rigidity for simply connected shaft cross sections with shear modulus unity by $\mathcal{T}_0(\Omega)$. The following geometric inequality shows when a system of coated fibers always decreases the torsional rigidity below that of the unreinforced shaft.

PROPOSITION 2.6 (geometric inequality II). *Suppose that* $\mathrm{G}_c^i \leq \mathrm{G}_f^i$*,* $i=1,\ldots,N$*. If* $\sum_{i=1}^N T_f^i \leq \sum_{i=1}^N \frac{\pi}{2}\mathrm{G}_m\, b_i^4$ *and* $G_{CCA}^i \leq \mathrm{G}_m$*, then*

$$\mathcal{T}^N(\Omega) \leq \mathrm{G}_m\mathcal{T}_0(\Omega). \tag{2.12}$$

The equality holds in (2.12) *when the shaft cross section is circular,* $\sum_{i=1}^N T_f^i = \sum_{i=1}^N \frac{\pi}{2}\mathrm{G}_m\, b_i^4$ *and* $G_{CCA}^i = \mathrm{G}_m$ *for* $i = 1,\ldots,N$*.*

This result follows from the upper bound on the torsional rigidity given by Proposition 5.2.

The following geometric inequality shows when a system of coated fibers always increases the torsional rigidity above that of the unreinforced shaft.

PROPOSITION 2.7 (geometric inequality III). *Suppose that* $\mathrm{G}_c^i \geq \mathrm{G}_f^i$*,* $i=1,\ldots,N$*. If* $\sum_{i=1}^N T_f^i \geq \sum_{i=1}^N \frac{\pi}{2}\mathrm{G}_m\, b_i^4$ *and* $G_{CCA}^i \geq \mathrm{G}_m$*, then*

$$\mathcal{T}^N(\Omega) \geq \mathrm{G}_m\mathcal{T}_0(\Omega). \tag{2.13}$$

The equality holds in (2.13) *when the shaft cross section is circular,* $\sum_{i=1}^{N} T_f^i = \sum_{i=1}^{N} \frac{\pi}{2} \mathrm{G}_m\, b_i^4$ *and* $G_{CCA}^i = \mathrm{G}_m$ *for* $i = 1, \ldots, N$.

This result follows from the lower bound on the torsional rigidity given by Proposition 6.2.

3. Upper bounds on the torsional rigidity for shafts reinforced with circular coated fibers. In this section we develop trial warping functions for configurations of circular coated fibers. These are substituted into the variational principle (2.2) and deliver the upper bound presented in Proposition 2.1. The trial warping functions constructed here will be admissible for shaft cross sections of any shape. For circular shaft cross sections it is shown that the trial warping functions become the actual warping displacement in the shaft when $G_{CCA}^i = \mathrm{G}_m$ for $i = 1, \ldots, N$.

Consider a shaft of arbitrary cross section Ω reinforced with N circular coated fibers with centers at the points $\mathbf{x}_i$, $i = 1, \ldots, N$. The radius of the ith fiber is a_i, and the outer radius of the coated fiber is b_i. The coating occupies the annular shell with inner and outer radii a_i and b_i. The trial warping function $\tilde{w}$ is chosen such that $\tilde{w} = 0$ outside the coated fibers. In each coated fiber the function $\tilde{w}$ is required to be harmonic inside the fiber and harmonic inside the coating. It is required that $\tilde{w}$ be continuous across the interface separating the fiber and coating and

$$\mathrm{G}_f^i(\nabla \tilde{w}_{|_f} + \mathbf{x}^{\perp}) \cdot \mathbf{n} = \mathrm{G}_c^i(\nabla \tilde{w}_{|_c} + \mathbf{x}^{\perp}) \cdot \mathbf{n} \tag{3.1}$$

across the fiber–coating interface. Here the subscripts indicate the side of the interface over which the quantities are evaluated and $\mathbf{n}$ is the outward directed unit normal in the fiber–coating interface. The final requirement is that $\tilde{w}$ vanish on the boundary of the coated fiber. It is clear that the continuity conditions for $\tilde{w}$ at material interfaces ensure that it is an admissible trial field for (2.2).

We solve the transmission boundary value problem inside each coated fiber to obtain the explicit formula for $\tilde{w}$. The polar coordinates (θ, r) are chosen such that the axis $\theta = 0$ coincides with the direction given by $\mathbf{x}^{\perp}{}_i$ and origin with $\mathbf{x}_i$. In these coordinates, the transmission condition (3.1) on the ith fiber–coating interface becomes

$$\mathrm{G}_c^i \partial_r \tilde{w}_{|_c} - \mathrm{G}_f^i \partial_r \tilde{w}_{|_f} = (\mathrm{G}_f^i - \mathrm{G}_c^i)|\mathbf{x}_i| \cos\theta \text{ on } r = a_i. \tag{3.2}$$

Since $\tilde{w}$ is required to be harmonic inside each fiber and coating it follows that

$$\tilde{w} = C_1 r \cos\theta \text{ for } r \le a_i \tag{3.3}$$

and

$$\tilde{w} = (C_2 r + C_3 r^{-1}) \cos\theta \text{ for } a_i \le r \le b_i. \tag{3.4}$$

The transmission conditions at $r = a_i$ and boundary condition at $r = b_i$ require that

$$\begin{gathered} C_2 b_i^2 + C_3 = 0, \\ C_1 a_i^2 - C_2 a_i^2 - C_3 = 0, \\ \mathrm{G}_c^i a_i^2 C_2 - \mathrm{G}_c^i C_3 - \mathrm{G}_f^i a_i^2 C_1 = (\mathrm{G}_f^i - \mathrm{G}_c^i) a_i^2 |\mathbf{x}_i|. \end{gathered} \tag{3.5}$$

The solution of (3.5) shows that inside each coated fiber the trial warping function is given by

$$\tilde{w} = C_1^i r \cos\theta \text{ for } r = |\mathbf{x} - \mathbf{x}_i| \le a_i \tag{3.6}$$

and

$$\tilde{w} = (C_2^i r + C_3^i r^{-1})\cos\theta \text{ for } a_i \le r = |\mathbf{x} - \mathbf{x}_i| \le b_i, \tag{3.7}$$

where $\Delta_i = \mathrm{G}_c^i(a_i^2 + b_i^2) + \mathrm{G}_f^i(b_i^2 - a_i^2)$ and

$$\begin{aligned} C_1^i &= (\mathrm{G}_c^i - \mathrm{G}_f^i)|\mathbf{x}_i|(b_i^2 - a_i^2)/\Delta_i, \\ C_2^i &= (\mathrm{G}_f^i - \mathrm{G}_c^i)|\mathbf{x}_i|a_i^2/\Delta_i, \\ C_3^i &= (\mathrm{G}_c^i - \mathrm{G}_f^i)|\mathbf{x}_i|b_i^2 a_i^2/\Delta_i. \end{aligned} \tag{3.8}$$

Outside the coated fibers $\tilde{w} = 0$.

The polar moment of inertia of the shaft cross section Ω with respect to the origin is written $I_0(\Omega)$. Here $I_0(\Omega) = \int_\Omega |\mathbf{x}|^2 d\mathbf{x}$. Substitution of $\tilde{w}$ into (2.2) delivers the upper bound given in the following.

PROPOSITION 3.1 (upper bound on rigidity for arbitrary shaft cross section).

$$\begin{aligned} \mathcal{T}^N(\Omega) \le \mathrm{G}_m I_0(\Omega) + \sum_{i=1}^N \Big(\frac{\pi}{2} T_f^i - \frac{\pi}{2}\mathrm{G}_m\, b_i^4\Big) \\ + \pi \sum_{i=1}^N |\mathbf{x}_i|^2 b_i^2 (G_{CCA}^i - \mathrm{G}_m). \end{aligned} \tag{3.9}$$

Next we consider shafts with a circular cross section of radius R. In order for the trial warping field $\tilde{w}$ to be the actual warping displacement in the shaft it must also satisfy the transmission condition on the coating–matrix interface $|\mathbf{x} - \mathbf{x}_i| = b_i$ given by

$$\mathrm{G}_m \mathbf{x}^\perp \cdot \mathbf{n} = \mathrm{G}_c^i(\nabla \tilde{w}_{|_c} + \mathbf{x}^\perp) \cdot \mathbf{n}. \tag{3.10}$$

This gives the extra condition

$$\mathrm{G}_c^i C_2^i b_i^2 - \mathrm{G}_c^i C_3^i = (\mathrm{G}_m - \mathrm{G}_c^i)|\mathbf{x}_i| b_i^2. \tag{3.11}$$

This condition together with the conditions given by (3.5) provide an overdetermined system of equations for the coefficients C_1^i, C_2^i, C_3^i in each coated fiber. It is easily seen that the overdetermined system has a solution when $G_{CCA}^i = \mathrm{G}_m$. For this case the function $\tilde{w}$ becomes the warping displacement in the shaft and we recover the formula

$$\mathcal{T}^N(D_R) = \mathcal{A}^N, \tag{3.12}$$

where $\mathcal{A}^N$ is given by (1.5).

4. Lower bounds on the torsional rigidity for circular shafts reinforced with circular coated fibers. In this section we develop trial stress potentials for configurations of circular coated fibers. These are substituted into the variational principle (2.1) to obtain the lower bound given in Proposition 2.1.

We consider a circular shaft cross section of radius R reinforced with N coated fibers. Outside the coated fibers the trial stress potential φ is taken to be $\varphi = \frac{1}{2}\mathrm{G}_m(R^2 - |\mathbf{x}|^2)$. The trial potential is taken to be continuous across the matrix–coating interface specified by $|\mathbf{x} - \mathbf{x}_i| = b_i$. It is easily seen that

$$\varphi = h(\mathbf{x}) = -\mathrm{G}_m(\mathbf{x} - \mathbf{x}_i)\cdot \mathbf{x}_i + \frac{1}{2}\mathrm{G}_m(R^2 - b_i^2 - |\mathbf{x}_i|^2) \tag{4.1}$$

on this interface in view of the condition $|\mathbf{x}-\mathbf{x}_i| = b_i$. The trial is taken to be continuous inside the coated fiber and is given by $\varphi = \psi^i + r^i$ in the ith fiber. Here ψ^i is chosen to be the stress potential generated inside the coated fiber when it is subject to torsion loading. It is the solution of the transmission problem inside the coated fiber given by

$$\mathrm{G}_f^{i\,-1}(x)\Delta\psi^i = -2 \text{ in the fiber, } |\mathbf{x}-\mathbf{x}_i| < a_i, \tag{4.2}$$

$$\mathrm{G}_c^{i\,-1}(x)\Delta\psi^i = -2 \text{ in the coating, } a_i < |\mathbf{x}-\mathbf{x}_i| < b_i. \tag{4.3}$$

ψ^i is continuous across the fiber–coating interface,

$$\mathrm{G}_f^{i\,-1}\nabla\psi^i_{|_f}\cdot\mathbf{n} = \mathrm{G}_c^{i\,-1}\nabla\psi^i_{|_c}\cdot\mathbf{n} \text{ on } |\mathbf{x}-\mathbf{x}_i| = a_i, \tag{4.4}$$

and $\psi^i = 0$ on $|\mathbf{x}-\mathbf{x}_i| = b_i$. It is easily seen that ψ^i is given by

$$\begin{aligned}
\psi^i &= -\frac{1}{2}\left(\mathrm{G}_f^i|\mathbf{x}-\mathbf{x}_i|^2 - \mathrm{G}_c^i(b_i^2-a_i^2) - \mathrm{G}_f^i a_i^2\right) \text{ for } |\mathbf{x}-\mathbf{x}_i| < a_i,\\
\psi^i &= -\frac{1}{2}\left(\mathrm{G}_c^i|\mathbf{x}-\mathbf{x}_i|^2 - \mathrm{G}_c^i b_i^2\right) \text{ for } a_i < |\mathbf{x}-\mathbf{x}_i| < b_i.
\end{aligned} \tag{4.5}$$

The function $r^i = h$ on the coating–matrix interface and is continuous inside the coated fiber. It is the solution to the transmission problem given by

$$\mathrm{G}_f^{i\,-1}(x)\Delta r^i = 0 \text{ in the fiber, } |\mathbf{x}-\mathbf{x}_i| < a_i, \tag{4.6}$$

$$\mathrm{G}_c^{i\,-1}(x)\Delta r^i = 0 \text{ in the coating, } a_i < |\mathbf{x}-\mathbf{x}_i| < b_i, \tag{4.7}$$

and

$$\mathrm{G}_f^{i\,-1}\nabla r^i_{|_f}\cdot\mathbf{n} = \mathrm{G}_c^{i\,-1}\nabla r^i_{|_c}\cdot\mathbf{n} \text{ on } |\mathbf{x}-\mathbf{x}_i| = a_i. \tag{4.8}$$

In the polar coordinates (θ, r) chosen such that the axis $\theta = 0$ coincides with the vector $\mathbf{x}_i$ and $r = |\mathbf{x}_i - \mathbf{x}|$, the solution of the transmission problem for r^i is given by

$$\begin{aligned}
r^i &= C_1^i r\cos\theta + k^i \text{ for } |\mathbf{x}-\mathbf{x}_i| < a_i,\\
r^i &= \left(C_2^i r + C_3^i r^{-1}\right)\cos\theta + k^i \text{ for } a_i < |\mathbf{x}-\mathbf{x}_i| < b_i,
\end{aligned} \tag{4.9}$$

where

$$\begin{aligned}
k^i &= \frac{\mathrm{G}_m}{2}\left(R^2 - b_i^2 - |\mathbf{x}_i|^2\right),\\
C_1^i &= -\mathrm{G}_m|\mathbf{x}_i|b_i^2 2\mathrm{G}_f^i/D_i,\\
C_2^i &= -\mathrm{G}_m|\mathbf{x}_i|b_i^2(\mathrm{G}_f^i + \mathrm{G}_c^i)/D_i,\\
C_3^i &= -\mathrm{G}_m|\mathbf{x}_i|a_i^2 b_i^2(\mathrm{G}_f^i - \mathrm{G}_c^i)/D_i,
\end{aligned} \tag{4.10}$$

and $D_i = (b_i^2+a_i^2)\mathrm{G}_f^i + (b_i^2-a_i^2)\mathrm{G}_c^i$. The lower bound in (2.3) follows from substitution of this trial potential into the variational principle (2.1).

In order for the trial potential field φ to be the actual stress potential in the shaft it must also satisfy the transmission condition on the coating–matrix interface given by

$$\mathrm{G}_m^{-1}\nabla\varphi_{|_m}\cdot\mathbf{n} = {\mathrm{G}_c^i}^{-1}(\nabla\psi^i_{|_c}+\nabla r^i_{|_c})\cdot\mathbf{n}. \tag{4.11}$$

Substitution and working in polar coordinates show that (4.11) gives the extra condition

$$ {\mathrm{G}_c^i}^{-1}\left(C_2^i - C_3^i b_i^{-2}\right) = -|\mathbf{x}_i|. \tag{4.12}$$

This condition together with the system of equations (4.10) overdetermines the coefficients C_1^i, C_2^i, C_3^i in each coated sphere. It is easily seen that the overdetermined system has a solution when $G^i_{CCA} = \mathrm{G}_m$. For this case the function φ becomes the stress potential in the shaft and we recover the formula

$$\mathcal{T}^N(D_R) = \mathcal{A}^N, \tag{4.13}$$

where $\mathcal{A}^N$ is given by (1.5).

5. Upper bounds on the torsional rigidity for $\mathbf{G_c^i \le G_f^i}$. In this section we focus on the case where $\mathrm{G}_c^i \le \mathrm{G}_f^i$, $i = 1, \ldots, N$. Here we are able to get tighter upper bounds on the torsional rigidity for shaft cross sections of arbitrary shape. Our approach follows the methodology developed in Lipton [5]. We fix the cross section of the shaft Ω and investigate the effects of adding a circular coated fiber to an already existing configuration of $N-1$ coated fibers. At present no assumptions on the geometry or shear moduli of the $N-1$ coated fibers are made. We denote the part of the shaft cross section already occupied by the coated fibers by A and the cross section of the circular coated fiber to be added by Σ. Here Σ is composed of a circular fiber of radius a_N with shear modulus G_f^N surrounded by a coating of outer radius b_N with shear modulus G_c^N. The torsional rigidity of the original configuration is denoted by $\mathcal{T}(A,\Omega)$. The rigidity associated with the added fiber is written as $\mathcal{T}(A\cup\Sigma,\Omega)$. We recall that the torsional rigidity obtained by replacing coating and fiber shear moduli with the matrix shear moduli in Σ is given by $\frac{\pi}{2}\mathrm{G}_m\, b_N^4$. Here b_N is the outer radius of the coating. The torsional rigidity of the coated fiber is $T_f^N = \frac{\pi}{2}(\mathrm{G}_c^N(b_N^4 - a_N^4) + \mathrm{G}_f^N a_N^4)$.

PROPOSITION 5.1 (upper rigidity inequality). *If* $\mathrm{G}_c^N \le \mathrm{G}_f^N$ *and if*

$$G^N_{CCA} \le \mathrm{G}_m, \tag{5.1}$$

then

$$\mathcal{T}(A\cup\Sigma,\Omega) \le \mathcal{T}(A,\Omega) + T_f^N - \frac{\pi}{2}\mathrm{G}_m\, b_N^4. \tag{5.2}$$

Proposition 5.1 is established with the aid of the variational principle given by (2.1). We remark that the methods used to establish this inequality apply to the case when the fiber cross section is multiply connected. One writes (2.1) as $\mathcal{T}(A,\Omega) = -2\mathcal{E}(A,\Omega)$, where

$$\mathcal{E}(A,\Omega) = \min_{\varphi}\left\{\frac{1}{2}\int_\Omega G^{-1}(\mathbf{x})|\nabla\varphi|^2\,d\mathbf{x} \;-\; 2\int_\Omega \varphi\,d\mathbf{x}\right\}. \tag{5.3}$$

Here the piecewise constant shear modulus $G(\mathbf{x})$ takes the value G_m in the matrix and takes the values G_f^i and G_c^i in the ith fiber and coating, respectively. The idea of the proof is to estimate the quantity $\mathcal{E}(A,\Omega)$ in terms of $\mathcal{E}(A\cup\Sigma,\Omega)$ associated with the additional fiber. We let $\mathcal{G}(\mathbf{x})$ denote the piecewise constant shear modulus for the configuration $A\cup\Sigma$. Here $\mathcal{G}(\mathbf{x})=G(\mathbf{x})$ outside of Σ and inside Σ the shear modulus $\mathcal{G}(\mathbf{x})=\mathrm{G}_f^N$ in the fiber and $\mathcal{G}(\mathbf{x})=\mathrm{G}_c^N$ in the coating. We regroup terms in the variational principle (5.3) and write

$$\begin{aligned}\mathcal{E}(A,\Omega)=\min_{\varphi}\Bigg\{&\frac{1}{2}\int_\Omega \mathcal{G}^{-1}(\mathbf{x})|\nabla\varphi|^2 d\mathbf{x}-2\int_\Omega\varphi d\mathbf{x}\\&+\frac{1}{2}\left(\int_\Sigma(\mathrm{G}_m^{-1}-\mathcal{G}^{-1}(\mathbf{x}))\right)|\nabla\varphi|^2 d\mathbf{x}\Bigg\}.\end{aligned}\tag{5.4}$$

We obtain an estimate by substitution of a suitable trial field in (5.4). Our choice is made as follows: We introduce the stress potential $\tilde{\Phi}$ for the configuration $A\cup\Sigma$. Here $\tilde{\Phi}$ is continuous in Ω,

$$-\mathcal{G}^{-1}(\mathbf{x})\Delta\tilde{\Phi}=2,\tag{5.5}$$

and satisfies the transmission conditions

$$\mathrm{G}_f^{i\,-1}\nabla\tilde{\Phi}_{|_f}\cdot\mathbf{n}=\mathrm{G}_c^{i\,-1}\nabla\tilde{\Phi}_{|_c}\cdot\mathbf{n}\ \text{on the fiber–coating interface}\ |\mathbf{x}-\mathbf{x}_i|=a_i\tag{5.6}$$

and

$$\mathrm{G}_f^{i\,-1}\nabla\tilde{\Phi}_{|_c}\cdot\mathbf{n}=\mathrm{G}_m{}^{-1}\nabla\tilde{\Phi}_{|_m}\cdot\mathbf{n}\ \text{on the matrix–coating interface}\ |\mathbf{x}-\mathbf{x}_i|=b_i.\tag{5.7}$$

The trial field φ is chosen to match $\tilde{\Phi}$ outside the coated fiber cross section Σ but inside we suppose that $\varphi=\tilde{\Phi}+\delta$, where δ is continuous, vanishes on the boundary of Σ, is square integrable inside Σ, and has a square integrable gradient over Σ. One easily checks that

$$\int_\Sigma\mathcal{G}^{-1}(\mathbf{x})\nabla\tilde{\Phi}\cdot\nabla\delta\,d\mathbf{x}=2\int_\Sigma\delta\,d\mathbf{x}.\tag{5.8}$$

Substitution of φ into (5.4) gives

$$\begin{aligned}\mathcal{E}(A,\Omega)\le\frac{1}{2}&\left(\int_{\Omega/\Sigma}\mathcal{G}^{-1}(\mathbf{x})|\nabla\tilde{\Phi}|^2 d\mathbf{x}+\int_\Sigma\mathcal{G}^{-1}(\mathbf{x})|\nabla\tilde{\Phi}+\nabla\delta|^2 d\mathbf{x}\right)\\&-2\int_\Omega\tilde{\Phi}d\mathbf{x}-2\int_\Sigma\delta d\mathbf{x}\\&+\frac{1}{2}\left(\int_\Sigma(\mathrm{G}_m^{-1}-\mathcal{G}^{-1}(\mathbf{x}))\right)|\nabla\varphi|^2 d\mathbf{x}.\end{aligned}\tag{5.9}$$

We apply (5.8) and expand the second term on the right-hand side of (5.9) to find

$$\int_\Sigma\mathcal{G}^{-1}(\mathbf{x})|\nabla\tilde{\Phi}+\nabla\delta|^2 d\mathbf{x}=\int_\Sigma\mathcal{G}^{-1}(\mathbf{x})|\nabla\tilde{\Phi}|^2 d\mathbf{x}+\int_\Sigma\mathcal{G}^{-1}(\mathbf{x})|\nabla\delta|^2 d\mathbf{x}+4\int_\Sigma\delta d\mathbf{x}.\tag{5.10}$$

Substitution of (5.10) into (5.9) yields

$$\mathcal{E}(A,\Omega) \leq \mathcal{E}(A\cup\Sigma,\Omega) + \frac{1}{2}\int_\Sigma \mathcal{G}^{-1}(\mathbf{x})|\nabla\delta|^2 d\mathbf{x} + \frac{1}{2}\left(\int_\Sigma (\mathrm{G}_m^{-1} - \mathcal{G}^{-1}(\mathbf{x}))\right)|\nabla\varphi|^2 d\mathbf{x}. \tag{5.11}$$

Multiplying by -2 and arranging terms, we find that

$$\mathcal{T}(A\cup\Sigma,\Omega) \leq \mathcal{T}(A,\Omega) + \int_\Sigma \mathcal{G}^{-1}(\mathbf{x})|\nabla\delta|^2 d\mathbf{x} + \int_\Sigma (\mathrm{G}_m^{-1} - \mathcal{G}^{-1}(\mathbf{x}))|\nabla\varphi|^2 d\mathbf{x}. \tag{5.12}$$

Next we minimize the right-hand side of (5.12) with respect to δ to obtain

$$\mathcal{T}(A\cup\Sigma,\Omega) \leq \mathcal{T}(A,\Omega) + \mathcal{U}, \tag{5.13}$$

where

$$\mathcal{U} = \int_\Sigma (\mathrm{G}_m^{-1} - \mathcal{G}^{-1}(\mathbf{x}))\nabla\hat{\varphi}\cdot\nabla\tilde{\Phi}\, d\mathbf{x}. \tag{5.14}$$

Here $\hat{\varphi} = \hat{\delta} + \tilde{\Phi}$ in Σ and $\hat{\delta}$ solves

$$\int_\Sigma \mathrm{G}_m^{-1}\nabla\hat{\delta}\cdot\nabla u\, d\mathbf{x} = \int_\Sigma (\mathcal{G}^{-1}(\mathbf{x}) - \mathrm{G}_m^{-1})\nabla\tilde{\Phi}\cdot\nabla u\, d\mathbf{x} \tag{5.15}$$

for every trial u vanishing on the boundary of Σ. From (5.15) and (5.5)–(5.7) we see that $\hat{\varphi}$ solves

$$\int_\Sigma \mathrm{G}_m^{-1}\nabla\hat{\varphi}\cdot\nabla u\, d\mathbf{x} = 2\int_\Sigma u\, d\mathbf{x} \tag{5.16}$$

for every trial u vanishing on the boundary of Σ. This is equivalent to the differential equation $-\mathrm{G}_m^{-1}\Delta\hat{\varphi} = 2$ over Σ.

We decompose the trial $\hat{\varphi}$ into two parts: $\hat{\varphi} = r + \psi^h$, where the function r satisfies

$$\Delta r = 0 \text{ in } \Sigma \text{ and } r = \tilde{\Phi} \text{ on the boundary of } \Sigma, \tag{5.17}$$

and ψ^h satisfies

$$\Delta\psi^h = -2\mathrm{G}_m \text{ in } \Sigma \text{ and } \psi^h = 0 \text{ on the boundary of } \Sigma. \tag{5.18}$$

Next we decompose $\tilde{\Phi}$ into two components over Σ. We write $\tilde{\Phi} = \psi - h$. Here ψ is continuous, vanishes on the boundary of Σ, and solves the torsion problem

$$-\mathcal{G}^{-1}(\mathbf{x})\Delta\psi = 2 \tag{5.19}$$

with the transmission condition

$$(\mathrm{G}_f^N)^{-1}\nabla\psi_{|_f}\cdot\mathbf{n} = (\mathrm{G}_c^N)^{-1}\nabla\psi_{|_c}\cdot\mathbf{n} \text{ on the fiber–coating interface } |\mathbf{x}-\mathbf{x}_N| = a_N. \tag{5.20}$$

The function h is continuous on Σ and $h = -\tilde{\Phi}$ on the boundary of Σ. It is the solution of

$$-\mathcal{G}^{-1}(\mathbf{x})\Delta h = 0 \tag{5.21}$$

and h satisfies the transmission condition

$$(\mathrm{G}_f^N)^{-1}\nabla h_{|_f} \cdot \mathbf{n} = (\mathrm{G}_c^N)^{-1}\nabla h_{|_c} \cdot \mathbf{n} \text{ on the fiber–coating interface } |\mathbf{x} - \mathbf{x}_N| = a_N. \tag{5.22}$$

Substitution of the functions ψ^h, ψ, r, and h into $\mathcal{U}$ and (5.13) gives

$$\begin{aligned}\mathcal{T}(A \cup \Sigma, \Omega) \leq \mathcal{T}(A, \Omega) + T_f^N - \frac{\pi}{2}\mathrm{G}_m\, b_N^4 \\ + \int_\Sigma \mathrm{G}_m^{-1}|\nabla r|^2\, d\mathbf{x} - \int_\Sigma \mathcal{G}^{-1}(\mathbf{x})|\nabla h|^2\, d\mathbf{x} - 4\int_\Sigma r + h\, d\mathbf{x}.\end{aligned} \tag{5.23}$$

For circular fiber cross sections calculation shows that $\int_\Sigma r + h\, d\mathbf{x} = 0$ and we obtain

$$\begin{aligned}\mathcal{T}(A \cup \Sigma, \Omega) \leq \mathcal{T}(A, \Omega) + T_f^N - \frac{\pi}{2}\mathrm{G}_m\, b_N^4 \\ + \int_\Sigma \mathrm{G}_m^{-1}|\nabla r|^2\, d\mathbf{x} - \int_\Sigma \mathcal{G}^{-1}(\mathbf{x})|\nabla h|^2\, d\mathbf{x}.\end{aligned} \tag{5.24}$$

It is clear that Proposition 5.1 holds when the indefinite term

$$\begin{aligned}D &= \int_\Sigma \mathrm{G}_m^{-1}|\nabla r|^2\, d\mathbf{x} - \int_\Sigma \mathcal{G}^{-1}(\mathbf{x})|\nabla h|^2\, d\mathbf{x} \\ &= \mathrm{G}_m^{-1}\left(\int_\Sigma |\nabla r|^2\, d\mathbf{x} - \int_\Sigma \frac{\mathrm{G}_m}{\mathcal{G}(\mathbf{x})}|\nabla h|^2\, d\mathbf{x}\right) \leq 0.\end{aligned} \tag{5.25}$$

If $\tilde{\Phi} = \text{const}$ on the boundary of Σ, then $r = \text{const}$ and $h = -\text{const}$ and $D = 0$. We now examine conditions for which $D \leq 0$ and $r \neq \text{const}$ and $h \neq -\text{const}$. To do this we search for the largest number β for which

$$\beta \int_\Sigma |\nabla r|^2\, d\mathbf{x} - \int_\Sigma \frac{\mathrm{G}_m}{\mathcal{G}(\mathbf{x})}|\nabla h|^2\, d\mathbf{x} \leq 0 \tag{5.26}$$

for every choice of r and h such that $h = -r$ on the boundary of Σ, r is harmonic inside Σ, and h is harmonic in the fiber and in the coating, and satisfies the transmission conditions

$$\frac{\mathrm{G}_m}{\mathrm{G}_f^N}\nabla h_{|_f} \cdot \mathbf{n} = \frac{\mathrm{G}_m}{\mathrm{G}_c^N}\nabla h_{|_c} \cdot \mathbf{n} \tag{5.27}$$

on the fiber–coating interface. The set of all such r and h for which $r \neq \text{const}$ and $h \neq \text{const}$ is denoted by $\mathcal{C}$. The largest β is given by

$$\hat{\beta} = \mathrm{G}_m \inf_{\mathcal{C}} \frac{\int_\Sigma \mathcal{G}(\mathbf{x})^{-1}|\nabla h|^2\, d\mathbf{x}}{\int_\Sigma |\nabla r|^2\, d\mathbf{x}}. \tag{5.28}$$

The stationary values for the quotient given in (5.28) are denoted by β_n, and the stationary conditions for the stationary functions (r_n, h_n) in $\mathcal{C}$ are given by

$$(\mathrm{G}_c^N)^{-1}\nabla h_n \cdot \mathbf{n} = -\beta_n \nabla r_n \cdot \mathbf{n} \tag{5.29}$$

on the coating matrix boundary $|\mathbf{x}-\mathbf{x}_N| = b_N$. Choosing polar coordinates (θ, r) such that the $\theta = 0$ axis is along $\mathbf{x}_N$ and $r = |\mathbf{x}-\mathbf{x}_N|$, one finds that the stationary functions are given by

$$\begin{aligned} r_n &= K_1^n r^n \exp(jn\theta) \text{ for } 0 \le r \le b_N, \\ h_n &= (K_2^n r^n) \exp(jn\theta) \text{ for } 0 \le r \le a_N, \\ h_n &= \left(K_3^n r^n + K_4^n r^{-n}\right) \exp(jn\theta) \text{ for } a_N \le r \le b_N. \end{aligned} \tag{5.30}$$

Here $j = \sqrt{-1}$ and both real and imaginary parts of r_n and h_n are stationary functions. The constants K_1^n are arbitrary and the remaining constants are given by

$$\begin{aligned} K_2^n &= -K_1^n \left(\frac{2\mathrm{G}_f^N}{\mathrm{G}_f^N - \mathrm{G}_c^N}\right) \frac{b_N^{2n}}{b_N^{2n}\frac{\mathrm{G}_c^N+\mathrm{G}_f^N}{\mathrm{G}_f^N-\mathrm{G}_c^N} + a_N^{2n}}, \\ K_3^n &= -K_1^n \left(\frac{\mathrm{G}_c^N + \mathrm{G}_f^N}{\mathrm{G}_f^N - \mathrm{G}_c^N}\right) \frac{b_N^{2n}}{b_N^{2n}\frac{\mathrm{G}_c^N+\mathrm{G}_f^N}{\mathrm{G}_f^N-\mathrm{G}_c^N} + a_N^{2n}}, \\ K_4^n &= -K_1^n \frac{a_N^{2n} b_N^{2n}}{b_N^{2n}\frac{\mathrm{G}_c^N+\mathrm{G}_f^N}{\mathrm{G}_f^N-\mathrm{G}_c^N} + a_N^{2n}}. \end{aligned} \tag{5.31}$$

The stationary values are given by

$$\beta_n = (\mathrm{G}_c^N)^{-1} \left(\frac{\mathrm{G}_f^N(b_N^{2n} - a_N^{2n}) + \mathrm{G}_c^N(b_N^{2n} + a_N^{2n})}{\mathrm{G}_f^N(b_N^{2n} + a_N^{2n}) + \mathrm{G}_c^N(b_N^{2n} - a_N^{2n})}\right). \tag{5.32}$$

One readily checks for $\mathrm{G}_f^N \ge \mathrm{G}_c^N$ that β_n is increasing with n and that $\beta_1 = 1/G_{CCA}^N$. It can also be easily checked that (5.32) gives all of the stationary values. Indeed one supposes there exists a stationary value $\tilde{\beta}$ not given by (5.32) to find that the only associated stationary functions are of the form $r = \text{const}$, $h = -\text{const}$. Thus we find that $\hat{\beta} = \mathrm{G}_m\beta_1 = \mathrm{G}_m/G_{CCA}^N$ to conclude that

$$\text{if } \frac{\mathrm{G}_m}{G_{CCA}^N} \ge 1, \text{ then } D \le 0, \tag{5.33}$$

and Proposition 5.1 follows.

The torsional rigidity for an arbitrary simply connected cross section reinforced with N circular coated fibers is denoted by $\mathcal{T}^N(\Omega)$ and repeated application of Proposition 5.1 gives the following.

PROPOSITION 5.2 (upper bound). *If* $\mathrm{G}_c^i \le \mathrm{G}_f^i$, $i = 1, \ldots, N$, *and*

$$G_{CCA}^i \le \mathrm{G}_m \text{ for } i = 1 \ldots, N, \tag{5.34}$$

then

$$\mathcal{T}^N(\Omega) \le \mathrm{G}_m \mathcal{T}_0(\Omega) + \sum_{i=1}^{N} \left(T_f^i - \frac{\pi}{2}\mathrm{G}_m\, b_i^4\right), \tag{5.35}$$

where $\mathcal{T}_0(\Omega)$ *is the torsional rigidity of the homogeneous cross section containing material with unit shear modulus.*

Proposition 2.6 follows directly from Proposition 5.2. To establish Proposition 2.4 one recalls the isoperimetric inequality

$$\mathcal{T}_0(\Omega) \leq \frac{\pi}{2} R^4, \tag{5.36}$$

which holds for all cross sections Ω with area πR^2; see Polya [7]. Proposition 2.4 then follows from (5.36) and (5.35) when $\mathrm{G}_f^i = \mathrm{G}_f$ and $\mathrm{G}_c^i = \mathrm{G}_c$, and $a_i/b_i = \nu^{1/2}$ for $i = 1, \ldots, N$.

6. Lower bounds on the torsional rigidity for $\mathbf{G_f^i \leq G_c^i}$. We focus on the case where $\mathrm{G}_f^i \leq \mathrm{G}_c^i$, $i = 1, \ldots, N$. We proceed as in the last section and investigate the effects of adding a circular coated fiber to an already existing configuration of $N-1$ coated fibers. At present no assumptions on the geometry or shear moduli of the $N-1$ coated fibers are made. The part of the shaft cross section already occupied by the coated fibers is denoted by A and the cross section of the circular coated fiber to be added by Σ. Here Σ is composed of a circular fiber of radius a_N with shear modulus G_f^N surrounded by a coating of outer radius b_N with shear modulus G_c^N. The torsional rigidity of the original configuration is denoted by $\mathcal{T}(A, \Omega)$. The rigidity associated with the added fiber is written as $\mathcal{T}(A \cup \Sigma, \Omega)$.

PROPOSITION 6.1 (lower rigidity inequality). *If* $\mathrm{G}_f^N \leq \mathrm{G}_c^N$ *and if*

$$\mathrm{G}_m \leq G_{CCA}^N, \tag{6.1}$$

then

$$\mathcal{T}(A, \Omega) + T_f^N - \frac{\pi}{2} \mathrm{G}_m \, b_N^4 \leq \mathcal{T}(A \cup \Sigma, \Omega). \tag{6.2}$$

Proposition 6.1 is established with the aid of the variational principle given by (2.1). One writes (2.1) as $\mathcal{T}(A \cup \Sigma, \Omega) = -2\mathcal{E}(A \cup \Sigma, \Omega)$, where

$$\mathcal{E}(A \cup \Sigma, \Omega) = \min_{\varphi} \left\{ \frac{1}{2} \int_\Omega \mathcal{G}^{-1}(\mathbf{x}) |\nabla \varphi|^2 \, d\mathbf{x} \, - \, 2 \int_\Omega \varphi \, d\mathbf{x} \right\}, \tag{6.3}$$

where the piecewise constant shear modulus $\mathcal{G}(\mathbf{x})$ is G_m in the matrix and takes the values G_f^i and G_c^i in the ith fiber and coating, respectively, for $i = 1, \ldots, N$. The idea of the proof is to estimate the quantity $\mathcal{E}(A \cup \Sigma, \Omega)$ in terms of $\mathcal{E}(A, \Omega)$ associated with the original configuration of $N-1$ fibers. We let $G(\mathbf{x})$ denote the piecewise constant shear modulus for the original configuration A of $N-1$ fibers. Here $G(\mathbf{x}) = \mathcal{G}(\mathbf{x})$ outside of Σ and inside Σ the shear modulus $G(\mathbf{x}) = \mathrm{G}_m$. We regroup terms in the variational principle (6.3) and write

$$\begin{aligned} \mathcal{E}(A \cup \Sigma, \Omega) = \min_{\varphi} \Bigg\{ & \frac{1}{2} \int_\Omega G^{-1}(\mathbf{x}) |\nabla \varphi|^2 d\mathbf{x} - 2 \int_\Omega \varphi d\mathbf{x} \\ & + \frac{1}{2} \left(\int_\Sigma (\mathcal{G}^{-1}(\mathbf{x}) - \mathrm{G}_m^{-1}) \right) |\nabla \varphi|^2 d\mathbf{x} \Bigg\}. \end{aligned} \tag{6.4}$$

We obtain an estimate by substitution of a suitable trial field in (6.4). Our choice is made as follows: We introduce the stress potential Φ_A for the configuration A. Here Φ_A is continuous in Ω,

$$-G^{-1}(\mathbf{x}) \Delta \Phi_A = 2, \tag{6.5}$$

and satisfies the transmission conditions

(6.6)
$$\mathrm{G}_f^{i\,-1}\nabla\tilde{\Phi}_{|_f}\cdot\mathbf{n}=\mathrm{G}_c^{i\,-1}\nabla\tilde{\Phi}_{|_c}\cdot\mathbf{n} \text{ on the fiber–coating interface } |\mathbf{x}-\mathbf{x}_i|=a_i, i=1,\dots,N-1,$$

and

(6.7)
$$\mathrm{G}_f^{i\,-1}\nabla\tilde{\Phi}_{|_c}\cdot\mathbf{n}=\mathrm{G}_m{}^{-1}\nabla\tilde{\Phi}_{|_m}\cdot\mathbf{n} \text{ on the matrix–coating interface, } |\mathbf{x}-\mathbf{x}_i|=b_i,\dots,N-1.$$

The trial field φ is chosen to match Φ_A outside the coated fiber cross section Σ but inside we suppose that $\varphi=\Phi_A+\delta$, where δ is continuous, vanishes on the boundary of Σ, is square integrable inside Σ, and has square integrable gradient over Σ. One easily checks that

$$\int_\Sigma \mathrm{G}_m{}^{-1}\nabla\Phi_A\cdot\nabla\delta\,d\mathbf{x}=2\int_\Sigma\delta\,d\mathbf{x}. \tag{6.8}$$

Application of (6.8) and rearranging terms as in the previous section yield

$$\begin{aligned}\mathcal{T}(A\cup\Sigma,\Omega)\geq\mathcal{T}(A,\Omega)&-\int_\Sigma \mathrm{G}_m{}^{-1}|\nabla\delta|^2d\mathbf{x}\\ &+\int_\Sigma(\mathrm{G}_m^{-1}-\mathcal{G}^{-1}(\mathbf{x}))|\nabla\varphi|^2d\mathbf{x}.\end{aligned}\tag{6.9}$$

On maximizing the right-hand side of (6.9) with respect to δ, we obtain

$$\mathcal{T}(A\cup\Sigma,\Omega)\geq\mathcal{T}(A,\Omega)+\mathcal{U}, \tag{6.10}$$

where

$$\mathcal{U}=\int_\Sigma(\mathrm{G}_m^{-1}-\mathcal{G}^{-1}(\mathbf{x}))\nabla\hat{\varphi}\cdot\nabla\Phi_A\,d\mathbf{x}. \tag{6.11}$$

Here $\hat{\varphi}=\hat{\delta}+\Phi_A$ in Σ and $\hat{\delta}$ solves

$$\int_\Sigma\mathcal{G}^{-1}(\mathbf{x})\nabla\hat{\delta}\cdot\nabla u\,d\mathbf{x}=\int_\Sigma(\mathrm{G}_m^{-1}-\mathcal{G}^{-1}(\mathbf{x}))\nabla\Phi_A\cdot\nabla u\,d\mathbf{x} \tag{6.12}$$

for every trial u vanishing on the boundary of Σ. From (6.8) and (6.12) we see that $\hat{\varphi}$ solves

$$\int_\Sigma\mathcal{G}^{-1}(\mathbf{x})\nabla\hat{\varphi}\cdot\nabla u\,d\mathbf{x}=2\int_\Sigma u\,d\mathbf{x} \tag{6.13}$$

for every trial u vanishing on the boundary of Σ. Proceeding as in the last section we introduce the continuous functions ψ, r, ψ^h, and h such that $\hat{\varphi}=\psi-h$ and $\Phi_A=r+\psi^h$. Here ψ^h and ψ are the same functions introduced in section 5. The function ψ^h is the solution of (5.18) and ψ solves the transmission problem given by (5.19) and (5.20). The function r is harmonic in Σ and $r=\Phi_A$ on the boundary of Σ. The function h is continuous on Σ and $h=-\Phi_A$ on the boundary of Σ. It is the solution of the transmission problem given by (5.21) and (5.22).

Substitution of the functions ψ^h, ψ, r, and h into $\mathcal{U}$ gives

$$\mathcal{T}(A\cup\Sigma,\Omega)\geq\mathcal{T}(A,\Omega)+T_f^N-\frac{\pi}{2}\mathrm{G}_m\,b_N^4+D, \tag{6.14}$$

where the indefinite quantity D is given by (5.25). Its clear that (6.2) holds when $D \geq 0$. We find conditions for which $D \geq 0$ when $r \neq$ const and $h \neq$ const. To do this we search for the largest number ρ for which

$$\int_\Sigma |\nabla r|^2 \, d\mathbf{x} - \rho \int_\Sigma \frac{\mathrm{G}_m}{\mathcal{G}(\mathbf{x})} |\nabla h|^2 \, d\mathbf{x} \geq 0 \tag{6.15}$$

for every choice of r and h in $\mathcal{C}$ and $r \neq$ const and $h \neq -$ const. The largest ρ is given by

$$\hat{\rho} = \inf_{\mathcal{C}} \frac{\int_\Sigma |\nabla r|^2 \, d\mathbf{x}}{\int_\Sigma \frac{\mathrm{G}_m}{\mathcal{G}(\mathbf{x})} |\nabla h|^2 \, d\mathbf{x}}. \tag{6.16}$$

Proceeding as in the previous section, we find that $\hat{\rho} = G^N_{CCA}/\mathrm{G}_m$. Thus $D \geq 0$ when $G^N_{CCA}/\mathrm{G}_m \geq 1$ and the proposition follows.

The torsional rigidity for an arbitrary simply connected cross section reinforced with N circular coated fibers is denoted by $\mathcal{T}^N(\Omega)$ and repeated application of Proposition 6.1 gives the following.

PROPOSITION 6.2 (lower bound). *If* $\mathrm{G}^i_f \leq \mathrm{G}^i_c$, $i = 1, \ldots, N$, *and*

$$\mathrm{G}_m \leq G^i_{CCA} \text{ for } i = 1 \ldots, N, \tag{6.17}$$

then

$$\mathrm{G}_m \mathcal{T}_0(\Omega) + \sum_{i=1}^{N} \left(T^i_f - \frac{\pi}{2} \mathrm{G}_m \, b_i^4 \right) \leq \mathcal{T}^N(\Omega). \tag{6.18}$$

Proposition 2.7 follows directly from Proposition 6.2.

REFERENCES

[1] A. ALVINO AND G. TROMBETTI, *Isoperimetric inequalities connected with torsion problems and capacity*, Boll. Un. Mat. Ital. B (6), 4 (1985), pp. 773–787.

[2] T. CHEN, Y. BENVENISTE, AND P. C. CHUANG, *Exact solution in torsion of composite bars: Thickly coated neutral inhomogeneities and composite cylinder assemblages*, Proc. Roy. Soc. London A, 458 (2002), pp. 1719–1759.

[3] Z. HASHIN AND B. W. ROSEN, *The elastic moduli of fiber–reinforced materials*, J. Appl. Mech., 31 (1964), pp. 222–232.

[4] J. P. JONES AND J. S. WHITTIER, *Waves at a flexibly bonded interface*, J. Appl. Mech., 34 (1967), pp. 905–909.

[5] R. LIPTON, *Optimal fiber configurations for maximum torsional rigidity*, Arch. Ration. Mech. Anal., 144 (1998), pp. 79–106.

[6] G. W. MILTON, *The Theory of Composites*, Cambridge University Press, Cambridge, UK, 2002.

[7] G. POLYA, *Torsional rigidity, principle frequency, electrostatic capacity, and symmetrization*, Quart. Appl. Math., 6 (1948), pp. 267–277.

[8] G. POLYA AND A. WEINSTEIN, *On the torsional rigidity of multiply connected cross sections*, Ann. of Math. (2), 52 (1950), pp. 154–163.

[9] G. POLYA AND G. SZEGO, *Isoperimetric Inequalities in Mathematical Physics*, Princeton University Press, Princeton, NJ, 1951.

[10] B. DE SAINT-VENANT, *Mémoire sur la torsion des prismes*, Mém. pres. divers savants Acad. Sci., 14 (1856), pp. 233–560.

SIAM J. APPL. MATH.
Vol. 65, No. 1, pp. 316–335

DELAYED COUPLING BETWEEN TWO NEURAL NETWORK LOOPS*

SUE ANN CAMPBELL†, R. EDWARDS‡, AND P. VAN DEN DRIESSCHE§

Abstract. Coupled loops with time delays are common in physiological systems such as neural networks. We study a Hopfield-type network that consists of a pair of one-way loops each with three neurons and two-way coupling (of either excitatory or inhibitory type) between a single neuron of each loop. Time delays are introduced in the connections between loops, and the effects of coupling strengths and delays on the network dynamics are investigated. These effects depend strongly on whether the coupling is symmetric (of the same type in both directions) or asymmetric (inhibitory in one direction and excitatory in the other). The network of six delay differential equations is studied by linear stability analysis and bifurcation theory. Loops having inherently stable zero solutions cannot be destabilized by weak coupling, regardless of the delay. Asymmetric coupling is weakly stabilizing but easily upset by delays. Symmetric coupling (if not too weak) can destabilize an inherently stable zero solution, leading to nontrivial fixed points if the gain of the neuron response function is not too negative or to oscillation otherwise. In the oscillation case, intermediate delays can restabilize the zero solution. At the borderline of the weak coupling region (symmetric or asymmetric), stability can change with delay ranges. When the coupling strengths are of the same magnitude, the oscillations of corresponding neurons in the two loops can be in phase, antiphase (symmetric coupling), or one quarter period out of phase (asymmetric coupling) depending on the delay.

Key words. neural network, coupled loops, time delay, bifurcation, oscillation

AMS subject classifications. 92B20, 34K20, 34K18, 92C20

DOI. 10.1137/S0036139903434833

1. Introduction.

Interacting loops that are capable of sustaining oscillation are common in physiological systems. One approach to modeling such systems is via coupled oscillators [13]. However, this approach does not lend itself to studying the patterns of connections between oscillators when each oscillator is itself a network. Furthermore, such networks may not be inherently oscillatory, but oscillations may arise as a result of the coupling between them. If the coupling between networks is slower than each network's internal dynamics, then additional effects can arise from the delay in the coupling. The coupling may also be faster than the internal dynamics, in which case each network could be modeled with internal delays, or both the internal connections and coupling between networks could have delays.

These questions arise in models of the brain's motor circuitry, where there are many interacting loops and feedback systems. For example, functionally separate parallel loops operate through the basal ganglia (e.g., through matrisomes in the striatum [10]) but may interact through crosstalk [2]. These loop interactions have been implicated in the generation of tremor oscillations in Parkinson's disease. The effect of the particular patterns of connections between parallel copies of a network was studied by Edwards and Gill [5], where synchrony of the network copies occurred

*Received by the editors September 16, 2003; accepted for publication (in revised form) April 22, 2004; published electronically October 28, 2004. This work was supported by the Natural Sciences and Engineering Research Council (Canada).

http://www.siam.org/journals/siap/65-1/43483.html

†Department of Applied Mathematics, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

‡Centre for Nonlinear Dynamics in Physiology and Medicine, McGill University, Montréal, QC H3A 2T5, Canada (sacampbell@uwaterloo.ca).

§Department of Mathematics and Statistics, University of Victoria, Victoria, BC V8W 3P4, Canada (edwards@math.uvic.ca, pvdd@math.uvic.ca).

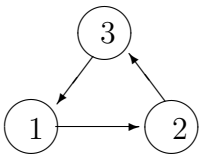
with appropriate crosstalk, but only when each network was in a periodic regime.

Previous work on these problems has not explicitly allowed for delays in connections. While analysis becomes more difficult with delays, their effects may be important for the applications. Research on Hopfield-type neural networks with delays, first introduced by Marcus and Westervelt [15], has shown that delays can modify dynamics in interesting ways. Delays have been inserted into various simple loop structures. Some work [14, 17, 18] has considered systems of two neurons with delayed connections. Shayer and Campbell [18] gave a detailed analysis of dynamics of two coupled units with delayed coupling and also delayed self-input, showing in particular how oscillation occurs when the interactions are strong enough, but also depending on the delays. A number of authors have studied loops of three or more neurons with delays, showing various types of behavior including oscillations, waves, steady states, and even chaos [1, 3, 7, 16, 21]. Other work has dealt with conditions for stability of steady states in Hopfield-type networks of arbitrary structure (see, for example, [4, 9, 18, 19, 20] and references therein).

The current study is an initial attempt to determine the effects of coupling (or crosstalk) between parallel copies of a network structure in the presence of delays. We focus on the simplest example that uses Hopfield network equations and in which each network copy is capable of oscillation, namely, a pair of simple loops of three neurons with one-way connections, with coupling between only one neuron of each loop. We consider the case in which the coupling between the loops, rather than the connections within the loops, is delayed. Of the previous studies mentioned above, this is perhaps most similar to that of Shayer and Campbell [18], in which the concern was also with delayed coupling between two potential oscillators, though the oscillators were single neurons with self-input rather than small loops. A single (one-way) loop of three Hopfield neurons can oscillate if their connections are inhibitory and sufficiently strong. The coupling between the loops can change this behavior, but we seek to determine how the behavior depends on the strength of coupling, the delay in coupling, and the internal gain or connection strengths within each loop.

We begin in section 2 with the dynamics of a single 3-loop (loop of three neurons with one-way connections). In section 3 we look at a pair of coupled 3-loops, giving results on stability of the trivial equilibrium and the presence of oscillation over the parameter space defined by the internal gain parameter (positive or negative) and a coupling strength parameter, allowing either inhibitory or excitatory coupling in either direction. Section 4 deals with delayed coupling between the loops. Most of the analysis is local, but where we do not have global results, numerical experiments support the conclusions. We summarize our results with a discussion (section 5).

2. Isolated 3-loop without delay. Consider a Hopfield-type network of three neurons connected in a (one-way) loop as in the following figure:



This can be described by the system of ordinary differential equations (with subscripts interpreted mod 3)

$$\frac{dx_j}{dt} = -x_j + \tanh(bx_{j-1}), \qquad j = 1, 2, 3, \tag{2.1}$$

together with initial condition $x(0) = (x_1(0), x_2(0), x_3(0))^t$. Here x_j represents the normalized voltage of neuron j and $b \in \mathbb{R}$ is the gain of the response function, assumed equal for each neuron. Interactions are inhibitory if $b < 0$ and excitatory if $b > 0$. System (2.1) always has the trivial equilibrium $(0, 0, 0)^t$, i.e., at the origin. The existence of nontrivial equilibria depends on the value of b, as in the following result.

THEOREM 2.1. *If $b > 1$, then system (2.1) has one positive symmetric and one negative symmetric equilibrium. If $b \leq 1$, then there is no nontrivial equilibrium.*

Proof. At an equilibrium of (2.1), $x_1 = \tanh(bx_3) \equiv f(x_3)$; thus $x_2 = \tanh(bx_1) = f(f(x_3))$ and $x_3 = \tanh(bx_2) = F(x_3)$, where $F(x) \equiv f(f(f(x)))$. Consider

$$h(x_3) \equiv x_3 - F(x_3) = 0. \tag{2.2}$$

Since $f'(0) = b$, $h'(0) = 1 - b^3$. Also $\lim_{x_3 \to \pm\infty} h(x_3) = \pm\infty$. From (2.2)

$$h'(x_3) = 1 - b^3 \operatorname{sech}^2(b\tanh(b\tanh(bx_3))) \operatorname{sech}^2(b\tanh(bx_3)) \operatorname{sech}^2(bx_3). \tag{2.3}$$

Thus $h'(x_3) \geq 1 - b^3$ for $b \geq 0$, which is nonnegative (for all x_3) if $b \leq 1$. Since $h(0) = 0$, there is no nontrivial equilibrium for $0 \leq b \leq 1$. For $b < 0$, (2.3) gives $h'(x_3) \geq 1$, which again shows that there is no nontrivial equilibrium.

If $b > 1$, then (2.3) gives $h''(x_3) > 0$ for all $x_3 > 0$, showing that $h(x_3)$ is concave up. This, together with $h(0) = 0$, $h'(0) < 0$, and $h(x_3) > 0$ for sufficiently large $x_3 > 0$, shows that there is a unique positive solution $x_3 = \overline{x_3} > 0$ to (2.2). The corresponding equilibrium values $\overline{x_1} > 0$, $\overline{x_2} > 0$ are determined from $\overline{x_3}$ and (2.1). Since (2.2) holds also for x_1 and x_2, it must be that $\overline{x_1} = \overline{x_2} = \overline{x_3} = \overline{x}$, giving the unique symmetric positive equilibrium as $(x_1, x_2, x_3)\,t = (\overline{x}, \overline{x}, \overline{x})^t$, with $0 < \overline{x} < 1$ from the equilibrium equation $\overline{x} = \tanh(b\overline{x})$. By symmetry, there is also a unique negative equilibrium $(x_1, x_2, x_3)\,t = (-\overline{x}, -\overline{x}, -\overline{x})^t$ if $b > 1$. ∎

The linear stability of an equilibrium $(\overline{x}, \overline{x}, \overline{x})\,t$ is governed by $\frac{dx}{dt} = Ax$, with

$$A = \begin{bmatrix} -1 & 0 & b\operatorname{sech}^2(b\overline{x}) \\ b\operatorname{sech}^2(b\overline{x}) & -1 & 0 \\ 0 & b\operatorname{sech}^2(b\overline{x}) & -1 \end{bmatrix}. \tag{2.4}$$

The following result shows that a Hopf bifurcation can occur at the trivial equilibrium.

THEOREM 2.2. *The trivial solution of (2.1) is locally asymptotically stable iff $-2 < b < 1$. At $b = -2$, the system undergoes a Hopf bifurcation and has stable limit cycle solutions for $b \lesssim -2$.*

Proof. The characteristic equation of A at $\overline{x} = 0$ in (2.4) is $-(1+\lambda)^3 + b^3 = 0$. For $-2 < b < 1$, all eigenvalues have negative real parts; thus the system is linearly stable. When $b = 1$, there is a zero eigenvalue, and for $b > 1$ there is a real positive eigenvalue. When $b = -2$, the eigenvalues are -3, $\pm\sqrt{3}i$, and for $b < -2$ there is a complex pair of eigenvalues with positive real part. At $b = -2$, matrix A is diagonalized by a matrix P of eigenvectors. Approximating $\tanh(bx_j)$ by $bx_j - b^3x_j^3/3$ (ignoring terms

of order ≥ 5) system (2.1) with $x = (x_1, x_2, x_3)^t$ is transformed by $y = P^{-1}x$ with $y = (y_1, y_2, y_3)^t$ at $b = -2$ to

$$\frac{dy}{dt} = \begin{bmatrix} 0 & -\sqrt{3} & 0 \\ \sqrt{3} & 0 & 0 \\ 0 & 0 & -3 \end{bmatrix} y + P^{-1}\left(\frac{8}{3}\right) \begin{bmatrix} (y_1 + y_3)^3 \\ \left(-y_1/2 - \sqrt{3}y_2/2 + y_3\right)^3 \\ \left(-y_1/2 + \sqrt{3}y_2/2 + y_3\right)^3 \end{bmatrix}. \tag{2.5}$$

The center manifold is given by $y_3 = H(y_1, y_2)$, with H third order because the third equation of (2.5) has no quadratic term. On the center manifold, (2.5) becomes

$$\begin{bmatrix} \frac{dy_1}{dt} \\ \frac{dy_2}{dt} \end{bmatrix} = \begin{bmatrix} 0 & -\sqrt{3} \\ \sqrt{3} & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} p(y_1, y_2) \\ q(y_1, y_2) \end{bmatrix}, \tag{2.6}$$

where p and q are determined by substituting $y_3 = H(y_1, y_2)$ in the first two equations of (2.5). The standard formula for the criticality coefficient ((3.4.11) of Guckenheimer and Holmes [11]) gives $a = -1 < 0$. Since $\mathrm{Re}\,(\partial\lambda/\partial b) = -1/2$ when evaluated at $\lambda = \pm\sqrt{3}i$, $b = -2$, the supercritical Hopf bifurcation gives rise to stable periodic solutions occurring for $b \lesssim -2$. □

For a Hopfield 2-loop, the corresponding characteristic equation, $-(1+\lambda)^2 + b^2 = 0$, cannot have pure imaginary solutions. Thus a Hopfield 3-loop without delay is the smallest that can undergo a Hopf bifurcation at the origin.

Global results for the trivial equilibrium when $-2 < b \leq 1$ are now stated. Theorem 2.1 of van den Driessche and Zou [20] can be used to show easily that if $|b| < 1$, then the origin is globally asymptotically stable. For system (2.1), a Lyapunov function $V = \sum_{j=1}^{3} x_j^2$ can be used to extend the range of global stability of the origin to $-\sqrt{2} \leq b \leq 1$. Note that $b = 1$ is included here, whereas it was not in Theorem 2.2. Numerical results indicate that the full range of global stability is $-2 < b \leq 1$.

Global results for the existence and stability of periodic solutions for $b < -2$ are more difficult to obtain. However, in the limit $b \to -\infty$, when the hyperbolic tangents become step functions, the problem is easier. Glass and Pasternack [8] showed that n-dimensional networks similar to (2.1) but with step functions have globally asymptotically stable periodic solutions for $n \geq 3$. Numerical simulations of (2.1) with $b < -2$ indicate that there is a unique globally asymptotically stable periodic solution for each $b \in (-\infty, -2)$.

Consider now the stability of the nontrivial equilibria (when they exist).

THEOREM 2.3. *For $b > 1$, the positive and negative symmetric equilibria of* (2.1) *are locally asymptotically stable.*

Proof. From (2.4), the characteristic equation of A is $-(1+\lambda)^3 + b^3\,\mathrm{sech}^6(b\overline{x}) = 0$, where $(\overline{x}, \overline{x}, \overline{x})^t$ with $\overline{x} > 0$ is the positive symmetric equilibrium of (2.1) that exists for $b > 1$ (by Theorem 2.1). Thus the eigenvalues are

$$\lambda_1(\overline{x}) = -1 + b\,\mathrm{sech}^2(b\overline{x}), \quad \lambda_{2,3}(\overline{x}) = -1 - \left(\frac{1}{2} \pm i\sqrt{\frac{3}{2}}\right) b\,\mathrm{sech}^2(b\overline{x}).$$

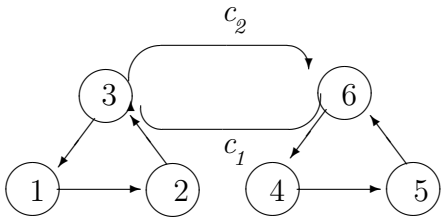
Since $b\,\mathrm{sech}^2(b\overline{x}) > 0$, local stability follows if $\lambda_1(\overline{x}) < 0$. By (2.1), $b\overline{x} = \tanh^{-1}(\overline{x})$ and $\mathrm{sech}^2(b\overline{x}) = 1 - \overline{x}^2$, giving $\lambda_1(\overline{x}) = -1 + \tanh^{-1}(\overline{x})\left(1 - \overline{x}^2\right)/\overline{x}$ for $0 < \overline{x} < 1$, which is equivalent to $1 < b < \infty$. Also, $\lambda_1(0) = 0$ at $b = 1$. Differentiating gives

$$\frac{d\lambda_1}{d\overline{x}}(\overline{x}) = \left(\overline{x} - \left(1 + \overline{x}^2\right)\tanh^{-1}(\overline{x})\right)/\overline{x}^2,$$

which is negative, since $\tanh^{-1}(\overline{x}) > \overline{x}$. Thus $\lambda_1(\overline{x}) < 0$, showing that the positive symmetric equilibrium is locally stable. Stability for $\overline{x} < 0$ follows by symmetry. ∎

Note that $\overline{x} \to 0^+$ as $b \to 1^+$, showing that the linearly stable positive and negative equilibria bifurcate from the trivial equilibrium as it loses stability. Thus the system has a supercritical pitchfork bifurcation at $b = 1$.

3. Coupled loops without delay. Consider a pair of coupled 3-loops:



The individual loops each follow the form of (2.1). Coupling strengths are given by c_1 and c_2, where $bc_j > 0$ implies excitatory and $bc_j < 0$ implies inhibitory coupling. The system of equations for the entire system is then

$$
\begin{aligned}
&\frac{dx_1}{dt} = -x_1 + \tanh(bx_3), \quad \frac{dx_2}{dt} = -x_2 + \tanh(bx_1),\\
&\frac{dx_3}{dt} = -x_3 + \tanh(bx_2) + c_1 \tanh(bx_6),\\
&\frac{dx_4}{dt} = -x_4 + \tanh(bx_6), \quad \frac{dx_5}{dt} = -x_5 + \tanh(bx_4),\\
&\frac{dx_6}{dt} = -x_6 + \tanh(bx_5) + c_2 \tanh(bx_3),
\end{aligned}
\tag{3.1}
$$

together with initial condition $x(0) = (x_1(0), x_2(0), x_3(0), x_4(0), x_5(0), x_6(0))^t$.

Equilibria of (3.1) satisfy $x_2 = f(x_1) = f(f(x_3))$ and $x_5 = f(x_4) = f(f(x_6))$, where $f(x_j) \equiv \tanh(bx_j)$. Using the other two equations gives

$$
x_3 = F(x_3) + c_1 f(x_6), \quad x_6 = F(x_6) + c_2 f(x_3), \tag{3.2}
$$

where $F(x) = f(f(f(x)))$ as before. This can be reduced (for $c_1 \neq 0$) to

$$
g(x_3) \equiv [x_3 - F(x_3)] - c_1 f\left(f\left(f\left(\frac{1}{c_1}[x_3 - F(x_3)]\right)\right) + c_2 f(x_3)\right) = 0. \tag{3.3}
$$

Any x_3 satisfying (3.3) determines x_6 and hence all the variables at an equilibrium. Clearly the origin $x_j = 0$, $j = 1, \ldots, 6$, is an equilibrium, and our interest mostly focuses on its stability properties. However, we first show the existence of nontrivial equilibria for some b values. Define $d \equiv b^2 c_1 c_2$ and $\beta \equiv b^3$. When $d > 0$, the coupling is either excitatory or inhibitory in both directions (symmetric coupling); when $d < 0$, the coupling is excitatory in one direction and inhibitory in the other (asymmetric coupling).

THEOREM 3.1. *If $d > (1-\beta)^2$, then system* (3.1) *has nontrivial equilibria. If either* (i) $0 < \beta < 1$ *and* $d < (1-\beta)^2$, *or* (ii) $\beta < 0$ *and* $d < 1$, *then system* (3.1) *has no nontrivial equilibrium.*

Proof. Assuming $c_1 \neq 0$, differentiating (3.3) gives

$$g'(x_3) = (1 - F'(x_3))\left(1 - \beta s_1^2 s_2^2 s_3^2\right) - d s_1^2 s_4^2,$$

where s_k^2, $k = 1, \ldots, 4$, represents $\operatorname{sech}^2(\cdot)$ evaluated at some point; thus $0 < s_k^2 \leq 1$. Since $F'(0) = \beta$, it follows from the derivative above that $g'(0) = (1-\beta)^2 - d$; thus $g(x_3)$ is strictly decreasing at the origin if $d > (1-\beta)^2$. Clearly $g(0) = 0$ and $\lim_{x\to\infty} g(x_3) = \infty$ because f (and therefore F) is bounded. Thus by continuity there is at least one positive value of x_3, namely, $\overline{x}_3 > 0$, such that $g(\overline{x}_3) = 0$. By symmetry, $g(-\overline{x}_3) = 0$ and these values determine the other variables at a nontrivial equilibrium.

For $x_3 > 0$ if $\beta > 0$, then $0 < F'(x_3) < \beta$. Thus $0 < \beta < 1$ implies that $(1-\beta)^2 - d < g'(x_3)$ if $d \geq 0$, and $(1-\beta)^2 < g'(x_3) < 1 - d$ if $d \leq 0$. Thus in case (i), $g(x_3)$ is strictly increasing for all $x_3 > 0$. Similarly $\beta < 0$ implies that $\beta < F'(x_3) < 0$ and $1 - d < g'(x_3) < (1-\beta)^2$ if $d \geq 0$, and $1 < g'(x_3) < (1-\beta)^2 - d$ if $d \leq 0$. Thus in case (ii), $g(x_3)$ is strictly increasing for all $x_3 > 0$. In both cases there is no nontrivial positive equilibrium and, by symmetry, no nontrivial negative equilibrium. If $c_1 = 0$ but $c_2 \neq 0$, then reversing the roles of x_3 and x_6 leads to the same conclusions. If $c_1 = c_2 = 0$, then the results follow from Theorem 2.1. □

Note that Theorem 3.1 does not specify all regions of parameter space in which nontrivial equilibria occur. It does not provide information about the regions where β is large and positive or where β is large and negative with $d > 1$. Moreover, the number and signs of equilibria may depend on the values of c_1 and c_2 for a given d. For example, if $b = 2$ (so that $\beta = 8$) and $c_1 = c_2 = 0$, then there is one positive and one negative nontrivial equilibrium for each uncoupled loop (see Theorem 2.1) so that for the full system (3.1) there are nine equilibria, three of which have x_3 positive. However, if the coupling goes only one way, e.g., $c_1 > 0$ but $c_2 = 0$, then there can be two or four different positive equilibrium values for x_3 when $b > 1$.

The special case of symmetric coupling $c_1 = c_2$ is now considered.

THEOREM 3.2. *Let $b > 1$. If $c_1 = c_2 > 0$, then system (3.1) has a positive equilibrium $x^* = (\overline{x}_1, \overline{x}_2, \overline{x}_3, \overline{x}_1, \overline{x}_2, \overline{x}_3)^t$ and an equilibrium $-x^*$; if $c_1 = c_2 < 0$, then it has equilibria $\tilde{x}^* = (\overline{x}_1, \overline{x}_2, \overline{x}_3, -\overline{x}_1, -\overline{x}_2, -\overline{x}_3)^t$ with $\overline{x}_j > 0$, $j = 1, 2, 3$, and $-\tilde{x}^*$.*

Proof. Consider $c_1 = c_2 > 0$ and suppose that $x_3 = x_6$; then (3.2) reduces to

$$G(x_3) \equiv x_3 - F(x_3) - c_1 f(x_3) = 0$$

at an equilibrium. Note that $G(0) = 0$, $\lim_{x_3\to\infty} G(x_3) = \infty$ and $G'(0) = 1 - b^3 - c_1 b < 0$. Thus there is at least one positive zero $\overline{x}_3$ of $G(x_3)$. Since $\overline{x}_3 = \overline{x}_6 > 0$, it follows that $\overline{x}_1 = \overline{x}_4 > 0$ and $\overline{x}_2 = \overline{x}_5 > 0$. By symmetry, the negative equilibrium follows.

Consider $c_1 = c_2 < 0$ and suppose that $x_3 = -x_6$; then (3.2) reduces to

$$\tilde{G}(x_3) \equiv x_3 - F(x_3) + c_1 f(x_3) = 0$$

at an equilibrium. By the above there is at least one positive zero $\overline{x}_3$ of $\tilde{G}(x_3)$. Then $\overline{x}_1, \overline{x}_2 > 0$ and $\overline{x}_4, \overline{x}_5, \overline{x}_6 < 0$. Symmetry gives the second equilibrium. □

Numerical solutions demonstrate that additional equilibria can occur: for $c_1 = c_2 > 0$ (respectively, < 0) there may be one or three equilibria with $\overline{x}_3 > 0$, $\overline{x}_6 < 0$ (respectively, $\overline{x}_3 < 0$, $\overline{x}_6 > 0$) and an equal number of symmetric equilibria.

The linear stability of the trivial equilibrium $x_j = 0$, $j = 1, \ldots, 6$, can be determined from $\frac{dx}{dt} = Ax$ with $x = (x_1, \ldots, x_6)^t$ and

$$A = \begin{bmatrix} -1 & 0 & b & 0 & 0 & 0 \\ b & -1 & 0 & 0 & 0 & 0 \\ 0 & b & -1 & 0 & 0 & bc_1 \\ 0 & 0 & 0 & -1 & 0 & b \\ 0 & 0 & 0 & b & -1 & 0 \\ 0 & 0 & bc_2 & 0 & b & -1 \end{bmatrix}. \tag{3.4}$$

The characteristic equation for this system at $x_j = 0$ with $d \equiv b^2 c_1 c_2$ and $\beta \equiv b^3$ is

$$\left[(1+\lambda)^3 - \beta\right]^2 - d\,(1+\lambda)^4 = 0. \tag{3.5}$$

First consider the case $d \geq 0$. The characteristic equation then factors as follows:

$$\Delta_+^+(\lambda)\Delta_-^+(\lambda) \equiv \left[(1+\lambda)^3 - \beta + \sqrt{d}\,(1+\lambda)^2\right]\left[(1+\lambda)^3 - \beta - \sqrt{d}\,(1+\lambda)^2\right] = 0. \tag{3.6}$$

From the single-loop results (see Theorem 2.2), if $d = 0$, then the origin is locally asymptotically stable for $-8 < \beta < 1$ and unstable for $\beta < -8$ and $\beta > 1$. To find the stability region for $\sqrt{d} > 0$, look for curves in the βd-plane on which there is a zero or pure imaginary eigenvalue.

From (3.6), zero eigenvalues occur when $1 - \beta = \pm\sqrt{d}$. Pure imaginary eigenvalues $\lambda = i\omega$ with $\omega > 0$ make $\Delta_+^+(\lambda) = 0$ when their real and imaginary parts are zero, namely,

$$1 - 3\omega^2 - \beta + \sqrt{d}\left(1 - \omega^2\right) = 0 \quad \text{and} \quad \omega\left(3 - \omega^2 + 2\sqrt{d}\right) = 0.$$

The second condition above gives $\omega^2 = 3 + 2\sqrt{d}$, which can be substituted into the first condition to yield $\beta = -2(2 + \sqrt{d})^2$. To make $\Delta_-^+(\lambda) = 0$, there is an analogous condition where $\sqrt{d}$ is replaced by $(-\sqrt{d})$, as long as $\omega^2 = 3 - 2\sqrt{d} > 0$, i.e., $\sqrt{d} < \frac{3}{2}$, namely, $\beta = -2(2 - \sqrt{d})^2$. Note that this curve intersects the parabola of zero eigenvalues at $\beta = -\frac{1}{2}$, $d = \frac{9}{4}$. These curves are shown in Figure 3.1. It is clear by continuity from the single-loop results ($d = 0$) that the region labeled "STABLE" in the figure corresponds to linear stability of the origin. By picking points in the other regions, it can easily be checked that the origin is unstable there.

In the case where $d \leq 0$ in (3.5), factor the characteristic equation as

$$\Delta_+^-(\lambda)\Delta_-^-(\lambda) \equiv \left[(1+\lambda)^3 - \beta + i\sqrt{-d}\,(1+\lambda)^2\right]\left[(1+\lambda)^3 - \beta - i\sqrt{-d}\,(1+\lambda)^2\right] = 0. \tag{3.7}$$

Now, $\lambda = 0$ when $1 - \beta \pm i\sqrt{-d} = 0$, i.e., only at the point $\beta = 1$ and $d = 0$. Working with $\Delta_+^-(\lambda) = 0$, $\lambda = i\omega$ implies that

$$\beta = 1 - 2\sqrt{-d}\omega - 3\omega^2 \quad \text{and} \quad \sqrt{-d} = \frac{\omega\left(\omega^2 - 3\right)}{\left(1 - \omega^2\right)} \quad (\text{if } \omega^2 \neq 1), \tag{3.8}$$

giving $\beta = 1 - 3\omega^2 - 2\omega^2\left(\omega^2 - 3\right)/\left(1 - \omega^2\right)$. Working with $\Delta_-^-(\lambda) = 0$ gives this same equation in β and ω.

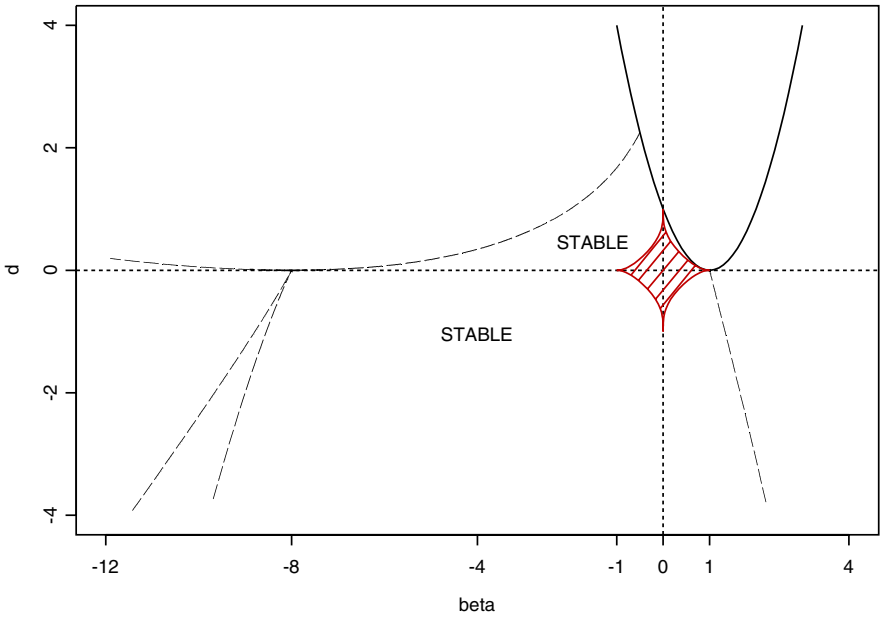


FIG. 3.1. *Regions in the parameter plane indicating stability of the origin in the coupled loop system (see* (3.1)*). The solid parabolic curve indicates parameter values where there is a zero eigenvalue. Dashed curves indicate parameter values where there are pure imaginary eigenvalues. The linearly stable regions defined by these boundaries are marked. The shaded area indicates where global stability of the origin has been proved.*

The fact that $\sqrt{-d} \geq 0$ implies that ω is in one of the intervals $\left(-\infty, -\sqrt{3}\right]$, $(-1, 0]$, or $\left(1, \sqrt{3}\right]$. In these intervals, (3.8) can be considered parametric equations for curves in the βd-plane. The curves are shown in Figure 3.1. Linear stability can again be checked by examining eigenvalues at points within each region. In the regions where $d > 0$ and $\beta > 1 - \sqrt{d}$, there are real positive eigenvalues, so the origin is an unstable node. Everywhere else outside the stability region it can be verified that there are complex conjugate pairs of eigenvalues with positive real parts, suggesting the existence of stable oscillations. Note that the pair of positive real eigenvalues becomes a complex pair with positive real part as d becomes negative $(\beta > 1)$.

For equal coupling strengths, as in the case of the single loop, global stability of the origin can be shown on a subset of the linear stability region.

THEOREM 3.3. *The origin is globally asymptotically stable for system* (3.1) *with* $|c_1| = |c_2| \equiv c > 0$ *when* $|b| < \frac{2}{\gamma}$, *where* $\gamma = 1 + 2c\,(B_1 + 1)$ *and* B_1 *is the positive root of the cubic* $8c^3 B_1\,(1 + B_1)^2 = 1$.

The proof uses the Lyapunov function $V = \sum_{j=1}^{6} a_j\, x_j^2$, where $a_j > 0$ are given as $a_1 = a_4 = \frac{B_1 + B_2}{2}$, $a_2 = a_5 = \frac{1}{4c} + \frac{B_2}{2}$, $a_3 = a_6 = \frac{1}{4c} + \frac{B_1 + 1}{2}$, and $B_2 = \sqrt{B_1/2c}$. Details of the proof are omitted. The condition $|b| < \frac{2}{\gamma}$ can be interpreted in terms of β and d by taking $d = \pm b^2 c_1^2$ (the sign depending on the sign of $c_1 c_2$, with $|c_1| = |c_2|$), and the resulting global stability region is the diamond-shaped region in Figure 3.1. Note that it covers most of the local stability region in the positive quadrant.

4. Coupled loops with delay. The case with delayed coupling connections between the 3-loops leads to the following system of delay differential equations:

$$
\begin{aligned}
&\frac{dx_1}{dt} = -x_1(t) + \tanh\left(bx_3(t)\right), \quad \frac{dx_2}{dt} = -x_2(t) + \tanh\left(bx_1(t)\right), \\
&\frac{dx_3}{dt} = -x_3(t) + \tanh\left(bx_2(t)\right) + c_1 \tanh\left(bx_6(t-\tau)\right), \\
&\frac{dx_4}{dt} = -x_4(t) + \tanh\left(bx_6(t)\right), \quad \frac{dx_5}{dt} = -x_5(t) + \tanh\left(bx_4(t)\right), \\
&\frac{dx_6}{dt} = -x_6(t) + \tanh\left(bx_5(t)\right) + c_2 \tanh\left(bx_3(t-\tau)\right),
\end{aligned} \tag{4.1}
$$

where $\tau \geq 0$ is the time delay, and when $\tau = 0$ this reduces to (3.1). To pose an initial value problem at $t = 0$, we must specify data for each variable on the interval $[-\tau, 0]$, i.e., $x_j(t) = \phi_j(t),\ -\tau \leq t \leq 0,\ j = 1, \ldots, 6$.

The equilibria for (4.1) are the same as for (3.1); in particular, Theorem 3.1 is also valid for (4.1). Using Theorem 2.1 of van den Driessche and Zou [20], we give one global stability result for system (4.1): If $|b| \max_i\{1 + |c_i|\} < 1$, then the origin is globally asymptotically stable for all values of delay $\tau \geq 0$. However, for other parameter ranges the stability of the equilibria may change due to the delay. In the next subsection we focus on the linear stability analysis of the trivial equilibrium. This then leads us to a discussion of the bifurcations of the trivial equilibrium.

4.1. Stability regions. Linearization of (4.1) about the origin gives

$$x'(t) = A_1 x(t) + A_2 x(t-\tau), \tag{4.2}$$

where

$$
A_1 = \begin{bmatrix}
-1 & 0 & b & 0 & 0 & 0 \\
b & -1 & 0 & 0 & 0 & 0 \\
0 & b & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 0 & b \\
0 & 0 & 0 & b & -1 & 0 \\
0 & 0 & 0 & 0 & b & -1
\end{bmatrix}, \quad
A_2 = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & bc_1 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & bc_2 & 0 & 0 & 0
\end{bmatrix}. \tag{4.3}
$$

The characteristic equation for this system with $d \equiv b^2 c_1 c_2$ and $\beta \equiv b^3$ is

$$\left[(1+\lambda)^3 - \beta\right]^2 - d(1+\lambda)^4 e^{-2\tau\lambda} = 0. \tag{4.4}$$

Section 3 describes the stability region of the trivial equilibrium when $\tau = 0$ (see Figure 3.1). To determine the stability region for $\tau > 0$, we determine curves in the $d\tau$-plane along which (4.4) has a zero root or a pair of pure imaginary roots. Given values of β and d for which the trivial equilibrium is stable at $\tau = 0$, it remains so for $0 \leq \tau \leq \tau_{crit}$, where τ_{crit} is the lowest value of τ on one of these curves.

First consider the case $d \geq 0$. The characteristic equation then factors as

$$
\begin{aligned}
&\Delta_+^+(\lambda)\Delta_-^+(\lambda) \\
&\quad \equiv \left[(1+\lambda)^3 - \beta + (1+\lambda)^2\sqrt{d}e^{-\tau\lambda}\right]\left[(1+\lambda)^3 - \beta - (1+\lambda)^2\sqrt{d}e^{-\tau\lambda}\right] = 0.
\end{aligned} \tag{4.5}
$$

As for the nondelayed case, zero roots occur when $d = (1-\beta)^2 \equiv d_0$.

LEMMA 4.1. *Let β and τ be fixed.*

(i) *If $\beta < 1$ and $\tau \neq 2 + \frac{3}{\beta-1}$, then $\Delta_-^+(\lambda)$ has a simple zero root when $d = d_0$; the number of roots of* (4.4) *with positive real part increases (decreases) by one as d increases through d_0 with $\tau > 2 + \frac{3}{\beta-1}$ ($\tau < 2 + \frac{3}{\beta-1}$).*

(ii) *If $\beta > 1$ and $\tau \neq 2 + \frac{3}{\beta-1}$, then $\Delta_+^+(\lambda)$ has a simple zero root when $d = d_0$; the number of roots of* (4.4) *with positive real part increases (decreases) by one as d increases through d_0 with $\tau > 2 + \frac{3}{\beta-1}$ ($\tau < 2 + \frac{3}{\beta-1}$).*

(iii) *If $\beta = 1$, both factors of* (4.5) *have a simple zero root when $d = d_0 = 0$.*

(iv) *If $\beta < 1$ ($\beta > 1$) and $\tau = 2 + \frac{3}{\beta-1}$, then $\Delta_-^+(\lambda)$ ($\Delta_+^+(\lambda)$) has a double zero root when $d = d_0$.*

Proof. The presence of zero roots follows from the facts that $\Delta_+^+(0) = 0$ when $\sqrt{d} = \beta - 1$ and $\Delta_-^+(0) = 0$ when $\sqrt{d} = 1 - \beta$. For case (iv), note that

$$\frac{d}{d\lambda}\Delta_\pm^+(\lambda) = 3(1+\lambda)^2 \pm 2(1+\lambda)\sqrt{d}e^{-\tau\lambda} \mp \tau(1+\lambda)^2\sqrt{d}e^{-\tau\lambda}.$$

Thus, if $\sqrt{d} = \pm(\beta - 1)$ and $\tau = 2 + \frac{3}{\beta-1}$, then $\frac{d}{d\lambda}\Delta_\pm^+(0) = 0$. The fact that zero is a simple root in cases (i)–(iii) also follows from this derivative.

To study the rate of change of the real part of a root, λ, of (4.5), consider either factor of this equation. For $d > 0$, differentiating with respect to d, keeping in mind that λ is a function of d, and rearranging give

$$\frac{d\lambda}{d\,d} = \frac{\pm(1+\lambda)^2\sqrt{d}e^{-\tau\lambda}}{-6d(1+\lambda)^2 \mp 4d(1+\lambda)\sqrt{d}e^{-\tau\lambda} \pm 2d\tau(1+\lambda)^2\sqrt{d}e^{-\tau\lambda}},$$

where the upper sign in $\pm$, $\mp$ refers to Δ_+^+ and the lower sign to Δ_-^+. Using (4.5) to eliminate $\pm\sqrt{d}e^{-\tau\lambda}$ and setting $\lambda = 0$ and $d = d_0$ yield

$$\left.\frac{d\lambda}{d\,d}\right|_{\lambda=0} = \frac{1}{2(\beta-1)[(\tau-2)(\beta-1)-3]}.$$

Consideration of the sign of the right-hand side completes the proofs of (i) and (ii). ∎

For $d > 0$, to find the curves where pure imaginary roots exist, set $\lambda = i\omega$ in each factor of (4.5) and separate into real and imaginary parts. Without loss of generality, take $\omega > 0$. For $\Delta_+^+(\lambda)$, isolating $\sin(\omega\tau)$ and $\cos(\omega\tau)$ yields

$$\begin{aligned} (1+\omega^2)^2\sqrt{d}\cos(\omega\tau) &= -\left((1+\omega^2)^2 - \beta(1-\omega^2)\right) \equiv -\mathcal{C}(\omega), \\ (1+\omega^2)^2\sqrt{d}\sin(\omega\tau) &= \omega\left((1+\omega^2)^2 + 2\beta\right) \equiv \mathcal{S}(\omega). \end{aligned} \tag{4.6}$$

To find d and τ in terms of β and ω, square the equations in (4.6) and add to give

$$d = d_{im}(\omega) \equiv \frac{(1+\omega^2)^3 + 2\beta(3\omega^2-1) + \beta^2}{(1+\omega^2)^2}. \tag{4.7}$$

Dividing the second equation of (4.6) by the first gives $\tan(\omega\tau) = -\mathcal{S}(\omega)/\mathcal{C}(\omega)$. However, this loses information about the signs of $\cos(\omega\tau)$ and $\sin(\omega\tau)$ that is in (4.6). Thus we introduce $y = \mathrm{Arctan}(u)$ as the branch of the arctangent function with range $(-\frac{\pi}{2}, \frac{\pi}{2})$. Note that this corresponds to $\cos(y) > 0$ and that the function

$\text{Arctan}(u) + \pi$ corresponds to $\cos(y) < 0$. The other branches of the arctangent function are obtained from these two by adding multiples of 2π. As can be seen from (4.6), the sign of $\cos(\omega\tau)$ is determined by $\mathcal{C}(\omega)$, and thus we define

$$\tau = \tau^+_{k+}(\omega) \equiv \frac{1}{\omega}\begin{cases} \text{Arctan}\left(-\frac{\mathcal{S}(\omega)}{\mathcal{C}(\omega)}\right) + 2k\pi, & \mathcal{C}(\omega) < 0, \\ \text{Arctan}\left(-\frac{\mathcal{S}(\omega)}{\mathcal{C}(\omega)}\right) + (2k+1)\pi, & \mathcal{C}(\omega) > 0, \end{cases} \tag{4.8}$$

where $k = 0, 1, \ldots$. (We do not take $k < 0$ as these branches always yield $\tau < 0$.)

In a similar manner it can be shown that the curves along which the second factor, $\Delta^+_-(\lambda)$, of (4.5) has a pair of pure imaginary roots are given by $(d, \tau) = (d_{im}(\omega), \tau^+_{k-}(\omega))$, where d_{im} is as above and

$$\tau^+_{k-}(\omega) \equiv \frac{1}{\omega}\begin{cases} \text{Arctan}\left(-\frac{\mathcal{S}(\omega)}{\mathcal{C}(\omega)}\right) + 2k\pi, & \mathcal{C}(\omega) > 0, \\ \text{Arctan}\left(-\frac{\mathcal{S}(\omega)}{\mathcal{C}(\omega)}\right) + (2k+1)\pi, & \mathcal{C}(\omega) < 0. \end{cases} \tag{4.9}$$

The zeros of $\mathcal{C}(\omega)$ define the points where the branches join. To see how the sign of $\mathcal{C}(\omega)$ varies with β and ω, rewrite the first equation of (4.6) as a quartic in ω, namely, $\mathcal{C}(\omega) = \omega^4 + (2+\beta)\omega^2 + 1 - \beta$. The roots of this quartic are $\pm\omega^+_{\mathcal{C}}, \pm\omega^-_{\mathcal{C}}$, where

$$\omega^\pm_{\mathcal{C}} = \sqrt{-1 - \frac{\beta}{2} \pm \frac{1}{2}\sqrt{\beta(\beta+8)}}. \tag{4.10}$$

All four roots exist if $\beta \le -8$ (with $\omega^+_{\mathcal{C}} = \omega^-_{\mathcal{C}}$ when $\beta = -8$), no roots exist if $-8 < \beta < 1$, and only $\pm\omega^+_{\mathcal{C}}$ exists if $\beta \ge 1$ ($\omega^+_{\mathcal{C}} = 0$ when $\beta = 1$). This yields the following ranges:

$$\begin{aligned} \beta < -8 \quad &: \quad \mathcal{C}(\omega) > 0 \quad \text{for} \quad 0 < \omega < \omega^-_{\mathcal{C}},\ \omega^+_{\mathcal{C}} < \omega, \\ &\quad\ \ \mathcal{C}(\omega) < 0 \quad \text{for} \quad \omega^-_{\mathcal{C}} < \omega < \omega^+_{\mathcal{C}}, \\ -8 \le \beta \le 1 \quad &: \quad \mathcal{C}(\omega) > 0 \quad \text{for} \quad 0 < \omega, \omega \ne \omega^\pm_{\mathcal{C}}, \\ 1 < \beta \quad &: \quad \mathcal{C}(\omega) < 0 \quad \text{for} \quad 0 < \omega < \omega^+_{\mathcal{C}}, \\ &\quad\ \ \mathcal{C}(\omega) > 0 \quad \text{for} \quad \omega^+_{\mathcal{C}} < \omega. \end{aligned}$$

Now consider the case $d < 0$. The characteristic equation factors as

$$\begin{aligned} &\Delta^-_+(\lambda)\Delta^-_-(\lambda) \\ &\quad \equiv \left[(1+\lambda)^3 - \beta + i(1+\lambda)^2\sqrt{-d}e^{-\tau\lambda}\right]\left[(1+\lambda)^3 - \beta - i(1+\lambda)^2\sqrt{-d}e^{-\tau\lambda}\right] = 0. \end{aligned} \tag{4.11}$$

Clearly, neither factor has a zero root. Note that λ is a root of $\Delta^-_+(\lambda)$ iff $\bar{\lambda}$ is a root of $\Delta^-_-(\lambda)$. This is a consequence of the fact that the roots of the unfactored characteristic equation (4.4) come in complex conjugate pairs. Following a similar procedure to that for $d > 0$, pure imaginary roots $i\omega, -i\omega$ with $\omega > 0$, of the first and second factors, respectively, exist along the curves $(d, \tau) = (-d_{im}(\omega), \tau^-_{k+}(\omega))$. Similarly, pure imaginary roots $-i\omega, i\omega$ with $\omega > 0$, of the first and second factors,

respectively, exist along the curves $(d,\tau)=(-d_{im}(\omega),\tau^-_{k\,-}(\omega))$. Here

$$\tau^-_{k\,+}(\omega)\equiv\frac{1}{\omega}\begin{cases}\operatorname{Arctan}\left(\frac{\mathcal{C}(\omega)}{\mathcal{S}(\omega)}\right)+2k\pi, & \mathcal{S}(\omega)<0,\\ \operatorname{Arctan}\left(\frac{\mathcal{C}(\omega)}{\mathcal{S}(\omega)}\right)+(2k+1)\pi, & \mathcal{S}(\omega)>0,\end{cases}\tag{4.12}$$

$$\tau^-_{k\,-}(\omega)\equiv\frac{1}{\omega}\begin{cases}\operatorname{Arctan}\left(\frac{\mathcal{C}(\omega)}{\mathcal{S}(\omega)}\right)+2k\pi, & \mathcal{S}(\omega)>0,\\ \operatorname{Arctan}\left(\frac{\mathcal{C}(\omega)}{\mathcal{S}(\omega)}\right)+(2k+1)\pi, & \mathcal{S}(\omega)<0.\end{cases}\tag{4.13}$$

The zeros of $\mathcal{S}(\omega)$ define the points where the branches join. To make the definitions of $\tau^-_{k\pm}$ more precise, the sign of $\mathcal{S}(\omega)$ with $\omega_{\mathcal{S}}=\sqrt{\sqrt{-2\beta}-1}$ is given as follows:

$$\begin{array}{rcll}\beta<-\frac{1}{2} & : & \mathcal{S}(\omega)<0 & \text{for}\quad 0<\omega<\omega_{\mathcal{S}},\\ & & \mathcal{S}(\omega)>0 & \text{for}\quad \omega_{\mathcal{S}}<\omega,\\ \beta\geq-\frac{1}{2} & : & \mathcal{S}(\omega)>0 & \text{for}\quad 0<\omega.\end{array}$$

To determine what these curves look like, we use the following results that are derived by using L'Hôpital's rule. Note that we consider only $\tau\geq 0$.

LEMMA 4.2. *For the functions in* (4.7)–(4.9), (4.12), (4.13),

$$d_{im}(0)=d_0,\ \lim_{\omega\to\infty}d_{im}(\omega)=\infty;\ \lim_{\omega\to\infty}\tau^{\pm}_{k\,\pm}=0;$$

$$\lim_{\omega\to0^+}\tau^+_{k\,\pm}=\infty,\ k>0;\quad \lim_{\omega\to0^+}\tau^+_{0\,+}=\begin{cases}2+\frac{3}{\beta-1}, & \beta>1,\\ \infty, & \beta\leq1;\end{cases}$$

$$\lim_{\omega\to0^+}\tau^+_{0\,-}=\begin{cases}\infty, & \beta>1,\\ -\infty, & \beta=1,\\ 2+\frac{3}{\beta-1}, & \beta<1;\end{cases}$$

$$\lim_{\omega\to0^+}\tau^-_{k\,\pm}=\infty,\ k>0;\quad \lim_{\omega\to0^+}\tau^-_{0\,+}=\begin{cases}-\infty, & \beta<-\frac{1}{2},\\ \infty, & \beta\geq-\frac{1}{2};\end{cases}$$

$$\lim_{\omega\to0^+}\tau^-_{0\,-}=\begin{cases}\infty, & \beta<1,\\ 1, & \beta=1,\\ -\infty, & \beta>1.\end{cases}$$

LEMMA 4.3. *For* $\frac{1}{2}(5-3\sqrt{3})\leq\beta\leq12-4\sqrt{5}$, *i.e.,* β *approximately* $\in[-0.0981, 3.0557]$, $d_{im}(\omega)$ *is a nondecreasing function of* ω. *Outside this interval it is nonmonotone and has the following behavior. For* $\beta<\frac{1}{2}(5-3\sqrt{3})$ *or* $\beta\geq\frac{1}{2}(5+3\sqrt{3})\approx5.0981$, *there exists* $\omega_c>0$ *such that* $d_{im}(\omega)$ *is decreasing for* $0<\omega<\omega_c$ *and increasing for* $\omega>\omega_c$. *For* $12-4\sqrt{5}<\beta<\frac{1}{2}(5+3\sqrt{3})$, *there exist* $0<\omega_{c1}<\omega_{c2}$ *such that* $d_{im}(\omega)$ *is increasing for* $0<\omega<\omega_{c1}$ *and* $\omega>\omega_{c2}$ *and decreasing for* $\omega_{c1}<\omega<\omega_{c2}$.

Proof. From (4.7) it is clear that

$$\frac{d\,d_{im}}{d\omega}=2\omega\frac{\omega^6+3\omega^4+3(1-2\beta)\omega^2+2\beta(5-\beta)+1}{(1+\omega^2)^3};\tag{4.14}$$

thus the sign of $\frac{d\,d_{im}}{d\omega}$ is determined by $\Omega^3+3\Omega^2+3(1-2\beta)\Omega+2\beta(5-\beta)+1$, where $\Omega=\omega^2$. Consideration of the sign of the constant term shows that the cubic has an

even number of positive roots if $\beta \in [\frac{1}{2}(5-3\sqrt{3}), \frac{1}{2}(5+3\sqrt{3}))$ and an odd number otherwise. The discriminant of this cubic is a positive multiple of $-\beta^2(\beta^2-24\beta+64)$, which is nonnegative for $\beta \in [12-4\sqrt{5}, 12+4\sqrt{5}]$ and negative otherwise. Thus outside this interval the cubic has one real root, and inside it has three. For $\beta \in [\frac{1}{2}(5-3\sqrt{3}), 12-4\sqrt{5})$, the cubic has no positive roots and at $\beta = 12-4\sqrt{5}$ it has a double positive root. Consideration of the graph of the cubic in Ω shows that $\frac{d\,d_{im}}{d\omega} \geq 0$ for $\beta \in [\frac{1}{2}(5-3\sqrt{3}), 12-4\sqrt{5}]$, and hence $d_{im}(\omega)$ is a nondecreasing function of ω. For $\beta < \frac{1}{2}(5-3\sqrt{3})$ or $\beta \geq \frac{1}{2}(5+3\sqrt{3})$ the cubic has one positive root, Ω_c. Let $\omega_c = \sqrt{\Omega_c}$. For $12-4\sqrt{5} < \beta < \frac{1}{2}(5+3\sqrt{3})$, the cubic has two positive roots $\Omega_{c1} < \Omega_{c2}$. Let $\omega_{cj} = \sqrt{\Omega_{cj}}$. The results follow from the graph of the cubic. ∎

LEMMA 4.4. *For fixed β, τ, the number of roots of* (4.4) *with positive real part increases (decreases) by two as τ increases through one of the curves $(d,\tau) = (d_{im}, \tau^+_{k\pm})$, where d_{im} is an increasing (decreasing) function of ω. The number of roots of* (4.4) *with positive real part increases (decreases) by two as τ increases through one of the curves $(d,\tau) = (-d_{im}, \tau^-_{k\pm})$, where $-d_{im}$ is a decreasing (increasing) function of ω.*

Proof. Consider the first factor of (4.5). Differentiating with respect to τ gives

$$\frac{d\lambda}{d\tau} = \frac{\lambda(1+\lambda)\sqrt{d}e^{-\tau\lambda}}{3(1+\lambda)+\sqrt{d}e^{-\tau\lambda}(2-\tau(1+\lambda))}.$$

Using (4.5) to eliminate $e^{-\tau\lambda}$ and setting $\lambda = i\omega$ yield

$$\left.\frac{d\lambda}{d\tau}\right|_{\lambda=i\omega} = \frac{i\omega\left[\beta+4\omega^2-(1-\omega^2)^2+i\omega(\beta-4(1-\omega^2))\right]}{1-3\omega^2+2\beta-\tau(\beta+4\omega^2-(1-\omega^2)^2)+i\omega[3-\omega^2-\tau(\beta-4(1-\omega^2))]}.$$

Taking the real part gives

$$\left.\frac{d\,[\mathrm{Re}(\lambda)]}{d\tau}\right|_{\lambda=i\omega} = \frac{\omega^2}{K_1^2+K_2^2}\left(\omega^6+3\omega^4+3(1-2\beta)\omega^2+2\beta(5-\beta)+1\right),$$

where

$$K_1 = 1-3\omega^2+2\beta-\tau(\beta+4\omega^2-(1-\omega^2)^2), \qquad K_2 = \omega[3-\omega^2-\tau(\beta-4(1-\omega^2))].$$

The second factor of (4.5) or either factor of (4.11) yields the same expression. Using (4.14) gives

$$\left.\frac{d\,[\mathrm{Re}(\lambda)]}{d\tau}\right|_{\lambda=i\omega} = \frac{\omega(1+\omega^2)^3}{2(K_1^2+K_2^2)}\,\frac{d\,d_{im}}{d\omega},$$

along the curves associated with pure imaginary roots of (4.5). Along the curves associated with pure imaginary roots of (4.11), d_{im} is replaced by $-d_{im}$ so the derivative is of opposite sign. The result follows. ∎

From section 3, when $\tau = 0$ and $\beta < -\frac{1}{2}$ the characteristic equation has a pair of pure imaginary roots $\lambda = \pm i\sqrt{-1+\sqrt{-2\beta}} = \pm i\omega_{\mathcal{S}}$ at the positive value $d = d^+ \equiv (2-\sqrt{-\beta/2})^2$. Similarly, when $\tau = 0$ and $\beta \leq -8$ the characteristic equation has pairs of pure imaginary roots $\lambda = \pm i\omega_{\mathcal{C}}^+, \pm i\omega_{\mathcal{C}}^-$, as defined in (4.10), at the following negative values of d:

$$d^-_{\pm} \equiv -\frac{\omega_{\mathcal{C}}^{\pm\,2}(\omega_{\mathcal{C}}^{\pm\,2}-3)^2}{(1-\omega_{\mathcal{C}}^{\pm\,2})^2}.$$

The roots at d_+^- also occur when $\beta > 1$. Note that d^+ must correspond to the value of d_{im} when $\tau = 0$ and $d_\pm^-$ to values of $-d_{im}$ when $\tau = 0$. More specifically, the definitions (4.8)–(4.9) and (4.12)–(4.13) of $\tau_{k\pm}^{\pm}$ give the following. For $\beta \le -8$, d^+ is the d intercept of the curve $(d_{im}(\omega), \tau_{0+}^{+}(\omega))$ and $d_\mp^-$ are the d intercepts of the curves $(-d_{im}(\omega), \tau_{0\pm}^{-}(\omega))$. For $-8 < \beta < -\frac{1}{2}$, d^+ is the d intercept of the curve $(d_{im}(\omega), \tau_{0-}^{+}(\omega))$. For $\beta > 1$, d_+^- is the d intercept of the curve $(-d_{im}(\omega), \tau_{0-}^{-}(\omega))$.

We now describe the region of stability of the trivial equilibrium in the $d\tau$-plane for intervals of values of β by finding bifurcation curves on which an eigenvalue has zero real part. This is the content of the rest of this section. Theorem 4.5 is illustrated (using Maple) in Figure 4.1(a) with $\beta = 1.5$, Theorem 4.6 is illustrated in Figure 4.1(b) with $\beta = 0.1$, and Theorem 4.7 is illustrated in Figure 4.4 with $\beta = -10$.

THEOREM 4.5. *Let $1 \le \beta \le 12 - 4\sqrt{5}$ be fixed. Then the trivial solution of* (4.1) *is linearly asymptotically stable for $d < d_+^-$, $0 \le \tau < \tau_{0-}^{-}$.*

Proof. From section 3, for $\tau = 0$ and $\beta \ge 1$, all roots of the characteristic equation have negative real parts if $d < d_+^-$ (see Figure 3.1). For fixed β and d, as τ is increased the number of roots with positive real parts remains the same until τ reaches the smallest value for which the characteristic equation has a pair of pure imaginary roots. This value is τ_{0-}^{-}. To see this, note from Lemma 4.4 that d_{im} is a continuous nondecreasing function of ω with $0 \le d_0 \le d_{im} < \infty$ and that $d_+^- < -d_0$. Thus for any fixed $d < d_+^-$ there is one positive value of ω such that $-d_{im}(\omega) = d$. Further, it is straightforward to show that, for any value of ω, $\tau_{0-}^{-}(\omega) < \tau_{0+}^{-}(\omega) < \tau_{k\pm}^{-}(\omega)$ for $k = 1, 2, \ldots$. Thus all the roots of the characteristic equation have negative real parts in the given range of d and τ. Applying the results of Lemma 4.4 shows there is at least one root of the characteristic equation with positive real part everywhere else in the $d\tau$-plane. ∎

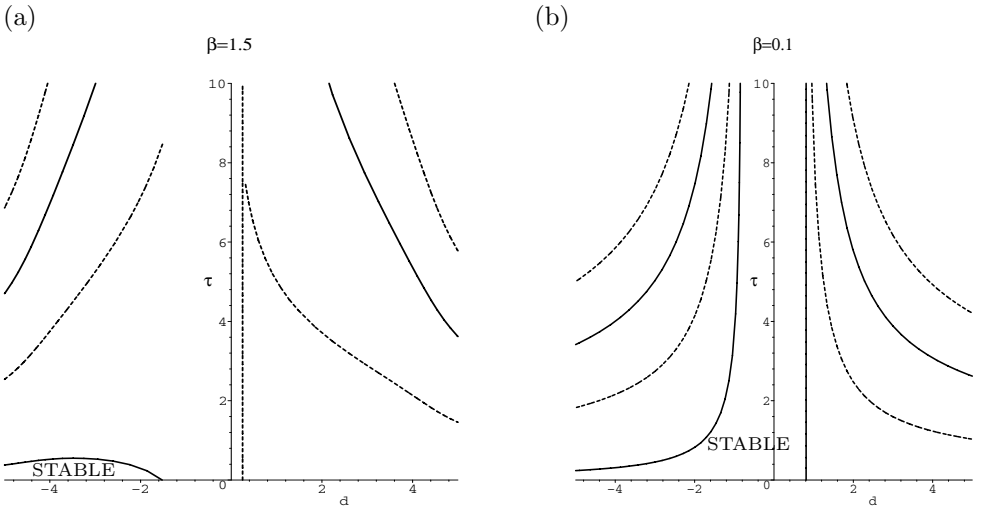


FIG. 4.1. *Bifurcation curves for the trivial solution of* (4.1) *for* (a) $\beta = 1.5$, (b) $\beta = 0.1$. *The stability region is qualitatively the same for* (a) $1 \le \beta \le 12 - 4\sqrt{5}$, (b) $\frac{1}{2}(5 - 3\sqrt{3}) \le \beta < 1$. *Along the solid (dashed) curves with $d > 0$, $\Delta_-^+(\lambda)$ ($\Delta_+^+(\lambda)$) has a pair of pure imaginary roots. Along the solid (dashed) vertical line $d = d_0$, $\Delta_-^+(\lambda)$ ($\Delta_+^+(\lambda)$) has a zero root. Along the solid (dashed) curves with $d < 0$, $\Delta_-^-(\lambda)$ ($\Delta_+^-(\lambda)$) has a root $i\omega$ with $\omega > 0$ and $\Delta_+^-(\lambda)$ ($\Delta_-^-(\lambda)$) has the complex conjugate root $-i\omega$.*

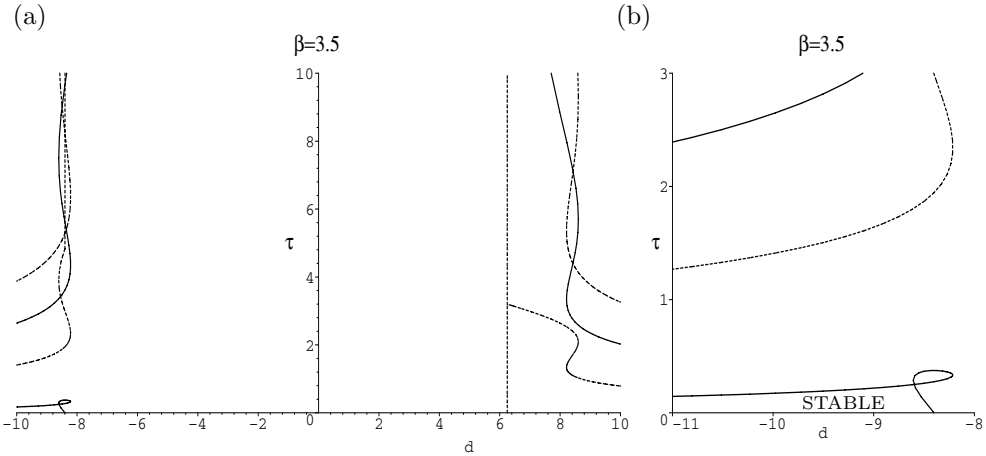


FIG. 4.2. (a) *Bifurcation curves for the trivial solution of* (4.1) *for* $\beta = 3.5$. (b) *Close-up showing stability region. The meaning of the solid and dashed curves are as described for Figure* 4.1.

THEOREM 4.6. *Let* $\frac{1}{2}(5-3\sqrt{3}) \le \beta < 1$ *be fixed. Then the trivial solution of* (4.1) *is linearly asymptotically stable for* $-d_0 \le d < d_0$, $0 \le \tau$ *or* $d < -d_0$, $0 \le \tau < \tau_{0-}^{-}$.

Proof. From section 3, for $\tau = 0$ and $\frac{1}{2}(5 - 3\sqrt{3}) \le \beta < 1$, all roots of the characteristic equation have negative real parts if $d < d_0$ (see Figure 3.1). Using the same argument as in the proof of Theorem 4.5, it can be shown that all roots of the characteristic equation have negative real parts for $d < -d_0$ and $0 \le \tau < \tau_{0-}^{-}$. From Lemma 4.3, d_{im} is a monotone function of ω for the assumed β range. Thus, using Lemma 4.2 for $\omega > 0$, $d_{im}(\omega) \ge d_{im}(0) = d_0$ and $-d_{im}(\omega) \le -d_{im}(0) = -d_0$. Hence for $-d_0 \le d < d_0$ and $\tau \ge 0$ all roots of the characteristic equation have negative real parts. The rest of the proof is the same as for Theorem 4.5. □

When β no longer lies in the first range given in Lemma 4.3, the curves along which the characteristic equation has pure imaginary roots become nonmonotone. This has two consequences. First, there will be values of ω such that $d_{im}(\omega) < d_0$, and second, there may exist intersection points of the curves $(d_{im}(\omega), \tau_{j\pm}^{+}(\omega))$ and $(-d_{im}(\omega), \tau_{j\pm}^{-}(\omega))$ with each other and with the line $d = d_0$. In this situation, the boundary of the stability region is made up of pieces of the curves $(d_{im}(\omega), \tau_{j\pm}^{+}(\omega))$ and $(-d_{im}(\omega), \tau_{j\pm}^{-}(\omega))$ and of the line $d = d_0$.

Consider first the case $\beta > 12 - 4\sqrt{5}$. For this range of β, we observe that $(d, \tau) = (-d_{im}, \tau_{0-}^{-})$ intersects itself. The stability region is still bounded by the d axis for $d < d_{+}^{-}$ and the curve $(-d_{im}(\omega), \tau_{0-}^{-}(\omega))$. However, part of the curve now forms a loop, inside which the trivial solution is unstable (this may be verified by applying Lemma 4.4). This is illustrated in Figure 4.2 with $\beta = 3.5$. We believe that the stability region is qualitatively the same for any $\beta > 12 - 4\sqrt{5}$.

Now consider the range $\beta < \frac{1}{2}(5 - 3\sqrt{3})$. For $\beta \le -\frac{1}{2}$ part of the curve of pure imaginary eigenvalues $(d_{im}(\omega), \tau_{0-}^{+}(\omega))$ enters the nonnegative τ region (this can be seen from the limits in Lemma 4.2). Using this fact, the discussion above, and the results of Lemmas 4.2 and 4.3, it can be shown that for $-\frac{1}{2} \le \beta < \frac{1}{2}(5 - 3\sqrt{3})$ the stability region looks qualitatively as depicted in Figure 4.3(a) and for $-8 < \beta < -\frac{1}{2}$ it looks qualitatively as depicted in Figure 4.3(b).

As β is decreased, the curves of pure imaginary eigenvalues approach the τ axis (and the stability region shrinks) until at $\beta = -8$ their points of minimal d value actually touch the τ axis. Recall that for $\beta < -8$, the curves $(-d_{im}(\omega), \tau_{0-}^{-}(\omega))$

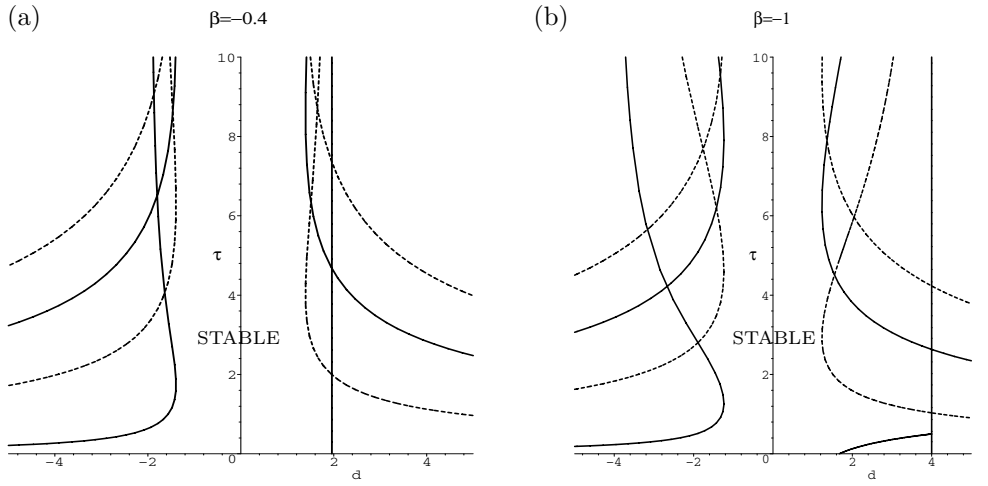


FIG. 4.3. *Bifurcation curves for the trivial solution of* (4.1) *for* (a) $\beta = -0.4$, (b) $\beta = -1$. *The stability region is qualitatively the same for* (a) $-\frac{1}{2} \le \beta < \frac{1}{2}(5 - 3\sqrt{3})$, (b) $-8 < \beta < -\frac{1}{2}$. *The meaning of the solid and dashed curves is as described for Figure* 4.1.

and $(-d_{im}(\omega), \tau_{0+}^{-}(\omega))$ intersect the d axis at d_+^- and d_-^-, respectively. From their definitions, (4.12)–(4.13), and the fact that $d_-^- < d_+^-$, these curves must have an intersection point. We denote this point by (d_{int}, τ_{int}) and have the following result, illustrated in Figure 4.4, for $\beta = -10$.

THEOREM 4.7. *Let $\beta \le -8$ be fixed. Then the trivial solution of* (4.1) *is linearly asymptotically stable for $d < d_-^-$, $0 \le \tau < \tau_{0-}^{-}$ or $d_-^- \le d < d_{int}$, $\tau_{0+}^{-} < \tau < \tau_{0-}^{-}$.*

Proof. From section 3, for $\tau = 0$ and $\beta \le -8$, all the roots of the characteristic equation have negative real parts if $d < d_-^-$ (see Figure 3.1). Using the same argument as in the proof of Theorem 4.5, it can be shown that all roots of the characteristic equation have negative real parts for $d < d_-^-$ and $0 \le \tau < \tau_{0-}^{-}$. For $\tau = 0$ and $d_-^- < d < d_+^-$, the characteristic equation has two roots with positive real parts.

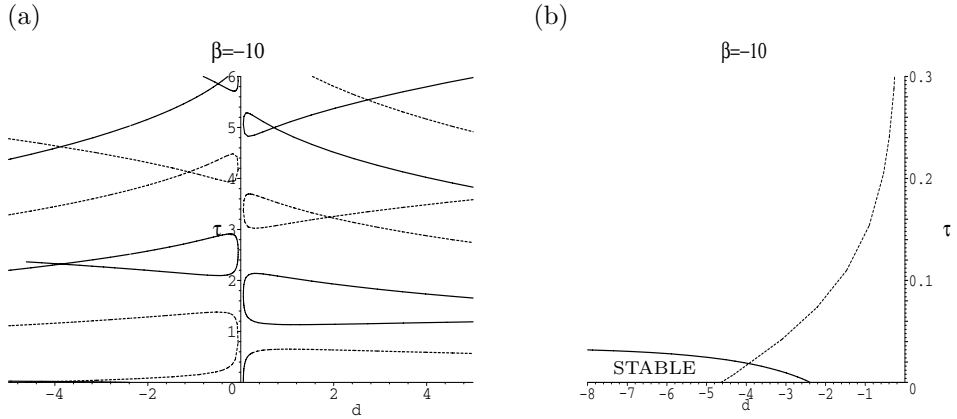


FIG. 4.4. (a) *Bifurcation curves for the trivial solution of* (4.1) *for* $\beta = -10$. (b) *Close-up showing stability region. The stability region is qualitatively the same for any* $\beta \le -8$. *The meaning of the solid and dashed curves is as described for Figure* 4.1.

Applying Lemma 4.4 shows that the number of roots with positive real parts decreases by two along the part of $(-d_{im}(\omega), \tau_{0+}^{-}(\omega))$, where $-d_{im}$ is increasing, and increases by two along the part of $\left(-d_{im}(\omega), \tau_{0-}^{-}(\omega)\right)$, where $-d_{im}$ is decreasing. The rest of the proof is the same as for Theorem 4.5. □

4.2. Bifurcations. In the previous subsection, we determined all points in parameter space where the trivial solution of (4.2) has eigenvalues with zero real parts. The bifurcations that may occur at such points as a system parameter is varied are important, particularly when they lie on the boundary of the stability region, as they determine the observable behavior of the system.

Consider first the case when a zero root of (4.4) exists. This occurs for parameter values along the line $d = d_0$. For $\beta \neq 1$, it can be shown that the conditions for a pitchfork bifurcation to occur are satisfied at almost all points on this line. In particular, taking d as the bifurcation parameter, Lemma 4.1 shows that the root is simple for $\beta \neq 1$ and $\tau \neq 2 + \frac{3}{\beta - 1}$. To ensure that the characteristic equation has no other roots with zero real part, the points of intersection of the line $d = d_0$ with the curves $(d_{im}(\omega), \tau_{\pm}^{+}(\omega))$ and $(-d_{im}(\omega), \tau_{\pm}^{-}(\omega))$ must also be excluded. In terms of the original model parameters, taking d as the bifurcation parameter is equivalent to fixing b, τ, and one of the c_j and using the other c_j as the bifurcation parameter.

Consider now the case when a pair of pure imaginary roots of (4.4) exists. This occurs for parameter values on the curves $(d_{im}(\omega), \tau_{\pm}^{+}(\omega))$ and $(-d_{im}(\omega), \tau_{\pm}^{-}(\omega))$. A statement of the Hopf bifurcation theorem for delay equations can be found in [12, Chapter 11]. It can be shown that this theorem is satisfied at almost all points on these curves. In particular, taking τ as the bifurcation parameter, Lemma 4.4 shows that the roots are simple at all points where $\frac{dd_{im}}{d\omega} \neq 0$. To ensure that the characteristic equation has no other roots with zero real part, the points of intersection of each curve with $d = d_0$ and the other curves where pure imaginary roots exist must be excluded.

If there is slightly more symmetry in the model, then some interesting patterns in the bifurcating solutions emerge. Suppose that $c_1 = c_2 = c$, as in parts of section 3 (e.g., Theorem 3.2), implying that $d = b^2c^2 > 0$. In this case, only the pitchfork bifurcation and the Hopf bifurcations along $(d_{im}(\omega), \tau_{\pm}^{+}(\omega))$ can occur. Consider the bifurcations that occur at a point in parameter space where $\Delta_{-}^{+}(\lambda)$ has a root with zero real part. (This corresponds to the solid curves in the figures of the previous subsection.) When $bc > 0$ it is straightforward to show that the solution of (4.2) corresponding to a root λ of $\Delta_{-}^{+}(\lambda)$ has the form $e^{\lambda t}(y_1, y_2, y_3, y_1, y_2, y_3)^t$. Thus we expect that the bifurcating solutions have a similar property—namely, the corresponding elements of the two loops are in phase, or synchronized. Similarly, when $bc < 0$ the solution of (4.2) corresponding to a root λ of $\Delta_{-}^{+}(\lambda)$ has the form $e^{\lambda t}(y_1, y_2, y_3, -y_1, -y_2, -y_3)^t$, and we expect the bifurcating solutions have corresponding elements of the two loops antiphase (or half a period out of phase). The solutions corresponding to roots of $\Delta_{+}^{+}(\lambda)$ have just the opposite property. When $bc > 0$ they are antiphase and when $bc < 0$ they are in-phase. When $c_1 \neq c_2$ but $c_1 \approx c_2$, we expect that bifurcating solutions are almost in-phase or almost antiphase. Such behavior was observed in [18].

Now suppose that $c_1 = -c_2 = c$, implying that $d = -b^2c^2 < 0$. In this case, only Hopf bifurcations along the curves $(-d_{im}(\omega), \tau_{\pm}^{-}(\omega))$ occur. When the characteristic equation has a pair of roots $\lambda = \pm i\omega$, corresponding solutions of (4.2) have the form $e^{\lambda t}(y_1, y_2, y_3, \pm iy_1, \pm iy_2, \pm iy_3)^t$. Thus corresponding elements of the bifurcating periodic orbits are one quarter period out of phase. When $c_1 \neq -c_2$ but $c_1 \approx -c_2$,

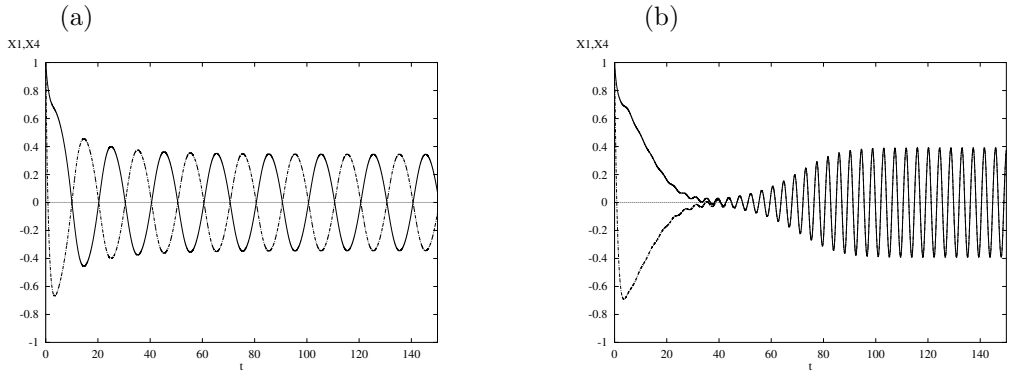


FIG. 4.5. *Numerical simulations of* (4.1) *with* $b = -1$, $c_1 = c_2 = 1.75$. *The plots in each case show* x_1 *(solid line) and* x_4 *(dot-dash line) vs.* t. (a) $\tau = 0.3$*; periodic orbit with* $x_4(t) = -x_1(t)$. (b) $\tau = 1.5$*; periodic orbit with* $x_4(t) = x_1(t)$. *Initial conditions for both cases* $x(t) = (1, -0.7, -0.9, 1.1, 0.8, 1.2)^t$, $-\tau \le t \le 0$.

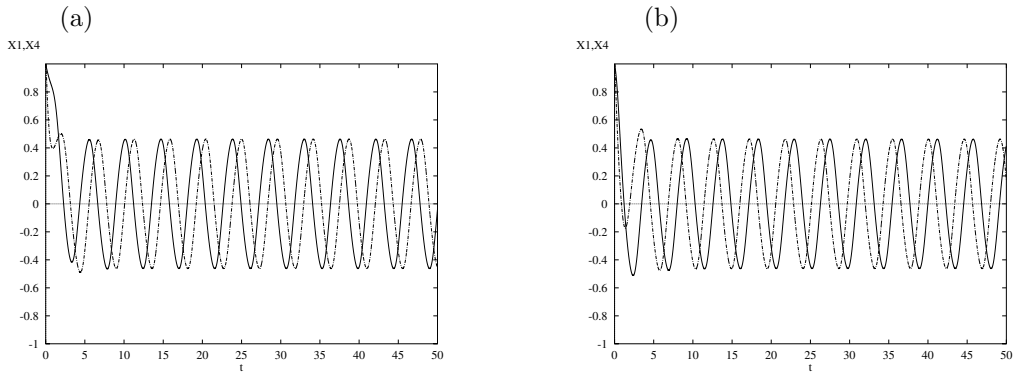


FIG. 4.6. *Numerical simulations of* (4.1) *with* $b = -1$, $\tau = 0.5$, *and* $c_1 = -c_2$. *The plots in each case show* x_1 *(solid line) and* x_4 *(dot-dash line) vs.* t. (a) $c_1 = 1.75$*; periodic orbit with* $x_4(t) = x_1(t - \frac{T}{4})$, *where* T *is the period.* (b) $c_1 = -1.75$*; periodic orbit with* $x_4(t) = x_1(t + \frac{T}{4})$. *Initial conditions as for Figure* 4.5.

we expect that bifurcating solutions are close to one quarter period out of phase.

These results are illustrated in Figures 4.5–4.6, showing numerical simulations of (4.1), with $b = -1$ and other parameters as indicated, which correspond to points in the stability diagram of Figure 4.3(b). Only x_1 and x_4 are shown in Figures 4.5–4.6; solutions for other pairs are similar. Simulations were performed in XPPAUT [6] using a fourth order Runge–Kutta integrator adapted for delay differential equations.

5. Discussion. Combining local and global results with numerical evidence, we arrive at the following summary of the dynamics of these loops. Results are given in terms of $\beta = b^3$ and $d = b^2 c_1 c_2$, where b is the gain of the response function for each neuron and c_i are the coupling strengths between the 3-loops. The loops are inherently (i.e., in isolation) oscillatory for $\beta < -8$. The origin is proved to be globally stable for $\beta \in (-2\sqrt{2}, 1)$, and numerical evidence extends this to $(-8, 1)$. For

$\beta > 1$ the solutions approach a nontrivial stable fixed point (the origin is unstable).

The effect of coupling depends on whether it is symmetric (excitatory in both directions or inhibitory in both directions) or asymmetric (excitatory in one direction and inhibitory in the other). It is interesting that the linear stability analysis is identical for excitatory and inhibitory coupling, as long as it is the same in both directions, as it depends essentially on the product of the two coupling coefficients. This was also noted in the somewhat similar situation studied by Shayer and Campbell [18]. Symmetric coupling of sufficient strength (not necessarily very strong) can destabilize the origin in the middle (inherently stable) β range. When $\beta \in (-\frac{1}{2}, 1)$, the system goes to nontrivial fixed points, but when $\beta \in (-8, -\frac{1}{2})$, it first goes to oscillation as coupling is increased. Asymmetric coupling of sufficient strength (and here it needs to be quite strong) can stabilize the origin in either of the two inherently unstable ranges. The further β is from the inherently stable range, the stronger the coupling needs to be to accomplish this stabilization. In the case of symmetric coupling, nontrivial equilibria exist when β is large enough, but there are no nontrivial equilibria for smaller β when the coupling is weak. In the case of asymmetric coupling, there are no nontrivial equilibria for $\beta < 1$. It is not clear whether nontrivial equilibria occur for other regions of parameter space. For most regions, oscillation of the system is suggested when the linear results show that the origin is unstable.

We have observed five main delay-related phenomena in this system.

1. When coupling is asymmetric ($d < 0$) and large, the stability of the origin is weak in the sense that only a small delay is needed to destabilize it and produce oscillation. This is delay-induced oscillation or delay-induced instability, which has commonly been observed in delayed networks since the early work of Marcus and Westervelt [15].
2. In the inherently stable range $\beta \in (-8, 1)$, delay independent stability exists for weak enough coupling ($|d|$ small) whether symmetric or asymmetric.
3. For intermediate values of $|d|$ and $\beta \in (-8, -0.098)$, whether the system oscillates or settles at the origin depends on the delay in a complex way. For some delay ranges, the origin is stable, and for others it is unstable, and there can be stability/instability switches as the delay increases.
4. For $\beta \in (-8, -\frac{1}{2})$, if coupling is symmetric and fairly strong ($d > 0$ and large but still $< (1-\beta)^2$), in the region where coupling has destabilized the origin to create oscillation, there is an intermediate range of delays (not including zero but not too large) that stabilizes the origin again and suppresses the oscillations. This is delay-induced stability or *oscillator death*.
5. For equal and symmetric coupling strengths, oscillatory solutions in the two loops bifurcating from the origin may be in phase or antiphase depending on the value of the delay. For asymmetric coupling with equal strengths, corresponding neurons in the two loops oscillate one quarter period out of phase.

Some of these results are similar to those found by Shayer and Campbell [18] for a simpler coupled system. However, their work focused on the symmetric coupling case.

Some properties of coupled systems that can each potentially oscillate begin to emerge from these studies—in particular, the ways in which oscillation or instability depends on the interaction between coupling strength and coupling delay. Although the system studied here is too simple to draw definite conclusions about physiological systems, results do show that complicated effects can occur even in the simplest

coupled loops with delay. This study could be extended by investigating other patterns of coupling between two loops, such as "lateral" coupling between each corresponding pair of units in the two loops (if the loops have the same structure), or "forward" coupling as studied without delays by Edwards and Gill [5]. For applications in which the units are far apart, it would be worthwhile to include delays in connections within each loop.

Acknowledgment. We wish to acknowledge the assistance of Daisuke Shinki in carefully checking the results and proofreading the paper.

REFERENCES

[1] P. Baldi and A. Atiya, *How delays affect neural dynamics and learning*, IEEE Trans. Neural Networks, 5 (1994), pp. 612–621.

[2] H. Bergman, A. Feingold, A. Nini, A. Raz, H. Slovin, M. Abeles, and E. Vaadia, *Physiological aspects of information processing in the basal ganglia of normal and Parkinsonian primates*, Trends Neurosci., 21 (1998), pp. 32–38.

[3] S. A. Campbell, *Stability and bifurcation of a simple neural network with multiple time delays*, Fields Inst. Commun., 21 (1999), pp. 65–79.

[4] S. A. Campbell, *Delay independent stability for additive neural networks*, Differential Equations Dynam. Systems, 9 (2001) pp. 115–138.

[5] R. Edwards and P. Gill, *On synchronization and cross-talk in parallel networks*, Dyn. Contin. Discrete Impuls. Syst. Ser. B Appl. Algorithms, 10 (2003), pp. 287–300.

[6] B. Ermentrout, *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*, Software Environ. Tools 14, SIAM, Philadelphia, 2002.

[7] L. Glass and C. P. Malta, *Chaos in multi-looped negative feedback systems*, J. Theor. Biol., 145 (1990), pp. 217–223.

[8] L. Glass and J. S. Pasternack, *Stable oscillations in mathematical models of biological control systems*, J. Math. Biol., 6 (1978), pp. 207–223.

[9] K. Gopalsamy and X.-Z. He, *Delay independent stability in bidirectional associative memory networks*, IEEE Trans. Neural Networks, 5 (1994), pp. 998–1002.

[10] A. M. Graybiel, *Basal ganglia—input, neural activity, and relation to the cortex*, Curr. Opin. Neurobiol., 1 (1991), pp. 644–651.

[11] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

[12] J. Hale and S. M. V. Lunel, *Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.

[13] N. Kopell and G. B. Ermentrout, *Phase transitions and other phenomena in chains of coupled oscillators*, SIAM J. Appl. Math., 50 (1990), pp. 1014–1052.

[14] L. Olien and J. Bélair, *Bifurcations, stability, and monotonicity properties of a delayed neural network model*, Phys. D, 102 (1997), pp. 349–363.

[15] C. Marcus and R. Westervelt, *Stability of analog neural networks with delay*, Phys. Rev. A (3), 39 (1989), pp. 347–359.

[16] I. Ncube, S. A. Campbell, and J. Wu, *Change in criticality of synchronous Hopf bifurcation in a multiple-delayed neural system*, Fields Inst. Commun., 36 (2003), pp. 179–193.

[17] K. Pakdaman, C. Grotta-Ragazzo, C. P. Malta, O. Arino, and J.-F. Vibert, *Effect of delay on the boundary of the basin of attraction in a system of two neurons*, Neural Networks, 11 (1998), pp. 509–519.

[18] L. P. Shayer and S. A. Campbell, *Stability, bifurcation, and multistability in a system of two coupled neurons with multiple time delays*, SIAM J. Appl. Math., 61 (2000), pp. 673–700.

[19] P. van den Driessche, J. Wu, and X. Zou, *Stabilization role of inhibitory self-connections in a delayed neural network*, Phys. D, 150 (2001), pp. 84–90.

[20] P. van den Driessche and X. Zou, *Global attractivity in delayed Hopfield neural network models*, SIAM J. Appl. Math., 58 (1998), pp. 1878–1890.

[21] J. Wu, T. Faria, and Y. S. Huang, *Synchronization and stable phase-locking in a network of neurons with memory*, Math. Comput. Modelling, 30 (1999), pp. 117–138.

SIAM J. APPL. MATH.
Vol. 65, No. 1, pp. 336–360

SINGULARITY FORMATION IN CHEMOTAXIS— A CONJECTURE OF NAGAI*

HOWARD A. LEVINE† AND JOANNA RENCŁAWOWICZ†‡

Abstract. Consider the initial-boundary value problem for the system (S) $u_t = u_{xx} - (uv_x)_x$, $v_t = u - av$ on an interval $[0,1]$ for $t > 0$, where $a > 0$ with $u_x(0,t) = u_x(1,t) = 0$. Suppose μ, v_0 are positive constants. The corresponding spatially homogeneous global solution $U(t) = \mu$, $V(t) = \mu/a + (v_0 - \mu/a)\exp(-at)$ is stable in the sense that if (μ', v_0') are positive constants, the corresponding spatially homogeneous solution will be uniformly close to $(U(\cdot), V(\cdot))$.

We consider, in sequence space, an approximate system (S') which is related to (S) in the following sense: The chemotactic term $(uv_x)_x$ is replaced by the inverse Fourier transform of the finite part of the convolution integral for the Fourier transform of $(uv_x)_x$. (Here the finite part of the convolution on the line at a point x of two functions, f, g, is defined as $\int_0^x (f(y)g(y-x)\,dy$.) We prove the following:

(1) If $\mu > a$, then in every neighborhood of (μ, v_0) there are (spatially nonconstant) initial data for which the solution of problem (S') blows up in finite time in the sense that the solution must leave $L^2(0,1) \times H^1(0,1)$ in finite time T. Moreover, the solution components $u(\cdot,t), v(\cdot,t)$ each leave $L^2(0,1)$.

(2) If $\mu > a$, then in every neighborhood of (μ, v_0) there are (spatially nonconstant) initial data for which the solution of problem (S) on $(0,1) \times (0, T_{max})$ must blow up in finite time in the sense that the coefficients of the cosine series for (u,v) become unbounded in the sequence product space $\ell^1 \times \ell^1_1$.

A consequence of (2) states that in every neighborhood of (μ, v_0), there are solutions of (S) which, if they are sufficiently regular, will blow up in finite time. (Nagai and Nakaki [*Nonlinear Anal.*, 58 (2004), pp. 657–681] showed that for the original system such solutions are unstable in the sense that if $\mu > a$, then in every neighborhood of $(\mu, \mu/a)$, there are spatially nonconstant solutions which blow up in finite *or* infinite time. They conjectured that the blow-up time must be finite.) Using a recent regularity result of Nagai and Nakaki, we prove this conjecture.

Key words. chemotaxis, finite time singularity formation, Keller–Segel model

AMS subject classifications. 35K55, 92C17

DOI. 10.1137/S0036139903431725

1. Introduction.

The classical equations of chemotaxis were introduced in [15, 16, 22]. A variant of them, which was later discussed in [14], takes the form

$$\begin{aligned} u_t &= D_1(\Delta u - \nabla \cdot (u\chi(v)\nabla v)), \\ v_t &= D_2 \Delta v + \lambda u - av \end{aligned} \tag{1.1}$$

when the diffusivities D_j are constant. The constants in (1.1) are presumed to be positive. Generally speaking, $u(\cdot,t)$ represents the cell concentration (or local population) of some species, while $v(\cdot,t)$ corresponds to the concentration of a chemotactic agent such as cyclic adenosine monophosphate (cAMP). The function $\chi(v)$ is called the chemotactic sensitivity.

*Received by the editors July 14, 2003; accepted for publication (in revised form) January 20, 2004; published electronically October 28, 2004. The first author was supported in part by Mathematical Sciences Biology Institute of Ohio State University. This material is based upon work supported by the National Science Foundation under agreement 0112050. The second author was supported by a NATO postdoctoral fellowship and by KBN 2-P03A-002-23.

http://www.siam.org/journals/siap/65-1/43172.html

†Department of Mathematics, Iowa State University, Ames, IA 50011 (halevine@iastate.edu, joannar@iastate.edu).

‡Institute of Mathematics, Polish Academy of Sciences, Śniadeckich 8, 00-956 Warsaw, Poland (jr@impan.gov.pl).

This system is not of standard reaction-diffusion type since the first equation will involve the Laplacian of the second dependent variable. Typically, the boundary conditions are of Neumann type for the first equation and of either mixed or Dirichlet type for the second equation. Precise conditions will be given later.

In the growing literature on singularity formation in chemotaxis, the problems studied tend to fall into one of two types.[1]

In the first type, the time scale of the second equation is assumed to be much smaller than that of the first, i.e., $D_2 \gg D_1$, or the cell species has infinite propagation speed, while the chemical species has diffused so rapidly that it has come to a steady state. In this case, the system simplifies into one of elliptic-parabolic type, namely,

$$\begin{aligned} u_t &= D_1(\Delta u - \nabla \cdot (u\chi(v)\nabla v)), \\ 0 &= D_2\Delta v + \lambda u - av. \end{aligned} \tag{1.2}$$

A number of papers have been concerned with the phenomenon of blowup for such systems. See [2, 1, 5, 9, 10, 8, 19], for example. The first such result in this direction seems to be contained in [13]. The rough idea of the approach to this system is to solve, at least in principle, the second (elliptic) equation in (1.2) for v as a nonlocal, but linear, function of u and then eliminate v from the first equation leaving a nonlinear, nonlocal dynamical equation for u.

In the second, and probably less well studied form, the time scale of the second equation is assumed to be much larger than that of the first, so fast in fact that the diffusion of the chemical species can be neglected. (That is, $D_2 \ll D_1$.) In this case the spatial movement of the chemical is being controlled by the movement of the particles which the chemical influences through its gradient. This was used as an example to illustrate the modelling approach to *Dictyostelium discoideum* movement taken in [21].

The reaction term $\lambda u - av$ is the cAMP saturated limit approximation of a reaction term that more accurately reflects Michealis–Menten reaction kinetics. A model of *D. discoideum* movement which views the cell receptors as the catalyzing agent for cAMP production can be found in [6, pp. 498ff.]. However, if cAMP is not in excess, then one must replace the reaction term by a term of the form $\frac{k_1 uv}{k_2+v} - av$, where $\lambda = k_1/k_2$, in order to more accurately describe the cell receptor kinetics involved (see [6, pp. 273ff.]). The system then takes the form

$$\begin{aligned} u_t &= D_1\nabla \cdot \left\{ u\nabla\left[\ln \frac{u}{\psi(v)}\right]\right\}, \\ v_t &= R(u,v) = \frac{k_1 uv}{k_2+v} - av. \end{aligned} \tag{1.3}$$

The advantage of writing the system in this form is that one can see that if the system is tending toward a steady state, u should follow $\psi(v)$.

Indeed, by using the principles of reinforced random walk [4], the authors of [21] derived the first of equations (1.3) ab initio. However, by writing the flux vector $\overrightarrow{J}$ as

$$\overrightarrow{J} = -D_1(\overrightarrow{\nabla} u - \chi(v)u\overrightarrow{\nabla} v),$$

[1] If one sets $\tau = D_1 t$, then D_1 drops out of the first equation, while the second equation becomes $D_1\partial_\tau v = D_2\Delta v + \lambda u - av$. One then lets $D_1 \to 0$ in this equation in the first case. In the second case, one lets $D_1, a, \lambda \to \infty$ in such a way that $a/D_1, \lambda/D_1$ remain constant.

defining $\psi(v)$ by the equation $\psi'(v)/\psi(v) = \chi(v)$, and using the continuity equation $u_t = -\overrightarrow{\nabla} \cdot \overrightarrow{J}$, we obtain the first equation in (1.3) by continuum mechanical considerations.

Here the solution approach, taken, for example in [17], is to solve the second equation for u as a nonlinear function of v, v_t and then eliminate it from the first equation, leaving a rather messy third order equation in v. Reference [17] is devoted to a detailed discussion and interpretation of the numerical results obtained there and earlier in [21] for the resulting equation.

In [17], the discussion begins with consideration of the following special case of (1.3) on an interval $[0, 1]$ for $t > 0$ with $\psi(v) = v$:

$$\begin{aligned} u_t &= \nabla \cdot \left\{ u \nabla \left[\ln \frac{u}{v} \right] \right\}, \\ v_t &= R(u, v) = uv \end{aligned} \tag{1.4}$$

with $u_x(0,t) = u_x(1,t) = 0$ and $v_x(0,t) = v_x(1,t) = 0$ (which then imply the zero flux conditions $u_x = uv_x$ at $x = 0, 1$). The vector $[\mu(x,t), v_0(x,t)]^t \equiv [1, e^t]^t$ is a spatially homogeneous solution of (1.4) with $[1, 1]$ as initial datum. The following statement is a consequence of the results of [17]: Let c be the positive root of $c^2 + Nc - 1 = 0$ and let $0 < \epsilon < 1$. Given any mode number N, there is a direction $[u_N, v_N]^t \equiv [Nc, 1]^t \cos(Nx)$ in the closed subspace of $L^2(0,1) \times H^1(0,1)$ consisting of the closure of functions which satisfy $u[\log(u/v)]_x = 0$ at $x = 0, 1$, and a curve given by $\overrightarrow{R}(\varepsilon) \equiv [u(\cdot, 0, \varepsilon), v(\cdot, 0, \varepsilon)]^t$ in $L^2(0,1) \times H^1(0,1)$ of initial data passing through $[1, 1]$ with the property that any solution initially emanating from this curve will blow up in a finite time.

This solution is given by $u = \psi_t(x,t)$, $v = \exp(\psi)$, where

$$\psi(x,t) = t - \ln[1 - 2\varepsilon \exp(Nct) \cos(N\pi x) + \varepsilon^2 \exp(2Nct)].$$

Moreover, this solution has the important biological property that it leaves the above space by aggregation, in particular, by virtue of the fact that $\|u(\cdot,t)\|_{L^2(0,1)}$ blows up in finite time. It is conceivable that for such systems, $\|u(\cdot,t)\|_{L^2(0,1)}$ can remain bounded, while $\|u_x(\cdot,t)\|_{L^2(0,1)}$ blows up in finite time. In [21], this possibility was demonstrated numerically, while in [17], a plausibility argument was given to show that these numerical results were not just artifacts of the simulations and were to be expected from the underlying dynamical system.

The result tells us that in every neighborhood of the initial data for the spatially homogeneous solution $[1, e^t]^t$, there are solutions of arbitrarily high initial total variation which begin in this neighborhood and blow up in finite time. The numerical evidence suggests that *every* arbitrarily small nonconstant perturbation of the initial data for $[1, e^t]^t$ (which must have a nontrivial projection onto at least one of the directions $[Nc, 1]^t \cos(Nx)$ for some N) must blow up in finite time. (This interpretation was not spelled out in [17].)

Clearly if we replace v by $\exp(v)$ in the system (1.4), there results

$$\begin{aligned} u_t &= u_{xx} - (uv_x)_x, \\ v_t &= u, \end{aligned} \tag{1.5}$$

which, when $\epsilon \geq 0$, is a special case of

$$\begin{aligned} u_t &= u_{xx} - (uv_x)_x, \\ v_t &= \epsilon v_{xx} + u - av. \end{aligned} \tag{1.6}$$

Equation (1.6) is the classical system studied in [3]. In its turn, this system contains as a special case the system of Nagai and Nakaki taken up in section 2. However, it is important to note that in [20], the authors establish the well posedness of the initial-boundary value problem for this system (with homogeneous Neumann boundary conditions) as well as the existence of a global attractor. Their proof demands that $\epsilon > 0$ in order to establish the existence of a global Lyapunov functional. An alternate proof of this result has been given in [12]. There the authors also provide an asymptotic profile of the solution.

Thus we are left with the question of what happens to solutions when the dissipation in v is weak, i.e., when $\epsilon = 0$ and $a > 0$. This is the problem raised by Nagai and partially addressed by him and Nakaki in [18].

The plan of the paper is as follows. In section 2 we discuss their system and the results they established recently [18]. There we also discuss a related initial value problem for their system and introduce a closely related approximate initial value problem. In section 3 we reformulate the Nagai–Nakaki system as an infinite system of nonlinear ordinary differential equations. In section 4, we introduce a second infinite system of ordinary differential equations which is closely related to the system of ordinary differential equations in section 3. This second system is related to the first in much the same way as the approximate initial value problem is related to the full initial value problem for their system.

In section 5 we establish the local existence and uniqueness of solutions of their system when $\mu > a$ in the sequence space $\ell_1 \times \ell_1^1$, i.e., in the space of pairs of sequences $(\{a_i\}, \{b_i\})$ such that $\sum_{i\geq 1}(|a_i| + i|b_i|)$ is finite. (This sequence space is continuously and injectively imbedded in $L^1(0,1) \times W^{1,1}(0,1)$. However, the inverse of the injection (restricted to the image) is not continuous. We discuss this point in more detail in section 7.)

In section 6, we demonstrate that the spatially homogeneous solutions of the system of ordinary differential equations for the approximate problem are unstable in the sense that in every neighborhood of the spatially homogeneous solution there are solutions in $L^2(0,1) \times H^1(0,1)$ with spatially inhomogeneous data which blow up in finite time in this space.

In section 7, we establish this conjecture. That is, if $\mu > a$, then in every neighborhood of (μ, v_0) there are (spatially nonconstant) initial data for which the corresponding solution of the Nagai–Nakaki problem in the cylinder must blow up in finite time (in a sense to be made precise below). This result will yield the Nagai–Nakaki conjecture for solutions that are sufficiently regular.

2. The system of Nagai and Nakaki [18]. Nagai, in a talk given at the International Conference on Partial Differential Equations and Mathematical Biology, Wuhan, China, 2001, considered the following initial-boundary value problem:

$$\begin{aligned} u_t &= u_{xx} - (uv_x)_x, \\ v_t &= u - av \end{aligned} \tag{2.1}$$

with $a \geq 0$. As boundary conditions he took

$$\begin{aligned} u_x(0,t) &= v_x(0,t)u(0,t), \\ u_x(1,t) &= v_x(1,t)u(1,t). \end{aligned} \tag{2.2}$$

These boundary conditions follow from the conditions $u_x(0,t) = u_x(1,t) = 0$ and $v_x(0,0) = v_x(1,0) = 0$ because the second equation of (2.1) implies that $v_x(0,t), v_x(1,t)$

satisfy $y'(t) = -ay(t)$. (In what follows we will refer to (2.1) and (2.2) as Nagai's problem or the Nagai–Nakaki problem.)

To set the notational stage for what follows, we introduce a potential function $\psi(x,t)$ as follows: We let $v(x,t) \equiv V(t)+\psi(x,t)$, $u(x,t) \equiv U(t)+\tilde{u}(x,t)$. The spatially homogeneous solution of Nagai's problem is

$$(V(t), U(t)) = (\mu/a + (v_0 - \mu/a)\exp(-at), \mu).$$

This and the second equation force the choice for $\tilde{u} = \psi_t + a\psi$ so that

$$(\psi, \psi_t + a\psi)^t = (v - V(t), u - \mu) = (v,u)^t - (V(t), U(t))^t. \tag{2.3}$$

In particular, in what follows, the reader is cautioned that $(\psi, \psi_t + a\psi)^t$ corresponds to the pair $(v, u + a\psi)^t$. (That is, with reference to [18], the pairing is (u,v), while in the theorems and proofs here, the pairing is (v,u).)

We are interested in those potential functions for which

$$\int_0^1 \psi\, dx = \int_0^1 \psi_t\, dx = 0 \tag{2.4}$$

in order to ensure that the mass, $\int_0^1 u\, dx$, is conserved.

Recently Nagai and Nakaki [18] have established the following statements for a restricted class of initial data perturbations for which the corresponding solutions are more regular than $L^2(0,1) \times H^1(0,1)$.

To save the reader a bit of time, we give a rough summary of their results. (By initial values, we mean initial values for which the mean value of u is μ and which are not identical with the stationary solution $(\mu, \mu/a)$.)

1. If the initial values are sufficiently regular (in particular if they are analytic functions) and if they satisfy the boundary conditions, then the solution components u, v will be in $H^2(0,1)$ and will be continuous or continuously differentiable in time into the appropriate range on the interval of existence.
2. If $\mu < a$ and the initial values are sufficiently regular, then (u,v) approaches $(\mu, \mu/a)$ exponentially rapidly in the norm of $H^1(0,1) \times H^2(0,1)$.
3. Suppose $\mu > a$ and suppose that the initial values are sufficiently regular.
   a. If the initial data satisfy

$$W(u(,\cdot,0), v(\cdot,0)) < \mu \ln \mu - \frac{\mu^2}{2a},$$

   where $f > 0, g \geq 0$ and where

$$W(f,g) \equiv \int_0^1 (f \ln f - fg + ag^2/2)\, dx,$$

   and if the solution exists for all time, then each component of (u,v) blows up in the H^1 norm as $t \to +\infty$.
   b. If the blow-up time is finite, then each component blows up in L^∞, i.e., pointwise.
3. There are also solutions in each such neighborhood which converge to the steady state solutions in infinite time in the norm of $H^1(0,1) \times H^2(0,1)$.

Stability results for related problems have been established in [7, 23]. All of the exact solutions found in [23] were found earlier in [17] in rescaled form, contrary to

the implication in [23, p. 776]. (The solutions in [23] follow from those of [17] after a shift in the time axis.)

In his talk, Nagai mentioned that he and Nakaki were unable to resolve whether or not the blowup occurred in finite time. Therefore, we shall refer to the following statement as *Nagai's conjecture.* There are choices of initial data $(u(\cdot,0), v(\cdot,0))$ for which the blow-up time must be finite.

In view of the results of [17], when $a = 0$, the blowup must occur in finite time. Motivated by this simple observation, we set about trying to establish this for Nagai's problem. However, we were unable to establish this claim in $L^2(0,1) \times H^1(0,1)$.

In the course of our investigations, we happened upon a problem, which is, in a sense, close to that of the problem of Nagai, for which Nagai's conjecture holds for $(u, v) \in L^2(0,1) \times H^1(0,1)$. Moreover, this result allows us to prove the conjecture of Nagai in a Banach space different from that proposed by Nagai. More precisely, if we denote by $\{(a_n(t), b_n(t))\}_{n=1}^{\infty}$ the sequence of cosine coefficients for the pair (ψ, ψ_t), we show that $\sum_n (n|a_n(t)| + |b_n(t)|)$ must blow up in finite time.

As remarked above, blowup in sequence space in this sense does *not* imply blowup of the L^1 norm of ψ, ψ_x, or ψ_t.

We see that ψ satisfies

$$\psi_{tt} + (\mu - a)\psi_{xx} = (\psi_{tx} - \psi_t\psi_x)_x - a(\psi_t + (\psi\psi_x)_x). \tag{2.5}$$

In order to motivate the approximate problem, we digress for a moment and consider the pure initial value problem for (2.5). If we compute the Fourier transform $\varphi(\xi, t) = \widehat{\psi(x,t)} = \int_{-\infty}^{\infty} e^{-i\xi x}\psi(x,t)dx$ and assume that ψ, ψ_x vanish at $x = \pm\infty$ on any interval $[0, T)$ where the solution of the initial value problem exists, we find (suppressing the second argument on the right)

$$\begin{aligned}
\varphi_{tt} + (a + \xi^2)\varphi_t + \xi^2(a - \mu)\varphi &= \frac{a\xi^2}{2}\varphi * \varphi(\xi) + \xi\varphi_t * (\eta\varphi) \\
&= \frac{1}{2}\int_0^{\xi} [a\xi^2\varphi(\xi - \eta)\varphi(\eta) + 2\xi(\xi - \eta)\varphi(\xi - \eta)\varphi_t(\eta)]\, d\eta \\
&\quad + \frac{1}{2}\int_{\xi}^{\infty} \{a\xi^2\varphi(\xi - \eta)\varphi(\eta) + \xi[(\xi - \eta)\varphi(\xi - \eta)\varphi_t(\eta) \\
&\qquad + \eta\varphi(\eta)\varphi_t(\xi - \eta)]\}\, d\eta.
\end{aligned} \tag{2.6}$$

If ψ, ψ_x are sufficiently regular, then

$$\lim_{|\xi|\to+\infty} \int_{\xi}^{\infty} a\xi^2\varphi(\xi - \eta)\varphi(\eta) + \xi[(\xi - \eta)\varphi(\xi - \eta)\varphi_t(\eta) + \eta\varphi(\eta)\varphi_t(\xi - \eta)]\, d\eta = 0. \tag{2.7}$$

We can estimate the terms in (2.7) as follows:

$$\left| \int_{\xi}^{\infty} \xi^2\varphi(\xi - \eta)\varphi(\eta)\, d\eta \right| \le \xi^2 \|\varphi\|_{L^\infty} \int_{\xi}^{\infty} |\varphi(\eta)|\, d\eta,$$

while

$$\frac{1}{2}\left| \int_{\xi}^{\infty} \xi[(\xi - \eta)\varphi(\xi - \eta)\varphi_t(\eta) + \eta\varphi(\eta)\varphi_t(\xi - \eta)]\, d\eta \right| \le |\xi| \|\eta\varphi(\eta)\|_{L^\infty} \int_{\xi}^{\infty} |\varphi_t(\eta)|\, d\eta.$$

These inequalities give us an idea of how rapidly the transform of the solution should decay.

The Fourier transform of the partial differential equation which approximates Nagai's problem is the equation

$$\varphi_{tt} + (a+\xi^2)\varphi_t + \xi^2(a-\mu)\varphi = \frac{1}{2}\int_0^\xi [a\xi^2\varphi(\xi-\eta)\varphi(\eta) + 2\xi(\xi-\eta)\varphi(\xi-\eta)\varphi_t(\eta)]\,d\eta \equiv \Phi(\phi,\phi_t)(\xi,t). \tag{2.8}$$

The nonlinear partial differential equation itself can be recovered from (2.8) in the form

$$\psi_{tt} + [a\psi + \psi_{xx}]_t + (\mu-a)\psi_{xx} = \widehat{\Phi}(\phi,\phi_t)(x,t), \tag{2.9}$$

where $\widehat{\Phi}$ denotes the inverse Fourier transform of Φ. We call Φ the finite part of the Fourier transform of $(uv_x)_x$. (Notice that, except for the factor ξ^2, Φ is the sum of the finite parts of two convolutions, one of ϕ with itself and the other of $\phi_t(\xi)$ with $\xi\phi(\xi)$.)

In section 4, we derive a system of ordinary differential equations for the cosine coefficients of the initial-boundary value problem for (2.9). This initial-boundary value problem corresponds to the natural initial-boundary value problem for (2.5) which arises from Nagai's problem, i.e., $\psi_x = \psi_{xt} = 0$ at $x = 0, 1$ and $t > 0$ with ψ, ψ_t prescribed at $t = 0$. The resulting system of ordinary differential equations for the cosine coefficients of the solution of (2.9) is related to the corresponding system of ordinary differential equations for the cosine coefficients of the solution of Nagai's problem in much the same way that (2.8) is related to (2.6).

In order to see how the former system comes about, we next derive the corresponding system of ordinary differential equations for Nagai's problem.

3. Reformulation of the Nagai–Nakaki system as a system of ordinary differential equations. We introduce some notation. Let $\beta \geq 0$ and $i \in \{1, 2\}$. We work in the spaces $\ell^i_\beta([0,T))$ of sequences of real valued functions $\{g_n(t)\}_{n=1}^\infty$ on $[0,T)$ for which $\sum_{n=1}^\infty n^\beta |g_n(t)|^i < \infty$. When $\beta = 0$ we omit the subscript. That is, $\ell^i_\beta = \ell^i_0 = \ell^i$.

We say that a sequence of differentiable functions $\{g_n(t)\}_{n=1}^\infty$ is in $\ell^i_\beta([0,T)) \times \ell^j_{\beta'}([0,T))$ if the sequence $\{g_n(t)\}_{n=1}^\infty$ is in $\ell^i{}_\beta([0,T))$ and the sequence $\{g_n'(t)\}_{n=1}^\infty$ is in $\ell^j_{\beta'}([0,T))$.

Assuming that $\mu > a$, we seek a solution of this equation in the form

$$\psi(x,t) = \sum_{n=1}^\infty g_n(t)\cos(Cnx), \tag{3.1}$$

where $C = 2\pi M$ for some integer M. With this choice of M the conservation conditions (2.4) hold.

Consequently,

$$(\psi_t\psi_x)_x = -C\sum_{n=2}^{\infty}\sum_{k+l=n} kg_kg_l'(\cos(Clx)\sin(Ckx))_x$$

$$= -\frac{1}{2}C^2\sum_{n=2}^{\infty}\sum_{k+l=n} kg_kg_l'[(k+l)\cos(C(k+l)x) + (k-l)\cos(C(k-l)x)] \tag{3.2}$$

$$= -\frac{1}{2}C^2\sum_{n=2}^{\infty}\sum_{k=1}^{n-1} kg_kg_{n-k}'[n\cos(Cnx) + (2k-n)\cos(C(2k-n)x)],$$

and

$$(\psi\psi_x)_x = -C\sum_{n=2}^{\infty}\sum_{k+l=n} kg_kg_l(\cos(Clx)\sin(Ckx))_x$$

$$= -\frac{1}{2}C^2\sum_{n=2}^{\infty}\sum_{k+l=n} kg_kg_l[(k+l)\cos(C(k+l)x) + (k-l)\cos(C(k-l)x)] \tag{3.3}$$

$$= -\frac{1}{4}C^2\sum_{n=2}^{\infty}\sum_{k=1}^{n-1} g_kg_{n-k}[n^2\cos(Cnx) + (2k-n)^2\cos(C(2k-n)x)].$$

Then (2.5) can be rewritten in the form

$$\sum_{n=1}^{\infty}(g_n'' + (C^2n^2+a)g_n' - (\mu-a)C^2n^2g_n)\cos(Cnx)$$

$$= \frac{1}{2}C^2\sum_{n=2}^{\infty}\sum_{k=1}^{n-1}\{kg_kg_{n-k}'\,[n\cos(Cnx) + (2k-n)\cos(C(2k-n)x)]$$

$$+\frac{a}{2}g_kg_{n-k}\left[n^2\cos(Cnx) + (2k-n)^2\cos(C(2k-n)x)\right]\}.$$

The terms involving $\cos(C(2k-n)x)$ on the right-hand side can be rewritten by switching the order of summation and setting $l = 2k-n$ to obtain

$$\sum_{n=2}^{\infty}\sum_{k=1}^{n-1}(2k-n)kg_kg_{n-k}'\cos(C(2k-n)x) = \sum_{k=1}^{\infty}\sum_{n=k+1}^{\infty}(2k-n)kg_kg_{n-k}'\cos(C(2k-n)x)$$

$$= \sum_{k=1}^{\infty}\sum_{l=-\infty}^{k-1} lkg_kg_{k-l}'\cos(Clx)$$

$$= \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}(-l)kg_kg_{k+l}'\cos(Clx)$$

$$+\sum_{k=1}^{\infty}\sum_{l=1}^{k-1} lkg_kg_{k-l}'\cos(Clx)$$

$$= -\sum_{l=1}^{\infty}\sum_{k=1}^{\infty} lkg_kg_{k+l}'\cos(Clx)$$

$$+\sum_{l=1}^{\infty}\sum_{k=l+1}^{\infty} lkg_kg_{k-l}'\cos(Clx).$$

Likewise,

$$\begin{aligned}\sum_{n=2}^{\infty}\sum_{k=1}^{n-1} g_k g_{n-k}(2k-n)^2\cos(C(2k-n)x) &= \sum_{k=1}^{\infty}\sum_{n=k+1}^{\infty} g_k g_{n-k}(2k-n)^2\cos(C(2k-n)x)\\ &= \sum_{k=1}^{\infty}\sum_{l=-(k-1)}^{\infty} g_k g_{l+k} l^2\cos(Clx)\\ &= 2\sum_{l=1}^{\infty}\left(l^2\sum_{k=1}^{\infty} g_k g_{l+k}\right)\cos(Clx),\end{aligned}$$

the third line following from the second by breaking up the inner sum on the right of the second line into an infinite sum over positive integers and a finite sum over negative integers $l=-1,\dots,-(k-1)$. The latter inner sum is then rewritten as a finite sum over positive indices and the order changed in the resultant double sum.

Therefore, we have

$$\begin{aligned}&\sum_{n=1}^{\infty}(g_n''+(C^2n^2+a)g_n'-(\mu-a)C^2n^2g_n)\cos(Cnx)\\ &\quad=\frac{1}{2}C^2\sum_{n=2}^{\infty}\sum_{k=1}^{n-1}\left(nkg_kg_{n-k}'+\frac{a}{2}n^2g_kg_{n-k}\right)\cos(Cnx)\\ &\qquad+\frac{1}{2}C^2\sum_{n=1}^{\infty}n\left[\sum_{k=n+1}^{\infty}kg_kg_{k-n}'-\sum_{k=1}^{\infty}kg_kg_{k+n}'+an\sum_{k=1}^{\infty}g_kg_{n+k}\right]\cos(Cnx).\end{aligned}\tag{3.4}$$

We obtain the following (infinite) system of ordinary differential equations:

$$\begin{aligned}\mathfrak{L}_n g_n &\equiv g_n''+(C^2n^2+a)g_n'-(\mu-a)C^2n^2g_n\\ &=\frac{1}{2}C^2n\left\{\sum_{k=1}^{n-1}\left(kg_kg_{n-k}'+\frac{a}{2}ng_kg_{n-k}\right)\right.\\ &\qquad\left.+\sum_{k=1}^{\infty}\left[(n+k)g_{n+k}g_k'-kg_kg_{k+n}'+ang_kg_{n+k}\right]\right\}\quad\text{for}\quad n=1,2,\dots.\end{aligned}\tag{3.5}$$

In order to rewrite (3.5) in a more compact form, we introduce the notation

$$\begin{aligned}|g| &= \{|g_k|\}_{k=1}^{\infty},\\ T_ng &= \{g_{n+k}\}_{k=1}^{\infty}\ \text{(shift operator)},\\ g' &= \{g_k'\}_{k=1}^{\infty}\quad\text{(differentiation)},\\ \mathcal{M}g &= \{kg_k\}_{k=1}^{\infty}\quad\text{(multiplication by the transform variable)},\\ g*h &= \left\{\sum_{k=1}^{n-1}g_kh_{n-k}\right\}_{n=1}^{\infty}\quad\text{(convolution)},\\ (g,h) &= \sum_{k=1}^{\infty}g_kh_k\ \ \text{(scalar product in } \ell^2).\end{aligned}$$

Then one can solve the Nagai–Nakaki system in the aforementioned function space if and only if one can solve the initial value problem for the following system in the

corresponding sequence space:

$$\begin{aligned}\mathfrak{L}_n g_n &\equiv g_n'' + (C^2n^2 + a)g_n' - (\mu - a)C^2n^2 g_n \\ &= \frac{1}{2}C^2 n\Big\{(\mathcal{M}g * g')_n + n\frac{a}{2}(g*g)_n + [(T_n\mathcal{M}g, g') - (\mathcal{M}g, T_n g')] + an(g, T_n g)\Big\}.\end{aligned} \tag{3.6}$$

4. A system of ordinary differential equations related to the Nagai–Nakaki system. If we consider (3.6) without the last three terms on the right-hand side, we obtain the following system of ordinary differential equations:

$$\begin{aligned}\mathfrak{L}_n g_n &\equiv g_n'' + (C^2n^2 + a)g_n' - (\mu - a)C^2n^2 g_n \\ &= \frac{1}{2}C^2 n\left\{\sum_{k=1}^{n-1}\left(k g_k g_{n-k}' + \frac{a}{2}n g_k g_{n-k}\right)\right\} \\ &\equiv \frac{1}{2}C^2 n\Big\{(\mathcal{M}g * g')_n + n\frac{a}{2}(g*g)_n\Big\}.\end{aligned} \tag{4.1}$$

That is, (4.1) is the discrete version of (2.8). It is the system satisfied by the cosine coefficients of solutions of (2.9).

Comparing (4.1) and (2.8), we infer that the three terms in (3.6) that do not appear in (4.1) can be viewed as "tail ends" of integrals. That is, they are analogous to the three integrals that have been dropped in passing from (2.6) to (2.8) under the assumption that ψ, ψ_x are sufficiently regular.

This suggests that $(T_n\mathcal{M}g, g')$, $(\mathcal{M}g, T_n g')$, $na(g, T_n g)$ can be neglected in comparison with the remaining terms on the right-hand side of (3.6).

Such a statement needs rigorous proof.

5. Local existence and uniqueness of solutions of the initial value problem for the Nagai–Nakaki system. We prove the following result in $\ell_1^1 \times \ell^1$ which was proved in [18] in a product of smoother spaces.

LEMMA 1. *The solution $\{g_n(\cdot)\}_{n=1}^{\infty}$ of the system* (3.5) *exists locally in time and is unique in $\ell_1^1 \times \ell^1$ on the interval of local existence. Moreover, the solution is uniformly bounded in the $\ell_1^1 \times \ell^1$ norm on compact subsets of the existence interval.*

Proof. First consider the question of uniqueness. Set $w_n = g_n - h_n$, where g_n and h_n are two solutions of the above system for $n \geq 2$ for which $g(0) = h(0)$, $g'(0) = h'(0)$. The difference w_n satisfies the equation

$$\begin{aligned}\mathfrak{L}_n w_n &\equiv w_n'' + (C^2n^2 + a)w_n' - (\mu - a)C^2n^2 w_n \\ &= \frac{1}{2}C^2 n\left\{\sum_{k=1}^{n-1}\left[k(w_k g_{n-k}' + h_k w_{n-k}') + ak(w_k g_{n-k} + h_k w_{n-k})\right]\right. \\ &\qquad + \sum_{k=1}^{\infty}\left[(n+k)(w_{n+k}g_k' + h_{n+k}w_k') - k(w_k g_{n+k}' + h_k w_{n+k}')\right] \\ &\qquad \left. + 2a\sum_{k=1}^{\infty}k(w_k g_{n+k} + h_k w_{n+k})\right\},\end{aligned} \tag{5.1}$$

which, in the above notation, becomes

$$\begin{aligned}\mathfrak{L}_n w_n = \frac{1}{2}C^2 n\{&(\mathcal{M}w * g')_n + (\mathcal{M}h * w')_n + a[(\mathcal{M}w * g)_n + (\mathcal{M}h * w)_n] \\ &+ (T_n\mathcal{M}w, g') + (T_n\mathcal{M}h, w') - (\mathcal{M}w, T_n g') \\ &- (\mathcal{M}h, T_n w') + 2a[(\mathcal{M}w, T_n g) + (\mathcal{M}h, T_n w)]\}.\end{aligned}$$

We abbreviate this as

$$\mathfrak{L}_n w_n = \mathcal{F}_n(w, w'), \tag{5.2}$$

where we suppress the dependence of the right-hand side on g, h, g', h' for the moment. The characteristic equation for each of the linear second order operators $\mathfrak{L}_n$ is

$$r^2 + (C^2n^2 + a)r - (\mu - a)C^2n^2 = 0.$$

The roots r_n^+, r_n^- are real with $r_n^+ > 0 > r_n^-$, with $r_n^+ \to 2(\mu - a) \equiv r^+$ as $n \to \infty$, and with $r_n^+ \le r^+$ for all n while $r_n^- \to -\infty$. Since the initial values for w vanish, the solution of (5.1) can be written as

$$w_n = A_n(\mathcal{F}_n, t)e^{r_n^+ t} + B_n(\mathcal{F}_n, t)e^{r_n^- t}, \tag{5.3}$$

where

(5.4)

$$A_n(\mathcal{F}_n, t) = \int_0^t \frac{-e^{r_n^- s}\mathcal{F}_n(s)}{W} ds = \frac{1}{\sqrt{(C^2n^2 + a)^2 + 4(\mu - a)C^2n^2}} \int_0^t e^{-r_n^+ s}\mathcal{F}_n(s)ds,$$
$$B_n(\mathcal{F}_n, t) = \int_0^t \frac{e^{r_n^+ s}\mathcal{F}_n(s)}{W} ds = \frac{-1}{\sqrt{(C^2n^2 + a)^2 + 4(\mu - a)C^2n^2}} \int_0^t e^{-r_n^- s}\mathcal{F}_n(s)ds.$$

Because

$$\max\{|B_n(\mathcal{F}_n, t)e^{r_n^- t}|,\ |A_n(\mathcal{F}_n, t)e^{r_n^+ t}|\} \le \frac{c}{n^2} e^{r_n^+ t} \int_0^t e^{-r_n^+ s}|\mathcal{F}_n(s)|ds$$

for some computable constant c, we have

$$\begin{aligned} \|\mathcal{M}w(t)\|_{\ell^1} = \sum_{n=1}^{\infty} n|w_n(t)| &\le \sum_{n=1}^{\infty} \frac{c}{n} e^{r_n^+ t} \int_0^t e^{-r_n^+ s}|\mathcal{F}_n(s)|ds \\ &\le e^{r^+ t} \int_0^t \sum_{n=1}^{\infty} \frac{c}{n}|\mathcal{F}_n(s)|ds. \end{aligned} \tag{5.5}$$

We need to estimate the terms in the last integral. For each index n, there are 10 sums arising from the 10 terms in the definition of $\mathcal{F}_n/n$. After use of the convolution inequality, we have that

$$\sum_{n=1}^{\infty} |(\mathcal{M}w * g')_n| \le \sum_{n=1}^{\infty} \sum_{k=1}^{n-1} k|w_k||g'_{n-k}| = \||\mathcal{M}w| * |g'|\|_{\ell^1} \le \|\mathcal{M}w(t)\|_{\ell^1} \|g'(t)\|_{\ell^1}.$$

The next three sums are bounded above by a constant multiple of $\|w'(t)\|_{\ell^1}\|\mathcal{M}h(t)\|_{\ell^1}$, $\|\mathcal{M}w(t)\|_{\ell^1}\|\mathcal{M}g(t)\|_{\ell^1}$, and $\|\mathcal{M}w(t)\|_{\ell^1}\|\mathcal{M}h(t)\|_{\ell^1}$, respectively.

The terms involving T_n are a bit trickier to estimate. We have, for the first of them,

$$\sum_{n=1}^{\infty} |(T_n\mathcal{M}w, g')| \le \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} (n+k)|w_{n+k}||g'_k| = \sum_{k=1}^{\infty} |g'_k| \sum_{l=k+1}^{\infty} l|w_l| \le \sum_{k=1}^{\infty} |g'_k| \|T_k\mathcal{M}w\|_{\ell^1}.$$

Since $\|T_k\mathcal{M}w\|_{\ell^1}$ is decreasing in k, we obtain

$$\sum_{k=1}^{\infty}|g_k'||T_k\mathcal{M}w|_{\ell^1}\le\|\mathcal{M}w(t)\|_{\ell^1}\|g'(t)\|_{\ell^1}.$$

In a similar fashion, the remaining five sums are found to be bounded above by a constant multiple of $\|w'(t)\|_{\ell^1}\|\mathcal{M}h(t)\|_{\ell^1}$, $\|\mathcal{M}w(t)\|_{\ell^1}\|g'(t)\|_{\ell^1}$, $\|w'(t)\|_{\ell^1}\|\mathcal{M}h(t)\|_{\ell^1}$, $\|\mathcal{M}w(t)\|_{\ell^1}\|\mathcal{M}g(t)\|_{\ell^1}$, and $\|\mathcal{M}w(t)\|_{\ell^1}\|\mathcal{M}h(t)\|_{\ell^1}$, respectively.

Thus, for some constant B

$$\|\mathcal{M}w(t)\|_{\ell^1}\le Be^{r^+t}\int_0^t(\|\mathcal{M}w(s)\|_{\ell^1}+\|w'(s)\|_{\ell^1})(\|\mathcal{M}g(s)\|_{\ell^1}+\|g'(s)\|_{\ell^1}+\|\mathcal{M}h(s)\|_{\ell^1}+\|h'(s)\|_{\ell^1})ds.$$

Assume that $g(\cdot),h(\cdot)$ are in $\ell^1_1([0,T])$ and g',h' are in $\ell^1([0,T])$; i.e., on every compact subset K of $[0,T)$ there is a finite constant $M(K)$ such that

$$\max\left\{\sum_{k=1}^{\infty}(k|g_k|+|g_k'|),\sum_{k=1}^{\infty}(k|h_k|+|h_k'|)\right\}<M(K). \tag{5.6}$$

There results

$$\|\mathcal{M}w(t)\|_{\ell^1}\le Be^{r^+t}\int_0^t(\|\mathcal{M}w(s)\|_{\ell^1}+\|w'(s)\|_{\ell^1})ds \tag{5.7}$$

for some new computable constant $B=B(M(K))$.

Next we estimate $\|w'(t)\|_{\ell^1}$. Using the representation formula (5.3) we obtain

$$w_n'(t)=\frac{1}{\sqrt{(C^2n^2+a)^2+4(\mu-a)C^2n^2}}\left\{r_n^+\int_0^t e^{r_n^+(t-s)}\mathcal{F}_n(s)ds-r_n^-\int_0^t e^{r_n^-(t-s)}\mathcal{F}_n(s)ds\right\}. \tag{5.8}$$

Consequently, for some positive computable constants c,d we have, noting that $r_n^-\approx -dn^2$,

$$|w_n'(t)|\le c\left\{\frac{r^+}{n^2}\int_0^t e^{r^+(t-s)}|\mathcal{F}_n(s)|ds+\int_0^t\left[ne^{-dn^2(t-s)}\right]\frac{|\mathcal{F}_n(s)|}{n}ds\right\}. \tag{5.9}$$

The first term on the right-hand side is treated exactly as above. We note that for $c_1=1/\sqrt{2ed}$, all positive integers n, and all $s\le t$,

$$ne^{-dn^2(t-s)}\le\frac{c_1}{\sqrt{t-s}}.$$

Using this in (5.9) and summing over n we obtain

$$\|w'(t)\|_{\ell^1}\le c\left\{\int_0^t r^+e^{r^+(t-s)}\sum_{n=1}^{\infty}\frac{|\mathcal{F}_n(s)|}{n^2}ds+\int_0^t\frac{c_1}{\sqrt{t-s}}\sum_{n=1}^{\infty}\frac{|\mathcal{F}_n(s)|}{n}ds\right\}. \tag{5.10}$$

Since we have already estimated the integrand sums $\sum_{n=1}^{\infty}\frac{|\mathcal{F}_n(s)|}{n}$ above, we obtain

$$\|w'(t)\|_{\ell^1}\le c_1\left\{e^{r^+t}\int_0^t\frac{c}{\sqrt{t-s}}(\|\mathcal{M}w(s)\|_{\ell^1}+\|w'(s)\|_{\ell^1})\,ds\right\}. \tag{5.11}$$

Setting $\varphi(t) = \|\mathcal{M}w(t)\|_{\ell^1} + \|w'(t)\|_{\ell^1}$, we have, with $c = c(T) = (cc_1 + B)e^{r^+T}$,

$$\varphi(t) \le c \int_0^t \frac{\varphi(s)}{\sqrt{t-s}}\, ds, \tag{5.12}$$

a Volterra integral inequality of Gronwall type with a weakly singular kernel. Thus, after an application of the Hölder inequality, it is easily shown that φ^q, with $1/p + 1/q = 1$ and $1 < p < 2$, satisfies a standard Gronwall inequality. That is,

$$\varphi^q(t) \le c^q \left(\frac{2}{2-p}\right)^{\frac{q}{p}} t^{\frac{q}{p}-\frac{q}{2}} \int_0^t \varphi^q(s)\, ds.$$

Thus, for $\phi(t) = \int_0^t \varphi^q(s)\, ds$ and $\tilde{c}(s) = c^q(\frac{2}{2-p})^{\frac{q}{p}} s^{\frac{q}{p}-\frac{q}{2}}$, we obtain

$$\frac{d}{dt}\left(\phi(t) e^{-\int_0^t \tilde{c}(s)\, ds}\right) \le 0.$$

Then

$$0 \le \phi(t) \le \phi(0) e^{\int_0^t \tilde{c}(s)\, ds}.$$

Since $\phi(0) = 0$, it follows that $\phi(t) \equiv 0$.

Consequently, $\varphi(t) \equiv 0$, and hence $g \equiv h$ on the existence interval.

Next, consider local existence. Since most of the estimates needed here have been worked through above, we will be brief.

Abbreviate (3.6) as $\mathfrak{L}g = F(g, g'), g(0) = g_0, g'(0) = g_0'$. Then the homogeneous solution $G(t)$ with inhomogeneous initial values solves

$$\begin{aligned} &\mathfrak{L}G = 0, \\ &G(0) = g_0, \quad G'(0) = g_0', \end{aligned}$$

while the function $H = g - G$ satisfies the nonlinear problem with homogeneous initial data:

$$\begin{aligned} &\mathfrak{L}H = F(G + H, G' + H'), \\ &H(0) = 0, \quad H'(0) = 0. \end{aligned}$$

Thus it suffices to show that the integral equation $H = \mathfrak{L}^{-1}F(G + H, G' + H')$ has a fixed point on some interval $[0, T_{exist})$. We see from the variation of parameters formula (5.3) that the components satisfy

$$H_n = A_n(\mathcal{F}_n, t)e^{r_n^+ t} + B_n(\mathcal{F}_n, t)e^{r_n^- t}.$$

Define the sequence of iterates $\{H^k\}_{k=1}^\infty$ as follows with $H^1(t) = \{H_n^1(t) = 0\}_{n=1}^\infty$ and, for $k \ge 1$,

$$H^{k+1} = \mathfrak{L}^{-1}F(G + H^k, G' + H^{k\prime}).$$

To construct a fixed point, we need to show that sequence $(\{H^k\}, \{H^{k\prime}\})$ is convergent in $\ell_1^1 \times \ell^1$. We do this by showing that we can apply the contraction mapping principle to the sequence. The argument usually goes in two steps.

1. There is a small time interval $[0,T]$ such that the sequence (of sequences) $(\{H^k\},\{H^{k'}\})_{k=1}^{\infty}$ is uniformly bounded in the norm on $\ell_1^1\times\ell^1$.
2. This sequence is contracting on this (or possibly smaller) time interval; i.e., there exists a constant $0\le\theta<1$ such that

$$\sup_{0\le s\le T}\|\mathcal{M}(H^{k+1}(s)-H^k(s))\|_{\ell^1}+\|(H^{k+1}-H^k)'(s)\|_{\ell^1}$$
$$\le\theta\sup_{0\le s\le T}\|\mathcal{M}(H^k-H^{k-1})(s))\|_{\ell^1}+\|(H^k-H^{k-1})'(s)\|_{\ell^1}.$$

To establish step 1, we find, as in the uniqueness proof, an inequality of the form

$$\|\mathcal{M}H^{k+1}(t)\|_{\ell^1}+\|(H^{k+1})'(t)\|_{\ell^1}$$
$$\le c\int_0^t\frac{1}{\sqrt{t-s}}(\|\mathcal{M}((G+H^k)(s))\|_{\ell^1}+\|(G+H^k)'(s)\|_{\ell^1})^2\,ds.$$

Because the roots $r_n^{\pm}$ are bounded above, we can assume that

$$\sup[0,T](\|\mathcal{M}G(s)\|_{\ell^1}+\|(G'(s)\|_{\ell^1})\le G_0(T)$$

for some constant G_0 depending only on T. Consequently, we have the estimate

$$\|\mathcal{M}H^{k+1}(t)\|_{\ell^1}+\|(H^{k+1})'(t)\|_{\ell^1}\le c\sqrt{t}\left\{G_0^2+[\|\mathcal{M}H^k(t)\|_{\ell^1}+\|(H^k)'(t)\|_{\ell^1}]^2\right\}.$$

If we set $Z_n=\sup_{[0,T]}[\|\mathcal{M}H^n(t)\|_{\ell^1}+\|(H^n)'(t)\|_{\ell^1}]$, then $Z_{n+1}\le c\sqrt{T}(G_0^2+Z_n^2)$. Let $G_1>0$ be any constant such that $Z_1\le G_1$. Then $Z_{n+1}\le G_1$ for all $n=1,2,3,\dots$, provided that T is so small that

$$\sqrt{T}\le\frac{G_1}{c(G_0^2+G_1^2)}.$$

Since both G_0,G_1 depend on the initial values for the free solution, we have the desired a priori bound on the iterates. For step 2, we examine the difference:

$$\mathfrak{L}(H^{k+1}-H^k)=F(G+H^k,G'+H^{k'})-F(G+H^{k-1},G'+H^{k-1'})\equiv\mathcal{F}(W^k,W^{k'}),$$

where $W^k=H^k-H^{k-1}$ and $\mathcal{F}$ depends in fact on $H^k,H^{k'},H^{k-1},G,G'$. That is,

$$\begin{aligned}\mathfrak{L}_nW_n^{k+1}=\frac{1}{2}C^2n\{&(\mathcal{M}W^k*(G+H^k)')_n+(\mathcal{M}(G+H^{k-1})*W^{k'})_n\\&+a[(\mathcal{M}W^k*(G+H^k))_n+(\mathcal{M}(G+H^{k-1})*W^k)_n]\\&+(T_n\mathcal{M}W^k,(G+H^k)')+(T_n\mathcal{M}(G+H^{k-1}),W^{k'})\\&-(\mathcal{M}W^k,T_n(G+H^k)')-(\mathcal{M}(G+H^{k-1}),T_nW^{k'})\\&+2a[(\mathcal{M}W^k,T_n(G+H^k))+(\mathcal{M}(G+H^{k-1}),T_nW^k)]\}.\end{aligned}$$

Consequently, as in the previous proof, after obtaining a similar expression for $W^{k'}$ we can then use the estimates in the previous part of the proof to derive the estimate of the form

$$\|\mathcal{M}W^{k+1}(t)\|_{\ell^1}+\|(W^{k+1})'(t)\|_{\ell^1}\le c\int_0^t\frac{1}{\sqrt{t-s}}(\|\mathcal{M}W^k(s)\|_{\ell^1}+\|(W^k)'(s)\|_{\ell^1})\,ds,$$

where

$$c = c(\|\mathcal{M}(G + H^{k-1})\|_{\ell^1}, \|G + H^k\|_{\ell^1}, \|(G + H^k)'\|_{\ell^1}).$$

By step 1, we may assume that c is uniformly bounded above by a constant $\tilde{c}(T, G(0), G'(0))$ for a sufficiently small time interval $[0, T]$. We proceed as in the previous proof, using the Hölder inequality to obtain the Gronwall inequality. From this, we easily show that on a sufficiently small time interval one can apply the contraction mapping principle to the sequence $(\{H^k\}, \{H^{k'}\})$ in $\ell_1^1 \times \ell^1$. We omit the details. □

6. Local existence and blowup of solutions of the approximate system. We turn next to the local existence theorem and blow-up theorem for the initial value problem for the approximate system (4.1). We choose the special initial sequence $g_n(0) = a_n$, $g_n'(0) = n\lambda a_n$, where a_n and λ are to be chosen in such a manner that $\{g_n(t) = a_n e^{n\lambda t}\}_{n=1}^{\infty}$ is a solution of the approximate system which must blow up in finite time.

In what follows, we adopt the following notation. Let M be a positive integer such that $\frac{a}{4\pi^2 M^2} \equiv a_* \leq 1$. Suppose that ε, δ are such that

$$0 < 2\varepsilon \leq a_1 \leq \frac{\lambda}{\lambda + a}\delta,$$

where λ is given by (6.3) below.

We establish the following theorems.

THEOREM 1 (local existence). *Suppose $\mu > a$ and the sequence $\{a_n\}_{n=2}^{\infty}$ solves the recurrence relation given by* (6.1) *below. Then the sequence $\{g_n(t) = a_n e^{n\lambda t}\}_{n=1}^{\infty}$ solves* (4.1) *on an interval $[0, T_e]$, where $T_e \geq T_* = -\frac{\ln \delta}{\lambda}$. Moreover, $(\psi(\cdot, t), \psi_t(\cdot, t))$ is in $H^1(0,1) \times L^2(0,1)$ on $[0, T_*)$, where $(\psi(\cdot, t))$ is given by* (3.1). *The function ψ is analytic on $(0,1) \times [0, T_*)$.*

THEOREM 2 (finite time blowup). *The function $(\psi(\cdot, t), \psi_t(\cdot, t))$ of the previous theorem must leave $H^1(0,1) \times L^2(0,1)$ in finite time $T_\infty \leq -\frac{\ln \varepsilon}{\lambda}$.*

Before proving these theorems, we show that for $n \geq 2$

$$\lambda(4\pi^2 M^2 n - a)(n-1)a_n = 2\pi^2 M^2 \sum_{k=1}^{n-1} [\lambda(n-k) + a] k a_k a_{n-k}. \tag{6.1}$$

This recurrence formula defines the sequence a_n. However, unlike the situation in [11, 17], we cannot find a simple expression for the coefficients a_n in terms of a_1 and n. Nonetheless, we can find upper and lower bounds for the series which sum to the related solutions found in [17] for the case $a = 0$. These estimates provide the necessary comparison functions for the existence and blowup of ψ in $H^1(0,1) \times L^2(0,1)$.

For convenience of notation, we use $\xi = x - 1/2$ in (3.1) instead of x. Then $\partial_x = \partial_\xi$ and single point blowup at $\xi = 0$ corresponds to blowup at $x = 1/2$. Therefore, if $g_n = a_n e^{n\lambda t}$ and $C = 2\pi M$, equation (4.1) reads

$$\begin{aligned} &\sum_{n=1}^{\infty} a_n \{4\pi^2 M^2 \lambda n^3 + \lambda^2 n^2 - (\mu - a)4\pi^2 M^2 n^2 + a\lambda n\} e^{n\lambda t} \cos(2\pi M n \xi) \\ &\qquad = 2\pi^2 M^2 \sum_{n=2}^{\infty} n \left\{ \sum_{k=1}^{n-1} (\lambda(n-k) + a) k a_k a_{n-k} \right\} e^{n\lambda t} \cos(2\pi M n \xi). \end{aligned} \tag{6.2}$$

Comparing coefficients for $n \geq 2$,

$$[4\pi^2 M^2 \lambda n^2 + n\left(\lambda^2 - (\mu - a)4\pi^2 M^2\right) + a\lambda]a_n = 2\pi^2 M^2 \sum_{k=1}^{n-1} [\lambda(n-k) + a]k a_k a_{n-k}.$$

For $n = 1$,

$$a_1\{\lambda^2 + \lambda(4\pi^2 M^2 + a) - (\mu - a)4\pi^2 M^2\} = 0.$$

Since $\mu > a$, the roots are real. Let

$$\lambda = \frac{1}{2}(\sqrt{(4\pi^2 M^2 + a)^2 + 16\pi^2 M^2(\mu - a)} - 4\pi^2 M^2 - a) \tag{6.3}$$

denote the positive root.[2] Then the relation for a_n, $n \geq 2$, simplifies to (6.1) as claimed. With the values of λ, a_* above, we have that

$$2\lambda(n - a_*)(n-1)a_n = \sum_{k=1}^{n-1} (\lambda(n-k) + a)k a_k a_{n-k}. \tag{6.4}$$

We are now in a position to prove the theorems. We begin with Theorem 1.

Proof. From (6.1), since $n \geq k + 1$,

$$\begin{aligned} 2\lambda(n - a_*)(n-1)a_n &\leq \sum_{k=1}^{n-1} (\lambda(n-k) + a)k a_k a_{n-k} \\ &\leq (\lambda + a) \sum_{k=1}^{n-1} k(n-k)\, a_k a_{n-k}. \end{aligned}$$

If $a_1 \leq b_1$, by induction it follows that $a_n \leq b_n$, where

$$2(n - a_*)(n-1)b_n = \frac{\lambda + a}{\lambda} \sum_{k=1}^{n-1} k(n-k) b_k b_{n-k}.$$

Because $n \geq 2$, we have $a_* \leq n/2$ and

$$n(n-1)b_n \leq \frac{\lambda + a}{\lambda} \sum_{k=1}^{n-1} k(n-k) b_k b_{n-k}.$$

Comparing this sequence with b'_n, it again follows that if $b_1 = b'_1$, then $b_n \leq b'_n$, where

$$n(n-1)b'_n = \frac{\lambda + a}{\lambda} \sum_{k=1}^{n-1} k(n-k) b'_k b'_{n-k}.$$

[2] As in [17], the choice of the negative root yields a global solution which converges to the spatially homogeneous solution as $t \to +\infty$. Notice also that if $\mu < a$, both roots have negative real part and the constructed solution must not only be global, it must converge to the spatially homogeneous solution, an observation consistent with the results of Nagai and Nakaki [18]. When $\mu = a$, $\lambda = 0$ and the constructed solution will be global but will not converge to the spatially homogenous solution.

This recurrence relation can be solved explicitly with $b'_n = \frac{1}{n}\left(\frac{\lambda+a}{\lambda}\right)^{n-1}(b'_1)^n$. For $b'_1 = \frac{\lambda}{\lambda+a}\delta$, we have $b'_n = \frac{\lambda}{\lambda+a}\frac{\delta^n}{n}$. The sequence $\{b'_n\}$ defines a convergent series of the form

$$\begin{aligned}\overline{\psi}(x,t) &= \sum_{n=1}^{\infty} b'_n e^{n\lambda t}\cos(2\pi Mn\xi) = \frac{\lambda}{\lambda+a}\sum_{n=1}^{\infty}\frac{1}{n}\delta^n e^{n\lambda t}\cos(2\pi Mn\xi)\\ &= -\frac{\lambda}{\lambda+a}\ln[1-2\delta e^{\lambda t}\cos(2\pi M\xi)+\delta^2 e^{2\lambda t}]\end{aligned}$$

for $t < -\frac{\ln\delta}{\lambda}$. Consequently, the upper bound for ψ in L^2 holds by the comparison of the coefficients a_n and b'_n of series for ψ and $\overline{\psi}$. A similar norm estimate holds for ψ_t, $\bar{\psi}_t$. Thus ψ exists for all $t < T_*$. That is, the existence interval $[0, T_e] \supset [0, T_*)$ or $T_e \geq T_*$. □

We now turn to the proof of Theorem 2.

Proof. To obtain the lower bound, note that if $a_1 \geq c_1$, then $a_n \geq c_n$, where the c_n satisfy

$$2(n-a_*)(n-1)c_n = \sum_{k=1}^{n-1} k(n-k)c_k c_{n-k}.$$

Thus

$$2n(n-1)c_n \geq \sum_{k=1}^{n-1} k(n-k)c_k c_{n-k} + 2a_*(n-1)c_n \geq \sum_{k=1}^{n-1} k(n-k)c_k c_{n-k}.$$

Hence if $c_1 \geq c'_1 > 0$, then $c_n \geq c'_n$, where

$$2n(n-1)c'_n = \sum_{k=1}^{n-1} k(n-k)c'_k c'_{n-k}.$$

However, $c'_n = \frac{1}{n2^{n-1}}(c'_1)^n$. Setting $c'_1 = 2\varepsilon$, it follows that $a_n \geq c'_n = 2\frac{\varepsilon^n}{n}$.

The function

$$\begin{aligned}\underline{\psi}(x,t) &= 2\sum_{n=1}^{\infty}\frac{1}{n}\varepsilon^n e^{n\lambda t}\cos(2\pi Mn\xi)\\ &= -\ln[1-2\varepsilon e^{\lambda t}\cos(2\pi M\xi)+\varepsilon^2 e^{2\lambda t}]\end{aligned}$$

exists as long as $t < -\frac{\ln\varepsilon}{\lambda}$ because the series converges absolutely and uniformly. The function $\underline{\psi}$ blows up pointwise at those points for which $\cos(2\pi M\xi) = 1$. Although we cannot compare these functions pointwise, we can compare them in H^1. To see this, note that from Parseval's identity, it follows that $|\underline{\psi_x}|_{L^2}(t)$ blows up in finite time.

Again, from Parseval and the inequalities $a_n \geq 2\frac{\varepsilon^n}{n}$, we have

$$|\underline{\psi_x}|_{L^2(0,1)}(t) \leq |\psi_x|_{L^2(0,1)}(t)$$

and

$$|\underline{\psi_t}|_{L^2(0,1)}(t) \leq |\psi_t|_{L_2(0,1)}(t).$$

Thus, ψ "blows up" at $T_\infty \le -\frac{\ln \varepsilon}{\lambda}$ in the sense that the functions $t \to |\psi_x|_{L^2}(t)$ and $t \to |\psi_t|_{L^2}(t)$ cannot be locally bounded on $[0,\infty)$ and hence (u,v) cannot remain in $H^1(0,1) \times L^2(0,1)$ for all time.

We next require that

$$2\varepsilon \le a_1 \le \frac{\lambda}{\lambda + a}\delta$$

so that

$$\varepsilon \le \frac{\lambda}{2(\lambda + a)}\delta. \tag{6.5}$$

Then the blow-up time must satisfy $-\frac{\ln \delta}{\lambda} \le T_\infty \le -\frac{\ln \varepsilon}{\lambda}$.

To prove the last claim of the theorem, note that since $u = \mu + \psi_t + a\psi$, it follows yet again from Parseval that

$$|u(\cdot,t) - \mu|_{L^2(0,1)} = \sum_1^\infty |a_n|^2(1+n^2)e^{2n\lambda t} \ge 4\sum_1^\infty \varepsilon^{2n}(1+n^{-2})e^{2n\lambda t}.$$

Consequently u must leave L^2 in finite time. Notice that this blow-up time is at least as large as the time of escape from H^1.

Finally, note that $\psi(x,0)$ and $\psi_t(x,0)$ are uniformly bounded above by $[2\lambda/(\lambda + a)]\ln(1-\delta)$ and $[2\lambda/(\lambda+a)](\lambda\delta(2+\lambda\delta)/(1-\delta))$, respectively. Hence for sufficiently small δ the initial values for the perturbed solution are positive and uniformly close to those for the spatially homogeneous solution. □

COROLLARY 1. *If $a_n \ge 2\frac{\varepsilon^n}{n}$, the function $(\psi(\cdot,t), \psi_t(\cdot,t))$ in the theorem also blows up in finite time in the sense that the sequence $\{(g_n(t), g_n'(t))\}_{n=1}^\infty$ with $g_n(0) = a_n$ and $g_n'(0) = na_n\lambda$ must leave $\ell_1^1 \times \ell^1$ in finite time.*

Since $|ng_n(t)| + |g_n'(t)| \ge 4\varepsilon^n e^{n\lambda t}$, the result follows.

Remark 1. We give an argument in the next section that shows that in every neighborhood of the spatially homogeneous solution, there are solutions of Nagai's problem, which, if they agree initially with solutions of the approximate problem and are sufficiently regular, cannot be global. The key to this argument is the demonstration that one may neglect the terms which we have identified as "tail ends."

To explain why this might be reasonable, if we evaluate $(T_n\mathcal{M}g, g'), (\mathcal{M}g, T_ng'), na(g, T_ng)$ for $g_n(t) = \beta^n e^{n\lambda t}$ for $\beta \in (0,1)$ and $\lambda > 0$, one has $g_n' = n\lambda g_n$ so that the sum of the three neglected terms is bounded above by a constant multiple of $\beta^{n+2}e^{\lambda(n+2)t}/(1-\beta^2e^{2\lambda t})$, while the sum of the convolution terms behaves like $n\beta^n e^{n\lambda t}$. Therefore on any compact subinterval of $[0, -\ln\beta/\lambda)$ the terms involving T_n are small in comparison to the convolution terms for all sufficiently large n. Elementary calculations with series for which $g_n(t),\ g_n'(t) \approx A/n^{(2+\delta)}$ show that the terms involving the T_n are not necessarily small when compared with the convolution terms.

A loose interpretation of this is the following: The partial differential equation $\psi_{tt} + (\mu - a)\psi_{xx} = (\psi_{tx} - \psi_t\psi_x)_x - a(\psi_t + (\psi\psi_x)_x)$ can be written in the form

$$\psi_{tt} + [a\psi - \psi_{xx}]_t + a(\psi\psi_x)_x + (\psi_t\psi_x)_x + (\mu - a)\psi_{xx} = 0,$$

which can be viewed as a quasi-linear second order partial differential equation with a strong damping term $(a\psi - \psi_{xx})_t$. Suppose that $\mu > a$. Linearizing this equation about $\psi = 0$ yields $\psi_{tt} + (a\psi - \psi_{xx})_t + (\mu - a)\psi_{xx} = 0$, an equation which is of

elliptic type in the second derivative terms. Without the damping term, the solutions are very regular, but the initial-boundary value problem is highly unstable. Even with the damping term, solutions of the linear equation can blow up in infinite time. The introduction of the nonlinear terms $a(\psi\psi_x)_x + (\psi_t\psi_x)_x$ can lend a hyperbolic character to the problem and force blowup in finite time by a focusing effect. See [17] for a discussion of this when $a = 0$.

7. Nagai's conjecture. From the local existence and uniqueness theorem we know that the cosine series for ψ, ψ_t satisfies the condition that

$$(\|\mathcal{M}h(s)\|_{\ell^1} + \|h'(s)\|_{\ell^1}) \tag{7.1}$$

is uniformly bounded on $[0, \tau]$ for all τ in the existence interval of the solution.

We also know that in every neighborhood of the homogeneous initial data, there is a solution of the approximate problem (4.1) with spatially nonconstant initial data for which the solution blows up in $\ell^1_1 \times \ell^1$ in finite time.

We establish the following theorem.

THEOREM 3 (Nagai's conjecture in $\ell^1_1 \times \ell^1$). *Suppose $\mu > a$. Then the corresponding solution of the Nagai–Nakaki problem, for which the cosine coefficients agree initially with the cosine coefficients of the aforementioned approximate problem, cannot be global. That is, Nagai's conjecture (in our sense) holds; i.e., such spatially inhomogeneous solutions become unstable by blowing up in finite time in $\ell^1_1 \times \ell^1$.*

Proof. Suppose that $h(t) \equiv \{h_n(t)\}_{n=1}^{\infty}$ satisfies $h(0) = a_n$, $h'(0) = n\lambda a_n$ and the system of ordinary differential equations (3.6) on some interval, say, $[0, T_{\max})$. Let $g(t) \equiv \{g_n(t)\}_{n=1}^{\infty}$ satisfy $g(0) = a_n$, $g'(0) = n\lambda a_n$ and satisfy (4.1) on $[0, T^*)$. Then $T_{\max} \le T^*$, and the solution of Nagai's problem must blow up in finite time $T_{\max}$ in $\ell^1_1 \times \ell^1$.

Suppose we could show that, on any time interval $[0, \tau)$ where $\{h_n(t)\}_{n=1}^{\infty}$ exists in the sense of the local existence theorem,

$$\sup_{[0,\tau]} \|\mathcal{M}(h(t) - g(t))\|_{\ell^1} + \sup_{[0,\tau]} \|h'(t) - g'(t)\|_{\ell^1} < C(h, \tau), \tag{7.2}$$

where $C(h, \tau)$ does not depend on g. Then this inequality, together with Theorem 2 and the triangle inequality, would lead to a lower bound for the $\ell^1_1 \times \ell^1$ norm of $\{h_n(t)\}_{n=1}^{\infty}$ in terms of the corresponding norm for $\{g_n(t)\}_{n=1}^{\infty}$ and thus would permit the establishment of Nagai's conjecture in the space $\ell^1_1 \times \ell^1$.

To this end, suppose that $\tau < T^* < T_{\max}$. We need to estimate $w_n(t) = h_n(t) - g_n(t)$ in the same fashion that we did in the proof of local existence and uniqueness where once again, $w_n(0) = w_n'(0) = 0$. Define, for any sequence $\{z_n(t)\}$,

$$\begin{aligned} \mathcal{G}_n(z, z') &= \frac{1}{2} C^2 n\{(\mathcal{M}z * z')_n + n\frac{a}{2}(z * z)_n\}, \\ \mathcal{H}_n(z, z') &= \frac{1}{2} C^2 n\{[(T_n\mathcal{M}z, z') - (\mathcal{M}z, T_n z')] + an(z, T_n z)\}. \end{aligned} \tag{7.3}$$

Then

$$
\begin{aligned}
\mathfrak{L}_n w_n &= \mathcal{G}_n(h,h') - \mathcal{G}_n(g,g') + \mathcal{H}_n(h,h') \\
&= \mathcal{G}_n(h,h') - \mathcal{G}_n(w+h,(w+h)') + \mathcal{H}_n(h,h') \\
&= \frac{1}{2}C^2 n\{-(\mathcal{M}w * h')_n - (\mathcal{M}h * w')_n - a[(\mathcal{M}w * h)_n + (\mathcal{M}h * w)_n]\} \\
&\quad + \mathcal{H}_n(h,h') \\
&\equiv \mathcal{K}_n(w,w';h,h') + \mathcal{H}_n(h,h') \\
&\equiv \mathcal{F}_n(w,w';h,h').
\end{aligned} \tag{7.4}
$$

This is the value of $\mathcal{L}_n$ that replaces the right-hand side of (5.2), (5.4)–(5.8). Notice that the terms in the definition of $\mathcal{K}_n$ can be estimated as in the local existence and uniqueness theorem. That is,

$$|\mathcal{K}_n(w,w';h,h')(s)| \le (\|\mathcal{M}w(s)\|_{\ell^1} + \|w'(s)\|_{\ell^1})(\|\mathcal{M}h(s)\|_{\ell^1} + \|h'(s)\|_{\ell^1}).$$

Likewise, we have

$$\sum_{n=1}^{\infty} \frac{|\mathcal{H}_n(h,h')|(s)}{n} \le \max\{a,1\} \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} [|h_k'(s)|j|h_j(s)| + |h_k(s)|j|h_j'(s)| + |h_k(s)|j|h_j(s)|].$$

For such functions, the conservation conditions (2.4) hold. Consequently,

$$\sum_{n=1}^{\infty} \frac{|\mathcal{H}_n(h,h')|(s)}{n} \le \max\{a,1\}(\|h(s)\|_{\ell^1} + \|h'(s)\|_{\ell^1})\|\mathcal{M}h(s)\|_{\ell^1}$$

by using estimates similar to those used for the estimates on the tail-end terms in the proof of uniqueness.

Thus we obtain an inequality of the form

$$
\begin{aligned}
&\|\mathcal{M}w(t)\|_{\ell^1} + \|w'(t)\|_{\ell^1} \\
&\le \int_0^t \frac{A(\|\mathcal{M}w(s)\|_{\ell^1} + \|w'(s)\|_{\ell^1}) + \sum_{n=1}^{\infty} |\mathcal{H}_n(h,h')|(s)/n}{\sqrt{t-s}}\, ds \\
&\le \int_0^t \frac{A(\|\mathcal{M}w(s)\|_{\ell^1} + \|w'(s)\|_{\ell^1}) + B(\|\mathcal{M}h(s)\|_{\ell^1} + \|h'(s)\|_{\ell^1})\|\mathcal{M}h(s)\|_{\ell^1}}{\sqrt{t-s}}\, ds
\end{aligned} \tag{7.5}
$$

for some constant A depending on $\tau, \|\mathcal{M}h\|_{\ell^1}, \|h'\|_{\ell^1}$ and for some constant B depending perhaps on τ but not on w, w', h, h'.

This inequality, (7.1), and an application of the Gronwall inequality will give us the estimate (7.2). Combining this observation with its consequence (7.2) and the triangle inequality gives the result. □

Remark 2. The injection $I : \ell^1 \to L^1(0,1)$ given by

$$I(\{a_n\}_{n=1}^{\infty})(x) = \sum_{n=1}^{\infty} a_n \cos{(n-1)\pi x}$$

is certainly continuous. However, the inverse is not. To see this, let

$$P(x,\epsilon) = \frac{1}{2}\frac{1-\epsilon^2}{1+\epsilon^2-2\epsilon\cos(\pi x)}$$

denote the Poisson kernel for $0 \leq \epsilon < 1$. The Poisson kernel is nonnegative, satisfies

$$\int_0^1 P(x,\epsilon)\,dx = 1 = \|P(\cdot,\epsilon)\|_{L^1},$$

and has Fourier cosine series

$$\sum_{n=1}^{\infty} a_n \cos{(n-1)\pi x} = \frac{1}{2} + \sum_{n=1}^{\infty} \epsilon^n \cos n\pi x$$

whose coefficient sequence satisfies

$$\|\{a_n\}_{n=1}^{\infty}\|_{\ell^1} = \frac{1+\epsilon}{1-\epsilon}.$$

Therefore, as ϵ increases to unity, the ℓ^1 norm of the coefficient sequence increases without bound, while the L_1 norms of $P(\cdot,\epsilon)$ remain bounded. (Indeed, they converge in measure to Dirac measure).

If $\mu \geq 0$, then we know from the first equation of the system that $\int_0^1 u(x,t)\,dx = \int_0^1 \mu(x)\,dx$ and hence must remain in $L^1(0,1)$ on the existence interval. The second equation tells us that the second component must likewise remain bounded in $L^1(0,1)$.

The remark tells us that it is possible for the solution to blow up in sequence space in finite time but remain bounded in the $H^1 \times L^1$ norm. If one knew that the solution components blow up in finite time in $\ell^2_\beta \times \ell^2_{\beta'}$, for large enough β, β', then Parseval's identity would tell us that the solution would blow up in $H^{\beta'}(0,1) \times H^{\beta'}(0,1)$. Thus we have the following corollary.

COROLLARY 2. *Let $\delta, \delta' > 0$. Suppose a solution of Nagai's problem blows up in finite time T in $\ell^1_1 \times \ell^1$. If the solution components belong to $H^{3/2+\delta}(0,1) \times H^{1/2+\delta'}(0,1)$ and are bounded on compact subsets of the existence interval in the norm of this product space, then the solution blows up in $H^{3/2+\delta}(0,1) \times H^{1/2+\delta'}(0,1)$ in finite time no larger than T.*

Proof. This result follows from Schwarz's inequality and Parseval's identity. ∎

Remark 3. The corollary states that certain very smooth solutions of Nagai's problem cannot be global. That is, they must lose regularity in finite time.

We can use the results of [18] to establish Nagai's conjecture.

COROLLARY 3. *Suppose $\mu > a$. Then in every neighborhood of the stationary solution, there are solutions which blow up in finite time in the sense that the $H^2 \times H^1$ norm blows up in finite time.*

Proof. In [18, Proposition 4.1] the authors prove that if the initial data for u, v are sufficiently smooth (in particular, if they are analytic) and satisfy the boundary conditions, then both components are continuous from $[0, T_{exist})$ into $H^2(0,1)$, while the first (corresponding to v in the notation of [18]) is continuously differentiable from $[0, T_{exist})$ into $H^2(0,1)$.

The initial data for the approximate problem which give a solution of the approximate problem that blows up in finite time are in fact analytic (the Fourier coefficients are bounded above by $Cn\epsilon^n$ for small ϵ) and consequently must satisfy the smoothness criteria of the initial data needed for [18, Proposition 4.1]. If we take the same initial values for solution of Nagai's problem, its components must belong to the same spaces.

Thus, by the preceding corollary, the solution cannot be global. This, together with the preceding corollary and $\delta = \delta' = 1/2$, establishes Nagai's conjecture for certain sufficiently smooth data in every neighborhood of the stationary data. ∎

Remark 4. These results say nothing about the pointwise finite time blowup for the solution components.

COROLLARY 4. *Suppose* $\mu > a$ *and* $0 < \varepsilon \le a_1 \le \frac{\lambda}{\lambda+a}\delta$, *where* λ *is given in* (6.3). *Then both components of the solution constructed in the preceding corollary blow up in* L^∞.

Proof. In [18, Theorem 7.1], the authors show that *if* the global existence time is finite, then both components blow up in L^∞. □

An illustrative computation is given in Figures 3 and 4.

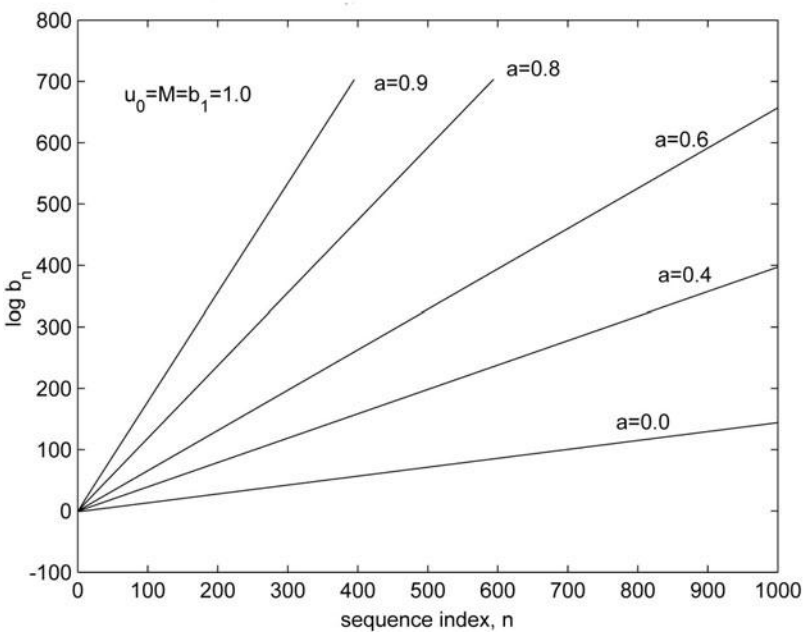
Remark 5. If we replace λ by the negative root of the quadratic it satisfies, then we can use the arguments of section 6 to obtain solutions of the approximate problem which decay uniformly and exponentially rapidly to zero. It is then possible, by appropriately modifying the arguments in Theorem 3, to show that the corresponding solution of the Nagai–Nakaki problem, *for which the cosine coefficients agree initially with the cosine coefficients of the aforementioned approximate problem*, must be global. That is, Nagai's conjecture (in our sense) holds; i.e., such spatially inhomogeneous solutions (ψ, ψ_t) exist for all time and must converge to the steady state $(\mu/a, \mu)$ in $\ell_1^1 \times \ell^1$ as $t \to \infty$. We omit the details.

8. Illustrative computations. One might well ask whether or not solutions of (6.1) are asymptotically of the form $A\epsilon^n$ for some $|\epsilon| \in (0,1)$ in the sense that there is $\epsilon \in [0,1)$ such that $\lim a_n/\epsilon^n = A$ for some constant A. We have shown that the solutions of (6.1) are bounded above and below by terms of this form but we have not yet established the asymptotics. However, the computations below provide a powerful argument for these asymptotics.

The dependence of the blow-up time on $M, \mu - a$ for the function ψ in Theorem 1 can be investigated numerically as follows. It is worth noting that $\lambda \to 0^+$ as $M \to +\infty$ for fixed $\mu - a$ or $\mu - a \to 0^+$ for fixed $M \ge 1$.

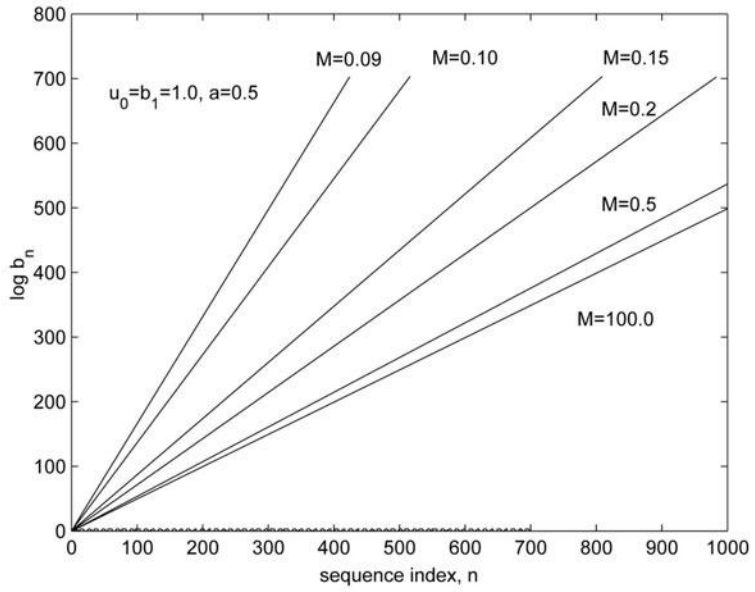
We begin by examining the growth of the terms of the sequence defined by (6.1). If we set $a_n(a) = b_n(a)\sigma^n/n$ with $b_1 = 1$ and $\sigma > 0$, it is not too hard to see that when $a = 0$, $b_n = 1$ for all $n \ge 1$. However, when $0 < a < \mu$, the terms b_n grow remarkably rapidly. As an illustrative example, with $M = \mu = 1$ and $a = 0.6$, $b_1 = 1$, $\ln b_{1000} \approx 656.9587$. In fact, numerical evidence suggests that $\ln b_n \approx 0.657616(n-1)(1+o(1/n))$. Let τ be this coefficient of $n-1$ (assuming it exists). If we take $\sigma = \exp(-\tau(1+\delta))$ for any small, positive δ, then the solution should blow up in finite time $t = \tau\delta/\lambda$. This will be the case if one can prove that the asymptotics for the b_n are as indicated by the numerical evidence. The numerical evidence indicates that as a increases to μ from below, τ increases without bound, and hence the blow-up time will increase without bound also. This is to be expected. (See Figure 1.) Likewise, the numerical evidence indicates that as the integer M is increased for fixed a, τ approaches a limiting value (which is to be expected since as M increases, $\lambda \to 0$). This corresponds to $a_* = 0$ so that the sequence behaves like the exact solution when $a = 0$. The solution for $a = 0$ has the smallest blow-up time possible for fixed ε, δ, μ and all $a \ge 0$. (See Figure 2.)

In Figures 3 and 4 we present a numerical simulation for the Nagai–Nakaki problem with $u_t = D(u_{xx} - (uv_x)_x)$, $v_t = u - av$ for $0 < x < 1, t > 0$ and zero flux boundary conditions. We took $D = 0.02$, $\mu = 4.0$, and $a = 1.0$. For initial values we used $u(x,0) = \mu - \epsilon\cos(2\pi x)$ where $\epsilon = 0.4$ and $v(x,0) = v_0 = \mu/a$. These figures provide some evidence that the solution does blow up pointwise in finite time as well as in the $\ell_1^1 \times \ell^1$ norm.



(The lines corresponding to a = 0.8, 0.9 do not continue due to exponential overflow.)

FIG. 1. *Growth of coefficients given by* (6.1) *for variable a.*



(The lines corresponding to M = 0.09, 0.1, 0.15, 0.2 do not continue due to exponential overflow.)

FIG. 2. *Growth of coefficients given by* (6.1) *for variable M.*

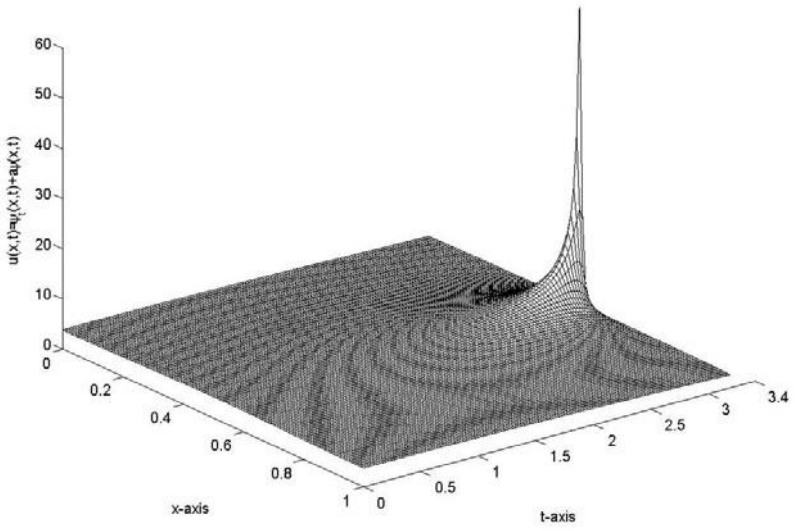


FIG. 3. *Partial density,* $a = 1.0, D = 0.02, \varepsilon = 0.25, \mu = 4.0$.

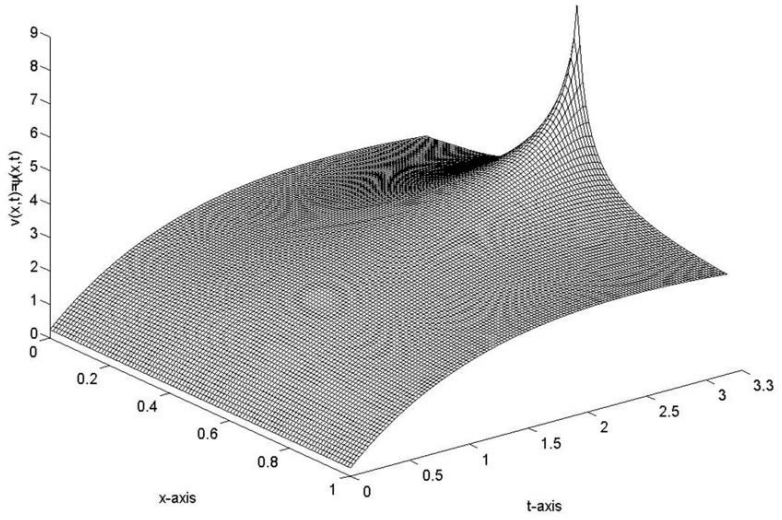


FIG. 4. *Chemical density,* $a = 1.0, D = 0.02, \varepsilon = 0.25, \mu = 4.0$.

Acknowledgments. The authors take pleasure in thanking Thomas Hillen, Peter Palacik, and Hans Weinberger and the referees for their constructive comments which considerably improved the earlier versions of this manuscript. The authors also take pleasure in thanking Toshitaka Nagai for sending us [18], for bringing to our attention Proposition 4.1 of that paper, and for clarifying a point that led to Corollary 4.

REFERENCES

[1] P. Biler, *Local and global solutions of a nonlinear nonlocal parabolic problem*, in Nonlinear Analysis and Applications (Warsaw, 1994), GAKUTO Internat. Ser. Math. Sci. Appl. 7, Gakkōtosho, Tokyo, 1996, pp. 49–66.

[2] P. Biler, *Global solutions to some parabolic-elliptic systems of chemotaxis*, Adv. Math. Sci. Appl., 9 (1999), pp. 347–359.

[3] S. Childress and J. K. Percus, *Nonlinear aspects of chemotaxis*, Math. Biosci., 56 (1981), pp. 217–237.

[4] B. Davis, *Reinforced random walk*, Probab. Theory Related Fields, 84 (1990), pp. 203–229.

[5] J. I. Diaz, T. Nagai, and J.-M. Rakotoson, *Symmetrization techniques on unbounded domains: Application to a chemotaxis system on* $\mathbf{R}^N$, J. Differential Equations, 145 (1998), pp. 156–183.

[6] L. Edelstein-Keshet, *Mathematical Models in Biology*, The Random House/Birkhäuser Mathematics Series, Random House Inc., New York, 1988.

[7] M. A. Fontelos, A. Friedman, and B. Hu, *Mathematical analysis of a model for the initiation of angiogenesis*, SIAM J. Math. Anal., 33 (2002), pp. 1330–1355.

[8] M. A. Herrero, E. Medina, and J. J. L. Velázquez, *Self-similar blow-up for a reaction-diffusion system*, J. Comput. Appl. Math., 97 (1998), pp. 99–119.

[9] M. A. Herrero and J. J. L. Velázquez, *Chemotactic collapse for the Keller-Segel model*, J. Math. Biol., 35 (1996), pp. 177–194.

[10] M. A. Herrero and J. J. L. Velázquez, *Singularity patterns in a chemotaxis model*, Math. Ann., 306 (1996), pp. 583–623.

[11] T. Hillen and H. A. Levine, *Blow-up and pattern formation in hyperbolic models for chemotaxis in* 1-D, Z. Angew. Math. Phys., 54 (2003), pp. 839–868.

[12] T. Hillen and A. Potapov, *The one-dimensional chemotaxis model: Global existence and asymptotic profile*, MMAS, to appear.

[13] W. Jäger and S. Luckhaus, *On explosions of solutions to a system of partial differential equations modelling chemotaxis*, Trans. Amer. Math. Soc., 329 (1992), pp. 819–824.

[14] E. F. Keller, *Assessing the Keller-Segel model: How has it fared?*, in Biological Growth and Spread (Heidelberg, 1979), Lecture Notes in Biomath. 38, Springer, Berlin, 1980, pp. 379–387.

[15] E. F. Keller and L. A. Segel, *Initiation of slime mold aggregation viewed as an instability*, J. Theoret. Biol., 26 (1970), pp. 399–415.

[16] E. F. Keller and L. A. Segel, *Model for chemotaxis*, J. Theoret. Biol., 30 (1971), pp. 225–234.

[17] H. A. Levine and B. D. Sleeman, *A system of reaction diffusion equations arising in the theory of reinforced random walks*, SIAM J. Appl. Math., 57 (1997), pp. 683–730.

[18] T. Nagai and T. Nakaki, *Stability of constant steady states and existence of unbounded solutions in time to a reaction-diffusion equation modelling chemotaxis*, Nonlinear Anal., 58 (2004), pp. 657–681.

[19] T. Nagai and T. Senba, *Global existence and blow-up of radial solutions to a parabolic-elliptic system of chemotaxis*, Adv. Math. Sci. Appl., 8 (1998), pp. 145–156.

[20] K. Osaki and A. Yagi, *Finite dimensional attractor for one-dimensional Keller-Segel equations*, Funkcial. Ekvac., 44 (2001), pp. 441–469.

[21] H. G. Othmer and A. Stevens, *Aggregation, blowup, and collapse: The ABC's of taxis in reinforced random walks*, SIAM J. Appl. Math., 57 (1997), pp. 1044–1081.

[22] C. S. Patlak, *Random walk with persistence and external bias*, Bull. Math. Biophys., 15 (1953), pp. 311–338.

[23] Y. Yang, H. Chen, and W. Liu, *On existence of global solutions and blow-up to a system of reaction-diffusion equations modelling chemotaxis*, SIAM J. Math. Anal., 33 (2001), pp. 763–785.

SIAM J. APPL. MATH.
Vol. 65, No. 2, pp. 361–391

FROM INDIVIDUAL TO COLLECTIVE BEHAVIOR IN BACTERIAL CHEMOTAXIS*

RADEK ERBAN† AND HANS G. OTHMER†

Abstract. Bacterial chemotaxis is widely studied from both the microscopic (cell) and macroscopic (population) points of view, and here we connect these very different levels of description by deriving the classical macroscopic description for chemotaxis from a microscopic model of the behavior of individual cells. The analysis is based on the velocity jump process for describing the motion of individuals such as bacteria, wherein each individual carries an internal state that evolves according to a system of ordinary differential equations forced by a time- and/or space-dependent external signal. In the problem treated here the turning rate of individuals is a functional of the internal state, which in turn depends on the external signal. Using moment closure techniques in one space dimension, we derive and analyze a macroscopic system of hyperbolic differential equations describing this velocity jump process. Using a hyperbolic scaling of space and time, we obtain a single second-order hyperbolic equation for the population density, and using a parabolic scaling, we obtain the classical chemotaxis equation, wherein the chemotactic sensitivity is now a known function of parameters of the internal dynamics. Numerical simulations show that the solutions of the macroscopic equations agree very well with the results of Monte Carlo simulations of individual movement.

Key words. chemotaxis equations, velocity-jump process, internal dynamics, transport equations, aggregation, bacterial chemotaxis

AMS subject classifications. 35Q80, 92B05, 92D25, 60J75

DOI. 10.1137/S0036139903433232

1. Introduction.

The ability to detect and respond to changes in the environment is a basic necessity for survival of all organisms, and as a result, a variety of mechanisms have evolved by which organisms sense their environment and respond to signals they detect. Often the response involves movement toward a more favorable environment or away from a noxious substance. The movement response can entail changing the speed of movement and the frequency of turning, which is called *kinesis*; it may involve directed movement, which is called *taxis*; or it may involve a combination of these. Taxes and kineses may be characterized as positive or negative, depending on whether they lead to accumulation at high or low points of the external stimulus that triggers the motion. A variety of both modes are known, and include responses to gradients of oxygen and other chemicals, gradients of adhesion to the substrate, and other effects. Both tactic and kinetic responses involve two major steps: (i) detection of the signal and (ii) transduction of the external signal into an internal signal that triggers the response. From the modeling and analysis standpoint, an important characteristic of both modes of response is whether or not the individual merely detects the signal or alters it as well, for example by amplifying it so as to relay the signal. When there is no significant alteration, the individual simply responds to the spatio-temporal distribution of the signal. However, when the individual produces

*Received by the editors August 12, 2003; accepted for publication (in revised form) June 9, 2004; published electronically December 16, 2004. The research of the first author was supported in part by NSF grant DMS 0317372. The research of the second author was supported in part by NIH grant GM 29123, NSF grant DMS 9805494, and NSF grant DMS 0317372. Both authors were supported in part by the Minnesota Supercomputing Institute.
http://www.siam.org/journals/siap/65-2/43323.html
†School of Mathematics, 270B Vincent Hall, University of Minnesota, Minneapolis, MN 55455 (erban@math.umn.edu, othmer@math.umn.edu).

or degrades the signal, there is coupling between the local density of individuals and the intensity of the signal. This occurs, for example, when individuals aggregate in response to a signal from "organizers" and relay the signal as well.

In several systems, including the flagellated bacterium *Escherichia coli* and the amoeboid cell *Dictyostelium discoideum*, a detailed understanding of how extracellular signals are transduced into behavioral changes is emerging from experimental work, while at the macroscopic level a great deal is known about solutions of the classical chemotaxis equations. However, the chemotaxis equations to date have been based on phenomenological descriptions of how cells respond to signals, and at present there is little understanding of how microscopic properties translate into the macroscopic parameters. The motion of *E. coli* has been studied for forty years, and much is known about how they sense and process environmental signals. *E. coli* alternates two basic behavioral modes, a more or less linear motion, called a run, and a highly erratic motion, called tumbling, the purpose of which is to reorient the cell. During a run the bacteria move at approximately constant speed in the most recently chosen direction. Run times are typically much longer than the time spent tumbling, and when bacteria move in a favorable direction (i.e., either in the direction of foodstuffs or away from harmful substances) the run times are increased further. These bacteria are too small to detect spatial differences in the concentration of an attractant on the scale of a cell length, and during a tumble they simply choose a new direction essentially at random, although it has some bias in the direction of the preceding run [7, 4]. The effect of alternating these two modes of behavior, and in particular, of increasing the run length when moving in a favorable direction, is that a bacterium executes a three-dimensional (3D) random walk with drift in a favorable direction when observed on a sufficiently long time scale [4, 25, 5]. Models for signal transduction and adaptation in this system are given in [40, 2, 28].

In the absence of external cues, many organisms use a random walk strategy to determine their pattern of movement. In this case the movement of organisms released at a point in a uniform environment can be described as an uncorrelated, unbiased random walk of noninteracting particles on a sufficiently long time scale. In an appropriate continuum limit the cell density n, measured in units of cells/L^N, where L denotes length and $N = 1, 2$, or 3, satisfies the diffusion equation

$$\frac{\partial n}{\partial t} = D\Delta n. \tag{1.1}$$

Here the cell flux is given by $j = -D\nabla n$, and the simplest description of cell motion in the presence of an attractant or repellent is obtained by adding a directed component to the diffusive flux to obtain

$$j = -D\nabla n + nu_c, \tag{1.2}$$

where u_c is the macroscopic chemotactic velocity. The taxis is positive or negative according to whether u_c is parallel or antiparallel to the direction of increase of the chemotactic substance. The resulting evolution equation for n is

$$\frac{\partial n}{\partial t} = \nabla \cdot (D\nabla n - nu_c), \tag{1.3}$$

and this is called a chemotaxis equation. In a phenomenological approach one postulates a constitutive relation for the chemotactic velocity of the form

$$u_c = \chi(S)\,\nabla S, \tag{1.4}$$

where S is the concentration of the chemotactic substance and the function $\chi(S)$ is called the chemotactic sensitivity. When $\chi > 0$, the tactic component of the flux is in the direction of ∇S and the taxis is positive. With this postulate, (1.3) takes the form

$$\frac{\partial n}{\partial t} = \nabla \cdot (D\nabla n - n\,\chi(S)\,\nabla S). \tag{1.5}$$

We call equations of this type classical chemotaxis equations, though frequently that term is used for a system of equations comprising (1.5) and a reaction-diffusion equation for the evolution of the signal substance. A recent review of the mathematical aspects of chemotaxis equations is given in [21].

A problem in using equations such as (1.5) to describe chemotaxis is how one justifies the constitutive assumption (1.4) and, in particular, how one incorporates microscopic responses of individual cells into the chemotactic sensitivity. A number of phenomenological approaches to the derivation of the chemotactic sensitivity or chemotactic velocity have been taken, including simply postulating the form in (1.4) [23, 33] or deriving the velocity directly in terms of forces exerted by the cell [35]. Other more fundamental approaches have also been used to relate the chemotactic velocity or sensitivity to a microscopic description of movement. In the first, one begins with a lattice walk or space jump process, either in discrete or continuous time, and postulates how the transition probabilities depend on the external signal. For a discrete time walk the chemotaxis equation is derived in the diffusion limit of this process, by letting the space step size h and the time step δt go to zero in such a way that the ratio $h^2/\delta t$ is a constant, namely D. A more general approach leads to a renewal equation, from which a partial differential equation is obtained by particular choices of the jump kernel and the waiting time distribution [29]. Another method, based on a continuous time reinforced random walk in which the walker modifies the transition probabilities of an interval for successive crossings, is developed in [31] for a single tactic substance.

A space jump process is suitable for certain organisms, but an alternative stochastic process that may be more appropriate for describing the motion of cells is called the velocity jump process [29]. In this process the velocity, rather than the spatial position, changes by random jumps at random instants of time. The governing evolution equation for the simplest version of this process is

$$\frac{\partial}{\partial t}p(x,v,t) + v \cdot \nabla p(x,v,t) = -\lambda p(x,v,t) + \lambda \int_V T(v,v')p(x,v',t)dv', \tag{1.6}$$

where $p(x,v,t)$ denotes the density of particles at spatial position $x \in \Omega \subset R^N$, moving with velocity $v \in V \subset R^N$ at time $t \geq 0$ [29]. Here λ is the (constant) turning rate, and $1/\lambda$ is a measure of the mean run length between velocity jumps. In general, the turning frequency λ must depend on the extracellular signal, as transduced through the signal transduction network and the motility control system. The turning kernel $T(v,v')$ gives the probability of a velocity jump from v' to v if a jump occurs, and implicit in the above formulation is the assumption that the choice of a new velocity is independent of the run length.

The forward equation (1.6) for a velocity jump process is similar to the Boltzmann equation, wherein the right-hand side is an integral operator that describes the collision of two particles, and is therefore quadratic in p [11]. The kernel of the integral operator is specified by the dynamics, and it is well known that an appropriate

scaling of space and time leads at least formally from the Boltzmann equation to a diffusion process [26, 17]. This also holds for transport equations and more general transport processes (see, e.g., [18, 34, 36]). The earliest derivation of the chemotactic sensitivity from a velocity jump process was done by Patlak [36], who used kinetic theory arguments to express u_c in terms of averages of the velocities and run times of individual cells. Alt [1] significantly extended Patlak's approach to the analysis of taxis and his results have been applied to *E. coli* using a phenomenological description of signal transduction [12].

In [19, 30] the kinetic equation approach for deriving chemotactic equations was further developed using a kernel T that may include an external bias. A general Perron–Frobenius property of the turning operator $\mathcal{T}$ defined by the right-hand side of (1.6) and a proper scaling of space and time lead to a Hilbert expansion of the long-term dynamics that produces a parabolic limiting equation. In certain cases there is no taxis, and the parabolic limit is anisotropic, in that the resulting equation for the macroscopic density,

$$n(x,t) = \int_V p(x,v,t)\mathrm{d}v, \tag{1.7}$$

is

$$\frac{\partial n}{\partial t} = \nabla \cdot D\nabla n, \tag{1.8}$$

where D is an $N \times N$ nondiagonal matrix. Necessary and sufficient conditions under which the diffusion matrix D reduces to a scalar times the identity were also obtained. In previous work the external bias enters the turning kernel and turning rate as an order ε term [1, 30], and the perturbation analysis done in [30] shows that the chemotaxis equation is obtained only in this case. In the approach used in [30] an external bias of order one in the turning kernel can be admitted, but with suitable restrictions this leads to (1.8) rather than the chemotaxis equation in the diffusion limit.

The prototypical organisms whose motion can be described as a velocity jump process are the flagellated bacteria such as *E. coli*. A bacterium runs at a constant velocity for a random length of time, then tumbles for a random length of time, chooses a new direction at random, and repeats the cycle. When motion is restricted to one space dimension and the tumble phase is neglected, this leads to a telegraph process described by the hyperbolic system

$$\begin{aligned} \frac{\partial p^+}{\partial t} + s\frac{\partial p^+}{\partial x} &= -\lambda p^+ + \lambda p^-, \\ \frac{\partial p^-}{\partial t} - s\frac{\partial p^-}{\partial x} &= \lambda p^+ - \lambda p^-, \end{aligned} \tag{1.9}$$

where $p^{\pm}(x,t)$ are the probabilities densities of particles that are at (x,t) and are moving to the right $(+)$ and left $(-)$ and s is the speed. This model was first analyzed by Goldstein [16], and subsequently by others [22, 27, 29]. It can be shown that if λ is a constant, the system reduces to a damped wave equation called the telegraph equation for the total density $p \equiv p^+ + p^-$, and on a finite domain with reflecting boundary conditions, solutions are asymptotically constant in space and time. Even if there is a fixed background signal and the turning rate depends on the signal but is

independent of the direction of travel, still there is no aggregation at extrema of the signal: all solutions are asymptotically constant [32].

It is not difficult to see formally that the turning rate for left-moving particles must be different from that for right-moving particles in order to produce a nonzero chemotactic velocity [32], and this has been analyzed in detail in [20]. However, at present there is little understanding of the interplay between the intracellular dynamical system that describes signal transduction and quantities such as the turning rate and turning kernel in a macroscopic, population-level description of motion. Our objective here is to develop a mathematical framework in which one can systematically extract information about population-level behavior for adapting walkers from microscopic models of individual behavior, and to apply it to a caricature of adapting intracellular dynamics. We use *E. coli* as a prototype system, but the methodology and the results apply more generally.

The paper is organized as follows. In the following section we briefly describe the signal transduction network in *E. coli* to motivate the simplified description used herein. Next we introduce the transport equations for systems with internal dynamics, which are the starting point for the derivation of the macroscopic limit. We derive the macroscopic moment equations (section 4), the modified version of the classical chemotaxis equation (section 6), and the classical chemotaxis equation (section 7). Finally, we show some illustrative numerical results, we apply our results to experiments with an exponential signal gradient, and we discuss generalizations of our approach. The extension of the results herein to higher space dimensions is done in [15].

2. Internal dynamics. *E. coli* have 4–6 flagella distributed uniformly over the cell surface and move by rotating them in a corkscrew-like manner [37, 41]. When rotated counterclockwise, the flagella coalesce into a propulsive bundle, resulting in a relatively straight "run" [8]. When rotated clockwise they fly apart, resulting in a "tumble" which reorients the cell but causes no significant change of location. The cell thus alternates between runs and reorienting tumbles. In the absence of stimuli, the bias or probability per unit time of a tumble (P_{CW}) is essentially independent of when the last tumble occurred [41]. The mean run interval is about 1 sec in the absence of chemotaxis, the mean tumble interval is about 0.1 sec, and both are distributed exponentially [6]. A chemoeffector (attractant or repellent) alters the probabilities that the flagella will rotate in a given direction, thus changing the frequencies and duration of runs and tumbles. *E. coli* respond chemotactically to a variety of attractants and repellents over a wide range of concentrations [6]. A typical response, which we define as a measurable change in bias from baseline following a transient increase in the concentration of an attractant or a decrease in that of a repellent, is as follows. After a brief latency period there is an increase in P_{CCW} (probability per unit time of a run) above the baseline probability of approximately 0.64 [8]. This early response, which is typically rapid, constitutes the excitation, and it is followed by a period of relatively slow adaptation to the stimulus. Adaptation eventually returns the bias to baseline, allowing the cell to respond to further changes.

The magnitude of the change in bias in response to an exponentially increasing attractant concentration increases approximately linearly with the ramp rate [9]. Assuming equilibrium binding, the fraction of receptors occupied is $\theta = S/(K_D + S)$, where S is the concentration and K_D is the dissociation constant. Therefore $\dot{\theta} = K_D S/(K_D + S)^2 \cdot d\ \mathrm{ln}S/dt$, and if $S \sim K_D$, then $4\dot{\theta} \sim d\ \mathrm{ln}S/dt$, which is the ramp rate. Thus the magnitude of the response is an approximately linear function of the *rate of change* in occupancy, which provides a superficial explanation of the observed

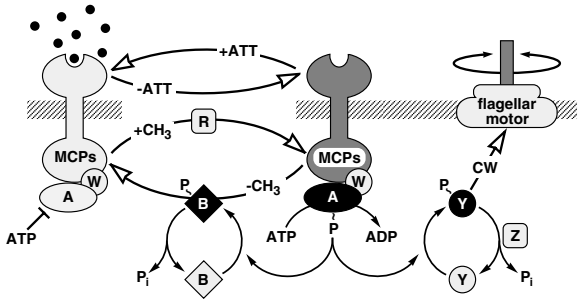


FIG. 2.1. *Signaling components and pathways for* E. coli *chemotaxis. Chemoreceptors (MCPs) span the cytoplasmic membrane (hatched lines), with a ligand-binding domain on the outside and a catalytic domain on the cytoplasmic side. MCP complexes have two alternative signaling states. In the attractant-bound form, the receptor inhibits CheA autokinase activity; in the unliganded form, the receptor stimulates CheA (A) activity. The overall flux of phosphoryl groups P to CheB (B) and CheY (Y) reflects the proportion of signaling complexes in the inhibited and stimulated states. Changes in attractant concentration shift this distribution, triggering a flagellar response. Adaptation occurs when the ensuing changes in the CheB phosphorylation state alter its methylesterase activity, producing a net change in the MCP methylation state that cancels the stimulus signal (cf.* [42] *for a review; figure reproduced from* [40]*, with permission).*

adaptation. Because of adaptation, the response is not directly dependent on the absolute concentration of chemoeffector [41], but instead the sensory system functions as a derivative detector.

E. coli is also extremely sensitive to small changes in chemoeffector levels. The cells can respond to slow exponential increases in attractant levels that correspond to rates of change in the fractional occupancy of chemoreceptors as small as 0.1% per second [9, 38]. Thus a cell can respond even when there is only a small change in the receptor occupancy over a typical sampling period. High sensitivity is also seen when cells are subjected to small impulses or step increases in attractant concentration, though the evidence is mixed. Segall, Block, and Berg [38] report that a change in receptor occupancy of 0.42% elicits a 23% change in bias—a ratio, or gain, of 55—but Khan et al. [24] report a maximum gain of only 6.

The main features of the *E. coli* chemotaxis excitation and adaptation pathways are as follows [10] (see Figure 2.1). Chemical stimuli are detected by transmembrane receptors, which in turn generate cytoplasmic signals that control the flagellar motors. Aspartate, the attractant chemoeffector most commonly used in experiments, binds directly to the periplasmic domain of its transducer, Tar. This initiates a complicated sequence of biochemical steps, the net effect of which is to temporarily reduce the level of the motor control protein CheY_P following an increase in attractant, thereby temporarily increasing $P_{CCW} = 1 - P_{CW}$ and increasing the fraction of time spent running as opposed to tumbling. Detailed models of this network are now available [40, 2, 28], and we refer the reader to the original literature. For our purposes we wish to abstract the essential features of the signal transduction and response processes.

2.1. Cartoon internal dynamics. The essential aspects that a simplified description must reproduce in order to make it useful for studying macroscopic phenomena are (i) it must exhibit excitation, which here means a change in bias in response to a stimulus, (ii) the bias must return to baseline levels (i.e., the response must adapt) on a time scale that is slow compared to excitation, and (iii) the signal transduction network should amplify signals appropriately. Let $y = (y_1, y_2, \ldots, y_m) \in \mathbb{R}^m$

denote the internal state variables, which can include the concentrations of receptors, proteins, etc., and let $S(x,t) = (S_1, S_2, \ldots, S_M) \in \mathbb{R}^M$ denote the signals in the environment. Then all current deterministic models of bacterial signal transduction pathways can be cast in the form of a system of ordinary differential equations that describe the evolution of the intracellular state, forced by the extracellular signal. Thus

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(y, S), \tag{2.1}$$

where $f : \mathbb{R}^m \times \mathbb{R}^M \to \mathbb{R}^m$ describes the particular model. The question is, given an accurate microscopic model, can we derive a macroscopic description, and can we use it to predict the effect on macroscopic behavior of changes in the microscopic parameters? At present this is very difficult to do with a full description of the internal dynamics for *E. coli*, which may involve 20 or more variables, and as a first step we use a simpler cartoon description of signal transduction that was developed in [32], which incorporates the essential features described above.

We describe the internal dynamics with two internal variables, i.e., $y \in Y \subset \mathbb{R}^2$, and we suppose that the internal state evolves according to the system of ordinary differential equations

$$\frac{\mathrm{d}y_1}{\mathrm{d}t} = \frac{g(S(x,t)) - (y_1 + y_2)}{t_e}, \tag{2.2}$$

$$\frac{\mathrm{d}y_2}{\mathrm{d}t} = \frac{g(S(x,t)) - y_2}{t_a}, \tag{2.3}$$

where t_e and t_a are constants, x is the current position of a cell, $S : \mathbb{R}^N \times [0,\infty) \to [0,\infty)$ is the concentration of the chemoattractant, and $g : [0,\infty) \to [0,\infty)$ models the first step of signal transduction. For any constant signal S these equations have the property that

$$\lim_{t\to\infty} y_1 = 0 \quad \text{and} \quad \lim_{t\to\infty} y_2 = g(S), \tag{2.4}$$

and therefore y_1 adapts perfectly to any constant stimulus. The time constants t_e and t_a are labeled in anticipation of using y_1 for the internal response and y_2 as the adaptation variable, and therefore we call t_e and t_a the excitation and adaptation time constants, respectively. In order to obtain the desired response, one must have $t_e < t_a$. In *E. coli* the excitation is much faster than adaptation, and we have $t_e \ll t_a$.

Since y_1 adapts perfectly, any continuous function $h : \mathbb{R} \to \mathbb{R}$ of y_1 can be used to model the response to changes in the extracellular signal, and the response will adapt; i.e., the steady state response will be independent of the magnitude of the stimulus S. In Figure 2.2 we compare the response of the cartoon model with the response predicted by a detailed model of the entire signal transduction pathway. It is clear that the cartoon model can capture the essential changes in the bias in *E. coli* using a suitable definition of the response. For the simplest velocity jump process in which tumbling is ignored, we identify the response with the turning frequency $\lambda(y)$. In a more detailed description in which the tumble phase is accounted for, one can relate the internal state more directly to experimental results on the switching frequency [13]. This will be done in section 9.1; here we use $\lambda(y) \equiv \mathit{Response} = h(y_1)$. Moreover,

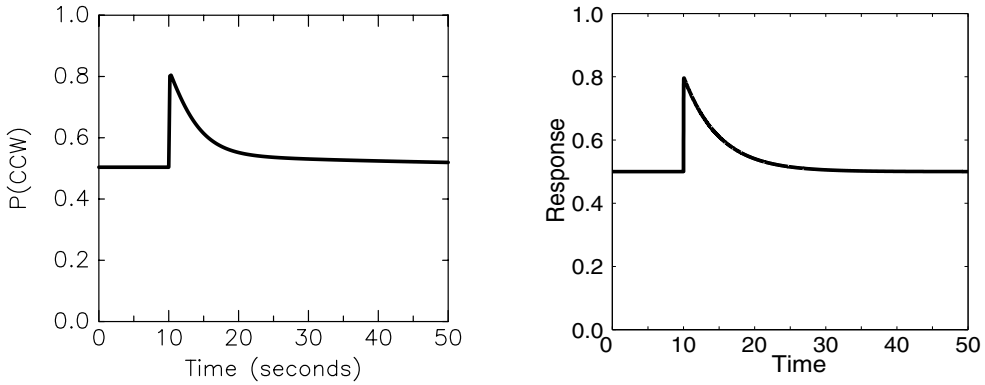


FIG. 2.2. *(left) The computed change in bias in response to a step change in attractant for a complete signal transduction model* [40]. *(right) The graph of the response of the internal dynamics* (2.2)–(2.3)*, given a step change of the signal. Here the response is defined as* $Response = 0.5 + y_1$; *moreover,* $g = Identity$, $t_e = 0.01$, *and* $t_a = 5$. *The signal function is* 0 *in the time interval* $[0, 10]$, *and the signal is equal to* 0.3 *in the time interval* $[10, 50]$.

for simplicity we will assume that h is a linear function of y_1 (which is always true for small responses y_1); i.e., we suppose that

$$\lambda(y) \equiv Response = \lambda_0 - by_1, \tag{2.5}$$

where λ_0 is the basal turning frequency for a fully adapted cell and b is a positive constant. The term by_1 describes the change in the turning frequency in response to a signal, and the negative sign accounts for the fact that an increase of y_1 should produce a decrease in the turning rate.

The function g in (2.2) and (2.3) describes the transduction of the signal, and a reasonable choice for this is to suppose that it depends on the fraction of receptors occupied, in which case

$$g(S) = G\left(\frac{S}{K_D + S}\right)$$

for some other function G, where K_D is the dissociation constant for the attractant [9]. We shall assume in the derivation that $g = \text{Identity}$, and that $t_e = 0$ in (2.2). The results for a general function g and $t_e \neq 0$ can be derived similarly, and we state them at the end of the corresponding sections (see (6.34), (7.12), etc.).

Whatever the choice of g, the formal solution to (2.2) and (2.3) can be obtained explicitly. However, because $x = x(t)$ is the cell position at time t in a given external concentration field, the integration must be along the cell trajectory, which is a biased random walk. Hence, $S(x, t)$ is a stochastic input to the signal transduction system.

3. Individual behavior. We suppose that the extracellular signal is specified as $S(x, t)$, and for the present we neglect the time spent tumbling; the tumble phase is incorporated in section 9.1. Let $p(x, v, y, t)$ be the density function of bacteria in a $(2N + m)$-dimensional phase space with coordinates (x, v, y), where $x \in \mathbb{R}^N$ is the position of a cell, $v \in V \subset \mathbb{R}^N$ is its velocity, and $y \in Y \subset \mathbb{R}^m$ is its internal state, which evolves according to (2.1). Thus $p(x, v, y, t)\mathrm{d}x\mathrm{d}v\mathrm{d}y$ is the number of cells with position between x and $x + \mathrm{d}x$, velocity between v and $v + \mathrm{d}v$, and internal state

between y and $y + \mathrm{d}y$. The evolution of p is governed by the following transport equation:

$$\frac{\partial p}{\partial t} + \nabla_x \cdot vp + \nabla_v \cdot Fp + \nabla_y \cdot fp = \mathcal{Q}, \tag{3.1}$$

where F denotes the external force acting on the individuals and $\mathcal{Q}$ is the rate of change of p due to reactions, random choices of velocity, collisions, etc. Here we ignore external forces and set $F \equiv 0$. Moreover, we assume that there is only one process represented in $\mathcal{Q}$: that which generates the random velocity change, and we assume that the changes are the result of a Poisson process of intensity $\lambda(y)$. Then

$$\mathcal{Q} = -\lambda(y)p(x, v, y, t) + \int_V \lambda(y)T(v, v', y)p(x, v', y, t)\mathrm{d}v',$$

where the kernel $T(v, v', y)$ gives the probability of a change in velocity from v' to v, given that a reorientation occurs. The kernel T is nonnegative and satisfies the normalization condition $\int_V T(v, v', y)\mathrm{d}v = 1$.

Consequently, the transport equation (3.1) takes the following form:

$$\frac{\partial p}{\partial t} + \nabla_x \cdot vp + \nabla_y \cdot fp = -\lambda(y)p + \int_V \lambda(y)T(v, v', y)p(x, v', y, t)\mathrm{d}v'. \tag{3.2}$$

The objective of this paper is to derive a macroscopic description for chemotaxis from the microscopic model, i.e., an evolution equation for the macroscopic density of individuals

$$n(x, t) = \int_Y \int_V p(x, v, y, t)\mathrm{d}v\mathrm{d}y. \tag{3.3}$$

Since we are primarily concerned with how the internal dynamics (2.1) influence the macroscopic behavior, we will only consider movement in one dimension. Considering 2D and 3D models does not alter the process of incorporating the internal dynamics into the macroscopic equations, but it does raise technical issues that will be discussed elsewhere [15]. Moreover, we assume that the speed is constant, and therefore we analyze the following generalization of the simple telegraph process described by (1.9): let $p^{\pm}(x, y, t)$ be the density of the particles that are at (x, t) with the internal state y and are moving to the right $(+)$ or left $(-)$, and suppose that the internal state evolves according to the system of equations (2.1). Then $p^{\pm}(x, y, t)$ satisfy the equations

$$\frac{\partial p^+}{\partial t} + s\frac{\partial p^+}{\partial x} + \sum_{i=1}^{m} \frac{\partial}{\partial y_i}\left[f_i(y, S)p^+\right] = \lambda(y)\left[-p^+ + p^-\right], \tag{3.4}$$

$$\frac{\partial p^-}{\partial t} - s\frac{\partial p^-}{\partial x} + \sum_{i=1}^{m} \frac{\partial}{\partial y_i}\left[f_i(y, S)p^-\right] = \lambda(y)\left[p^+ - p^-\right]. \tag{3.5}$$

Written as a system, this takes the form

$$\begin{aligned} &\frac{\partial}{\partial t}\begin{pmatrix} p^+ \\ p^- \end{pmatrix} + s\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix} p^+ \\ p^- \end{pmatrix} + \nabla_y \cdot \left[f(y, S)\begin{pmatrix} p^+ \\ p^- \end{pmatrix}\right] \\ &\quad = \lambda(y)\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} p^+ \\ p^- \end{pmatrix}. \end{aligned} \tag{3.6}$$

This is a hyperbolic system in diagonal form that has two independent characteristics for $s > 0$, and it is therefore strictly hyperbolic for $s > 0$.

To analyze the system (3.4)–(3.5), we must specify the internal dynamics (2.1) and the turning rate $\lambda(y)$. Here we shall use the internal dynamics (2.2)–(2.3) and the turning rate given by (2.5). Moreover, for simplicity we suppose that the signal $S(x)$ is a time-independent scalar function, that $t_e = 0$, and that $g =$ Identity. (The results for a general function g and $t_e \neq 0$ can be derived similarly, and we state them at the end of the corresponding sections; see (6.34), (7.12), etc.) Then the internal dynamics and the response are given as follows:

$$\frac{\mathrm{d}y_2}{\mathrm{d}t} = \frac{S(x) - y_2}{t_a}, \tag{3.7}$$

$$\lambda(y) \equiv \textit{Response} = \lambda_0 - b(S(x) - y_2). \tag{3.8}$$

It is convenient to define the new internal state variable z_2 as

$$z_2 = y_2 - S(x), \tag{3.9}$$

and then

$$\frac{\mathrm{d}z_2}{\mathrm{d}t} = \frac{S(x) - y_2}{t_a} - S'(x)\frac{\mathrm{d}x}{\mathrm{d}t} = -\frac{z_2}{t_a} \mp S'(x)s, \tag{3.10}$$

where the sign of the last term is determined by the sign of the velocity of the particle. Moreover,

$$\lambda(z_2) \equiv \lambda(y) = \lambda_0 - b(S(x) - y_2) = \lambda_0 + bz_2. \tag{3.11}$$

Later we will make use of an estimate on the internal state derived in the following lemma, and to avoid repetition, we introduce the following definition. Suppose that the cell moves in one dimension according to a velocity jump process with internal dynamics, and that the internal state z_2 of the cell evolves according to (3.10). We call this the *standard process*.

LEMMA 3.1. *Suppose that the cells execute the standard process, and suppose that*

$$|S'(x)| \leq K \ \text{ for } x \in \mathbb{R} \qquad \textit{and} \qquad |z_2(0)| \leq st_aK.$$

Then we have

$$|z_2(t)| \leq st_aK \ \text{ for } \ t \geq 0.$$

Proof. If $z_2(t) = -st_aK$, then the estimate $|S'(x)| \leq K$ implies $\frac{\mathrm{d}z_2}{\mathrm{d}t} \geq 0$. Similarly, if $z_2(t) = st_aK$, then the estimate $|S'(x)| \leq K$ implies $\frac{\mathrm{d}z_2}{\mathrm{d}t} \leq 0$. As $|z_2(0)| \leq st_aK$, we have $|z_2(t)| \leq st_aK$ for all $t \geq 0$. ∎

For a physically reasonable model we must ensure that the turning rate $\lambda(z_2)$ is always nonnegative, and for this we introduce the following *standing hypothesis*:

$$\text{Assume that} \quad |S'(x)| \leq \overline{C}, \quad \text{where } \overline{C} \text{ is given by} \quad \overline{C} = \frac{\lambda_0}{bst_a}. \tag{3.12}$$

Given (3.12), we have the following lemma.

LEMMA 3.2. *Suppose that the cells execute the standard process and that* (3.12) *is satisfied. Suppose that initially* $\lambda(z_2(0)) \geq 0$. *Then we have*

$$\lambda(z_2(t)) \geq 0 \qquad \text{for all } \ t \geq 0.$$

Proof. The linear turning rate (3.11) is nonnegative if and only if $z_2 \geq -\frac{\lambda_0}{b}$. As $\lambda(z_2(0)) \geq 0$, we have $z_2(0) \geq -\frac{\lambda_0}{b}$. Then Lemma 3.1 implies that $z_2(t) \geq -\frac{\lambda_0}{b}$ for $t \geq 0$. Consequently, the turning rate $\lambda(y)$ is nonnegative for all $t \geq 0$. ∎

In view of the preceding assumptions and simplifications, the evolution equations (3.4)–(3.5) for the densities $p^{\pm}(x, z_2, t)$ can be written as

$$\frac{\partial p^+}{\partial t} + s\frac{\partial p^+}{\partial x} + \frac{\partial}{\partial z_2}\left[\left(-\frac{z_2}{t_a} - sS'(x)\right)p^+\right] = (\lambda_0 + bz_2)\left[-p^+ + p^-\right], \tag{3.13}$$

$$\frac{\partial p^-}{\partial t} - s\frac{\partial p^-}{\partial x} + \frac{\partial}{\partial z_2}\left[\left(-\frac{z_2}{t_a} + sS'(x)\right)p^-\right] = (\lambda_0 + bz_2)\left[p^+ - p^-\right]. \tag{3.14}$$

In the following sections we use (3.13)–(3.14) to derive macroscopic equations. First, however, we address the question of existence and nonnegativity of the densities $p^{\pm}(x, z_2, t)$. In the following lemma we establish these properties for the general system (3.4)–(3.5), and this implies the result for (3.13)–(3.14).

LEMMA 3.3. *Suppose that* $f \in C^1(\mathbb{R}^m \times \mathbb{R}^M)$, *and let* $S : \mathbb{R} \times [0, \infty) \to \mathbb{R}^M$ *be continuous. Moreover, suppose that* $\lambda(y)$ *in* (3.4)–(3.5) *is always nonnegative, and that* $p_0^+ : \mathbb{R}^{m+1} \to [0, \infty)$ *and* $p_0^- : \mathbb{R}^{m+1} \to [0, \infty)$ *are given nonnegative compactly supported* C^1*-functions. Then there exists a domain* $Q \subset \mathbb{R}^{m+1} \times [0, \infty)$ *containing the entire plane* $t = 0$ *such that the system of equations* (3.4)–(3.5) *with initial conditions* $p^{\pm}(x, y, 0) = p_0^{\pm}(x, y)$ *has a unique* C^1*-solution in* Q. *Moreover, the functions* $p^{\pm}$ *are nonnegative wherever they are defined.*

Proof. As remarked earlier, the general system (3.6) has two independent characteristics, and this applies to (3.4)–(3.5) as well. Consequently, we can apply the modified implicit function theorem to show local existence of a unique classical solution (see [3, section 2.4.4]). To prove nonnegativity, let us consider that the solution is NOT nonnegative and define

$$t^0 = \inf\{\, \tau \ : \text{ there exists } (x, y) \text{ such that } p^+(x, y, \tau) < 0 \text{ or } p^-(x, y, \tau) < 0 \,\};$$

i.e., t^0 is the last time for which the nonnegativity of solutions is satisfied. In particular, we have $p^{\pm}(\cdot, \cdot, t^0) \geq 0$.

Let $\chi^+_{x,y,t^0}(\tau)$ be a characteristic through the point (x, y, t^0) for (3.4), and let $\chi^-_{x,y,t^0}(\tau)$ be a characteristic through the point (x, y, t^0) for (3.5). Thus $\chi^{\pm}_{x,y,t^0}(\tau)$ are curves in the $(m + 2)$-dimensional (x, y, t)-space along which (3.4)–(3.5) read as

$$\frac{d}{d\tau}p^+(\chi^+_{x,y,t^0}(\tau)) = -c(\chi^+_{x,y,t^0}(\tau))p^+(\chi^+_{x,y,t^0}(\tau)) + \lambda(\chi^+_{x,y,t^0}(\tau))p^-(\chi^+_{x,y,t^0}(\tau)), \tag{3.15}$$

$$\frac{d}{d\tau}p^-(\chi^-_{x,y,t^0}(\tau)) = \lambda(\chi^-_{x,y,t^0}(\tau))p^+(\chi^-_{x,y,t^0}(\tau)) - c(\chi^-_{x,y,t^0}(\tau))p^-(\chi^-_{x,y,t^0}(\tau)), \tag{3.16}$$

wherein

$$c(x,y,t) = \lambda(y) + \sum_{i=1}^{m} \frac{\partial f_i(y, S(x,t))}{\partial y_i}.$$

Let us suppose that $p^+(x,y,t^0) = 0$. Then (3.15) implies (using $\tau = t^0$, $\lambda(y) \geq 0$ and $p^-(x,y,t^0) \geq 0$)

$$\frac{d}{d\tau} p^+(x,y,t^0) \geq 0. \tag{3.17}$$

Similarly, if $p^-(x,y,t^0) = 0$, then (3.16) gives $\frac{d}{d\tau} p^-(x,y,t^0) \geq 0$. Consequently, there exists a constant $c > 0$ such that the solutions $p^{\pm}$ are nonnegative in the time interval $[t^0, t^0 + c)$. This is a contradiction with the choice of t^0. □

4. Moment equations. The next step is to derive evolution equations for macroscopic variables from the simplified system (3.13)–(3.14). Since there are only two velocities and one internal state variable, the density $n(x,t)$ is given by (cf. (3.3))

$$n(x,t) = \int_{\mathbb{R}} p^+(x,z_2,t) + p^-(x,z_2,t) \mathrm{d}z_2. \tag{4.1}$$

The objective is to derive an evolution equation involving only n, if possible. For this purpose define $\mathcal{N} = p^+(x,z_2,t) + p^-(x,z_2,t)$ and $\mathcal{J} = s(p^+(x,z_2,t) - p^-(x,z_2,t))$; the former is the microscopic particle density, obtained by integrating p over v, while the latter is a microscopic flux obtained similarly. In this notation, (4.1) can be written

$$n(x,t) = \int_{\mathbb{R}} \mathcal{N}(x,z_2,t) \mathrm{d}z_2, \tag{4.2}$$

and we define the additional moments

$$j(x,t) = \int_{\mathbb{R}} \mathcal{J}(x,z_2,t) \mathrm{d}z_2, \tag{4.3}$$

$$n_1(x,t) = \int_{\mathbb{R}} z_2 \mathcal{N}(x,z_2,t) \mathrm{d}z_2, \tag{4.4}$$

$$j_1(x,t) = \int_{\mathbb{R}} z_2 \mathcal{J}(x,z_2,t) \mathrm{d}z_2, \tag{4.5}$$

and

$$j_2(x,t) = \int_{\mathbb{R}} (z_2)^2 \mathcal{J}(x,z_2,t) \mathrm{d}z_2. \tag{4.6}$$

The quantity j is the macroscopic particle flux, n_1 and j_1 are first moments with respect to the slow component of the internal state of the microscopic density and flux, respectively, and j_2 is the second moment of the microscopic flux with respect to the slow component of the internal state. All moments with respect to z_2 are well defined by virtue of Lemma 3.1 and the standing assumption (3.12), which implies that p vanishes identically outside some sufficiently large interval in $|z_2|$.

Next, by multiplying (3.13) and (3.14) by 1 or z_2, integrating with respect to z_2, and adding or subtracting the resulting equations, we obtain the following four moment equations:

$$\frac{\partial n}{\partial t} + \frac{\partial j}{\partial x} = 0, \tag{4.7}$$

$$\frac{\partial j}{\partial t} + s^2 \frac{\partial n}{\partial x} = -2\lambda_0 j - 2b j_1, \tag{4.8}$$

$$\frac{\partial n_1}{\partial t} + \frac{\partial j_1}{\partial x} = -S'(x) j - \frac{1}{t_a} n_1, \tag{4.9}$$

$$\frac{\partial j_1}{\partial t} + s^2 \frac{\partial n_1}{\partial x} = -s^2 S'(x) n - \left(2\lambda_0 + \frac{1}{t_a}\right) j_1 - 2b j_2. \tag{4.10}$$

We see that the moment equations for a density-flux pair introduce a higher-order flux via the change in turning rate, as measured by b. If $S(x)$ is constant, the effect of the signal disappears and the second pair is uncoupled from the first. In section 6 we rescale the variables and then close the system of four moment equations with the assumption that

$$j_2 = 0; \tag{4.11}$$

i.e., we simply neglect the second-order flux. The moment closure (4.11) will be rigorously justified in the case of shallow gradients of the signal. The moment closures for arbitrary signal functions will be discussed in section 9.2.

Of course one can ask what a lower-order closure (i.e., the closure assumption on j_1) leads to, and it is easy to see that if we assume that $j_1 = 0$, we obtain the telegraph equation

$$\frac{1}{2\lambda_0} \frac{\partial^2 n}{\partial t^2} + \frac{\partial n}{\partial t} = \frac{s^2}{2\lambda_0} \frac{\partial^2 n}{\partial x^2}. \tag{4.12}$$

Since the external signal S is completely absent from this equation, this approximation is not suitable for studying the dependence of n on the signal. Clearly (4.12) applies if there is no effect of the signal on the turning rate, and in this case there can be no taxis.

From (4.7)–(4.10) we can derive evolution equations for various statistics of the motion that give insight into the asymptotics of solutions of the system of moment equations. These are derived in the following section.

5. Evolution of certain statistics of the motion. We denote by n_0 the total number of particles in the domain. This is a conserved quantity and is given by

$$n_0 = \int_{\mathbb{R}} n(x,t) \mathrm{d}x.$$

The mean position of the particles $\langle x \rangle(t)$ and the mean square displacement $\langle x^2 \rangle(t)$ are given by

$$\langle x \rangle(t) = \frac{1}{n_0} \int_{\mathbb{R}} x n(x,t) \mathrm{d}x, \qquad \langle x^2 \rangle(t) = \frac{1}{n_0} \int_{\mathbb{R}} \left(x - \langle x \rangle\right)^2 n(x,t) \mathrm{d}x. \tag{5.1}$$

We define the spatial moments of j, j_1, and j_2 as follows:

$$j^0 = \int_{\mathbb{R}} j(x,t)\mathrm{d}x, \quad j^x = \int_{\mathbb{R}} xj(x,t)\mathrm{d}x, \quad j_1^0 = \int_{\mathbb{R}} j_1(x,t)\mathrm{d}x, \quad j_2^0 = \int_{\mathbb{R}} j_2(x,t)\mathrm{d}x.$$

Then, multiplying (4.7) by x and by $(x - \langle x\rangle(t))^2$, and integrating the resulting equations with respect to x, we find that

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle x\rangle = \frac{j^0}{n_0} \qquad \text{and} \qquad \frac{\mathrm{d}}{\mathrm{d}t}\langle x^2\rangle = \frac{2j^x - 2\langle x\rangle j^0}{n_0}. \tag{5.2}$$

Integrating (4.8) and (4.10) with respect to x, we obtain the evolution equations for j^0 and j_1^0:

$$\frac{\mathrm{d}}{\mathrm{d}t}j^0 = -2\lambda_0 j^0 - 2bj_1^0, \tag{5.3}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}j_1^0 = -s^2\int_{\mathbb{R}} S'(x)n(x,t)\mathrm{d}x - \left(2\lambda_0 + \frac{1}{t_a}\right)j_1^0 - 2bj_2^0. \tag{5.4}$$

We can solve (5.3)–(5.4) explicitly for j^0 as a function of the quantities $\int_{\mathbb{R}} S'(x)n(x,t)\mathrm{d}x$ and j_2^0. Then the evolution equation (5.2) for $\langle x\rangle$ reads

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle x\rangle = \frac{e^{-2\lambda_0 t}}{n_0}\Bigg[j^0(0) + \int_0^t e^{t'/t_a}2b\Bigg(- j_1^0(0) + \int_0^{t'} e^{2\lambda_0 t'' + t''/t_a}\left\{s^2\int_{\mathbb{R}} S'(x)n(x,t)\mathrm{d}x + 2bj_2^0\right\}\mathrm{d}t''\Bigg)\mathrm{d}t'\Bigg].$$

Thus the mean displacement is driven by the flux j^0, which is in turn forced by the projection of the local density onto the gradient, as given by the integral term in S', as well as by the higher-order flux j_2^0. From the foregoing one can conclude that if $\langle x\rangle$ tends to a constant as $t \to \infty$, then the total flux j^0 must vanish as $t \to \infty$, and this in turn requires that the term $s^2\int_{\mathbb{R}} S'(x)n(x,t)\mathrm{d}x + 2bj_2^0$ must tend to zero. Thus this is a necessary, but not sufficient, condition for steady patterns. Similarly, one can derive the system of evolution equations for the mean square displacement.

In order to gain further insight into the evolution of the statistics of motion, let us suppose that

$$S'(x) = C = \text{constant} \qquad \text{and} \qquad j_2 = 0.$$

To derive equations for the mean position $\langle x\rangle$ and the mean square displacement $\langle x^2\rangle$ under this restriction we introduce some additional moments

$$n_1^0 = \int_{\mathbb{R}} n_1(x,t)\mathrm{d}x, \qquad n_1^x = \int_{\mathbb{R}} xn_1(x,t)\mathrm{d}x, \qquad \text{and} \qquad j_1^x = \int_{\mathbb{R}} xj_1(x,t)\mathrm{d}x.$$

Then, integrating (4.8), (4.9), and (4.10) with respect to x, we obtain the system

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} j^0 \\ n_1^0 \\ j_1^0 \end{pmatrix} = \begin{pmatrix} -2\lambda_0 & 0 & -2b \\ -C & -\frac{1}{t_a} & 0 \\ 0 & 0 & -\left(2\lambda_0 + \frac{1}{t_a}\right) \end{pmatrix}\begin{pmatrix} j^0 \\ n_1^0 \\ j_1^0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ -s^2Cn_0 \end{pmatrix}. \tag{5.5}$$

Similarly, multiplying (4.8), (4.9), and (4.10) by x and integrating with respect to x, we obtain

(5.6)
$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} j^x \\ n_1^x \\ j_1^x \end{pmatrix} = \begin{pmatrix} -2\lambda_0 & 0 & -2b \\ -C & -\frac{1}{t_a} & 0 \\ 0 & 0 & -\left(2\lambda_0 + \frac{1}{t_a}\right) \end{pmatrix}\begin{pmatrix} j^x \\ n_1^x \\ j_1^x \end{pmatrix} + \begin{pmatrix} s^2 n_0 \\ j_1^0 \\ s^2 n_1^0 - s^2 C\langle x\rangle n_0 \end{pmatrix}.$$

Together (5.5) and (5.6) form a system of six linear nonhomogeneous equations. The eigenvalues of the matrix of this 6×6 system, which has the 3×3 blocks of the separate system along the diagonal, are in fact just the diagonal entries, and are therefore real and negative. Since n_0 is constant, it follows that the system has a unique stable steady state given by

$$j^0 = \frac{bs^2 C t_a}{\lambda_0 + 2\lambda_0^2 t_a} n_0 \qquad \text{and} \qquad 2j^x - 2\langle x\rangle j^0 = \left(\frac{s^2}{\lambda_0} + \frac{2b^2 s^4 C^2 t_a^3}{(\lambda_0 + 2\lambda_0^2 t_a)^2}\right) n_0.$$

Thus, (5.2) implies that, asymptotically for $t \to \infty$, we have

$$\langle x\rangle(t) = \frac{bs^2 C t_a}{\lambda_0 + 2\lambda_0^2 t_a} t \qquad \text{and} \qquad \langle x^2\rangle(t) = \left(\frac{s^2}{\lambda_0} + \frac{2b^2 s^4 C^2 t_a^3}{(\lambda_0 + 2\lambda_0^2 t_a)^2}\right) t. \tag{5.7}$$

The second of these shows that when the gradient $S'(x) = C$, the standard process is asymptotically a diffusion process with diffusion constant

$$D = \frac{s^2}{2\lambda_0} + \frac{b^2 s^4 C^2 t_a^3}{(\lambda_0 + 2\lambda_0^2 t_a)^2}. \tag{5.8}$$

Later, using the scaling in (6.8), we will see that the second term in (5.8) is smaller than the first, and thus in (5.8) $D \sim s^2/2\lambda_0$.

6. The hyperbolic scaling and derivation of a hyperbolic chemotaxis equation. The macroscopic equations for n and j that can be obtained from the moment equations depend on the time and space scales of interest. In this section we use a hyperbolic scaling of space and time, which can capture the initial time evolution of the system. Using this scaling, we give a heuristic derivation of a hyperbolic version of the classical chemotaxis equation. Moreover, we also give an alternate random walk interpretation to the derived hyperbolic chemotaxis equation in section 6.1. In section 7, we use a parabolic scaling valid for large times, which leads to the classical chemotaxis equation. This equation is also a parabolic limit of the hyperbolic chemotaxis equation derived here in cases where the signal is fixed. When the signal itself evolves in time, this need not be true.

Let L, T, and s_0 be scale factors for the length, time, and velocity, respectively; let N_0 be a scale factor for the particle density; and define the dimensionless variables

$$\hat{x} = \frac{x}{L}, \qquad \hat{t} = \frac{t}{T}, \qquad \hat{n} = \frac{n}{N_0}, \qquad \hat{j} = \frac{j}{N_0 s_0},$$
$$\hat{n}_1 = \frac{n_1}{N_0}, \qquad \hat{j}_1 = \frac{j_1}{N_0 s_0}, \qquad \text{and} \qquad \hat{j}_2 = \frac{j_2}{N_0 s_0}. \tag{6.1}$$

Then the moment equations (4.7)–(4.10) can be written in the dimensionless form

$$\frac{\partial \hat{n}}{\partial \hat{t}} + \varepsilon \frac{\partial \hat{j}}{\partial \hat{x}} = 0, \tag{6.2}$$

$$\frac{\partial \hat{j}}{\partial \hat{t}} + \varepsilon \hat{s}^2 \frac{\partial \hat{n}}{\partial \hat{x}} = -2\hat{\lambda}_0 \hat{j} - 2\hat{b}\hat{j}_1, \tag{6.3}$$

$$\frac{\partial \hat{n}_1}{\partial \hat{t}} + \varepsilon \frac{\partial \hat{j}_1}{\partial \hat{x}} = -\varepsilon \hat{S}'(\hat{x})\hat{j} - \frac{1}{\hat{t}_a}\hat{n}_1, \tag{6.4}$$

$$\frac{\partial \hat{j}_1}{\partial \hat{t}} + \varepsilon \hat{s}^2 \frac{\partial \hat{n}_1}{\partial \hat{x}} = -\varepsilon \hat{s}^2 \hat{S}'(\hat{x})\hat{n} - \left(2\hat{\lambda}_0 + \frac{1}{\hat{t}_a}\right)\hat{j}_1 - 2\hat{b}\hat{j}_2, \tag{6.5}$$

where

$$\varepsilon \equiv \left(\frac{s_0 T}{L}\right), \quad \hat{s} \equiv \frac{s}{s_0}, \quad \hat{\lambda}_0 \equiv \lambda_0 T, \quad \hat{b} \equiv bT, \quad \hat{t}_a \equiv \frac{t_a}{T}, \quad \text{and} \quad \hat{S}'(\hat{x}) \equiv LS'(x). \tag{6.6}$$

In order to derive the macroscopic equations, we have to specify L, T, and s_0 and estimate the dimensionless parameters. The typical space scale of macroscopic experiments is several millimeters or centimeters, a typical speed of bacterium is $s = 10$ – 20 μm/sec, and a characteristic time scale depends on our interests. Here, we choose

$$T = 1 \text{ sec}, \qquad L = 1 \text{ mm}, \quad \text{and} \quad s_0 = 10\mu\text{m/sec}; \tag{6.7}$$

i.e., we use a time scale that is of the same order as the mean time between directional changes, since this characterizes the initial evolution. Assuming that the adaptation time and the bias are also of the same order as the mean run time, we get

$$\varepsilon \approx 10^{-2} \quad \text{and} \quad \hat{s} \sim \hat{\lambda}_0 \sim \hat{b} \sim \hat{t}_a \sim \mathcal{O}(1). \tag{6.8}$$

Using an approximation given later, this scaling will lead to a hyperbolic chemotaxis equation. For simplicity, we drop the hats on x, t, s, λ_0, b, t_a, S and the hats on moments, and use the same symbols for the dimensionless variables. Then the moment equations (6.2)–(6.5) read as follows:

$$\frac{\partial n}{\partial t} + \varepsilon \frac{\partial j}{\partial x} = 0, \tag{6.9}$$

$$\frac{\partial j}{\partial t} + \varepsilon s^2 \frac{\partial n}{\partial x} = -2\lambda_0 j - 2b j_1, \tag{6.10}$$

$$\frac{\partial n_1}{\partial t} + \varepsilon \frac{\partial j_1}{\partial x} = -\varepsilon S'(x) j - \frac{1}{t_a} n_1, \tag{6.11}$$

$$\frac{\partial j_1}{\partial t} + \varepsilon s^2 \frac{\partial n_1}{\partial x} = -\varepsilon s^2 S'(x) n - \left(2\lambda_0 + \frac{1}{t_a}\right) j_1 - 2b j_2. \tag{6.12}$$

In order to close this system we have to specify

$$j_2 = \mathcal{F}(n, j, n_1, j_1), \tag{6.13}$$

where the functional $\mathcal{F}$ is to be determined. To do that, we first rewrite our standing assumption (3.12) using hyperbolic scaling (6.7)–(6.8). We denote the dimensionless constant $\overline{C}$ again as $\overline{C}$ for simplicity, and then (3.12) for the dimensionless signal gradient reads as follows:

$$|S'(x)| \le \overline{C}, \qquad \text{where} \qquad \overline{C} \sim \mathcal{O}\left(\frac{1}{\varepsilon}\right). \tag{6.14}$$

Thus the maximal possible gradient that satisfies (3.12) is $\mathcal{O}(1/\varepsilon)$ on the hyperbolic scale. In other words, the restriction on the gradients that guarantees positivity of the turning rate is very weak, and we strengthen it as follows.

DEFINITION 6.1. *We call the signal gradient shallow on the hyperbolic scale if*

$$|S'(x)| \le \overline{K}, \qquad \textit{where} \qquad \overline{K} \sim \mathcal{O}(1). \tag{6.15}$$

In the following, we investigate the case of shallow gradients, and to do that, we have to estimate the moments in (6.9)–(6.12).

LEMMA 6.2. *Suppose that the signal gradient is shallow. Then the moments in* (6.9)–(6.12) *can be estimated as follows:*

$$\frac{j}{n} \le K_1, \qquad \frac{n_1}{n} \le \varepsilon K_2, \qquad \frac{j_1}{n} \le \varepsilon K_3, \qquad \frac{j_2}{n} \le \varepsilon^2 K_4, \tag{6.16}$$

where the constants K_1, K_2, K_3, and K_4 are $\mathcal{O}(1)$.

Proof. We use (4.6) rescaled by (6.1), (6.6), the nonnegativity of $p^{\pm}$, Lemma 3.1, (4.1), and (6.8) to estimate

$$\begin{aligned} j_2 &= \frac{s}{N_0}\int_{\mathbb{R}} (z_2)^2 \left[p^+(x,z_2,t) - p^-(x,z_2,t)\right] \mathrm{d}z_2 \\ &\le \frac{s}{N_0}\int_{\mathbb{R}} (z_2)^2 \left[p^+(x,z_2,t) + p^-(x,z_2,t)\right] \mathrm{d}z_2 \\ &\le \varepsilon^2(\overline{K}s\,t_a)^2 \frac{s}{N_0}\int_{\mathbb{R}} \left[p^+(x,z_2,t) + p^-(x,z_2,t)\right] \mathrm{d}z_2 = \varepsilon^2(\overline{K}t_a)^2 s^3\, n = \varepsilon^2 K_4\, n, \end{aligned}$$

where $K_4 \sim \mathcal{O}(1)$. This proves the last inequality in (6.16), and the proof of the other inequalities is similar. ∎

Therefore the term $2bj_2$ in equation (6.12) is $\mathcal{O}(\varepsilon^2)$ when the gradient is shallow, and we can close the moment equations (6.9)–(6.12) with the moment closure assumption

$$j_2 = 0. \tag{6.17}$$

This will introduce the error of order $\mathcal{O}(\varepsilon^2)$ into (6.12). The corresponding orders of the remaining moments are given in Lemma 6.2.

Next we show that one can obtain a hyperbolic chemotaxis equation for n, provided a certain assumption on the decay of modes holds. To do that, we write the system (6.9)–(6.12) in the matrix form

$$\frac{\partial v}{\partial t} + \varepsilon \frac{\partial}{\partial x}(Av) = B(x,\varepsilon)v + r, \tag{6.18}$$

where

$$v = \begin{pmatrix} n \\ j \\ n_1 \\ j_1 \end{pmatrix}, \qquad A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ s^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & s^2 & 0 \end{pmatrix}, \qquad r = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -2bj_2 \end{pmatrix}, \tag{6.19}$$

and

$$B(x,\varepsilon) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -2\lambda_0 & 0 & -2b \\ 0 & -\varepsilon S'(x) & -\frac{1}{t_a} & 0 \\ -\varepsilon s^2 S'(x) & 0 & 0 & -\left(2\lambda_0 + \frac{1}{t_a}\right) \end{pmatrix}. \tag{6.20}$$

Equation (6.18) holds for any signal function satisfying the standing assumption (3.12), since it is just a different formulation of the system (6.9)–(6.12). Assuming (6.17), the system (6.18) can be written in the form

$$\frac{\partial v}{\partial t} + \varepsilon \frac{\partial}{\partial x}(Av) = B(x,\varepsilon)v. \tag{6.21}$$

This is a hyperbolic system of four linear PDEs with nonconstant coefficients for four unknowns: n, j, n_1, and j_1. The matrix $B(x,\varepsilon)$ has the interesting property that its eigenvalues do not depend on the signal $S(x)$, and consequently the eigenvalues of $B(x,\varepsilon)$ are independent of ε and x. An easy calculation gives the following four (not necessarily distinct) eigenvalues of $B(x,\varepsilon)$:

$$\lambda_1 = 0, \quad \lambda_2 = -2\lambda_0, \quad \lambda_3 = -\frac{1}{t_a}, \quad \lambda_4 = -\frac{1+2\lambda_0 t_a}{t_a}. \tag{6.22}$$

Let us note that three of the eigenvalues, λ_2, λ_3, and λ_4, are negative. Moreover, $\lambda_4 < \lambda_2$ and $\lambda_4 < \lambda_3$.

System (6.21) cannot be solved explicitly, so we simplify it heuristically as follows. First consider the system (6.21) with $\varepsilon = 0$, in which case (6.21) reduces to the system of ordinary differential equations

$$\frac{\partial w}{\partial t} = B(x,0)w,$$

where the matrix $B(x,0)$ has four eigenvalues given by (6.22). Consequently, the long time behavior is given by the eigenvectors corresponding to the largest eigenvalues. Next, let us consider the system (6.21) with $\varepsilon \neq 0$. As ε is a small parameter, we use the following heuristic argument to derive a hyperbolic chemotaxis equation.

The eigenvectors of $B(x,\varepsilon)$ are

$$\lambda_1 = 0: \quad \vartheta_1 = \begin{pmatrix} \lambda_0 + 2\lambda_0^2 t_a \\ b\varepsilon s^2 S'(x) t_a \\ -b(\varepsilon s S'(x) t_a)^2 \\ -\varepsilon s^2 S'(x) t_a \lambda_0 \end{pmatrix}, \qquad \lambda_2 = -2\lambda_0: \quad \vartheta_2 = \begin{pmatrix} 0 \\ -1 + 2\lambda_0 t_a \\ \varepsilon S'(x) t_a \\ 0 \end{pmatrix}, \tag{6.23}$$

$$\lambda_3 = -\frac{1}{t_a}: \quad \vartheta_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \lambda_4 = -\frac{1+2\lambda_0 t_a}{t_a}: \quad \vartheta_4 = \begin{pmatrix} 0 \\ 2b\lambda_0 t_a \\ \varepsilon S'(x) b t_a \\ \lambda_0 \end{pmatrix}. \tag{6.24}$$

Let us suppose that $2\lambda_0 \neq \frac{1}{t_a}$. Then we can write the unknown vector function $v(x,t)$ as a linear combination of the eigenvectors ϑ_i, $i = 1, \ldots, 4$, i.e.,

$$v(x,t) = c_1(x,t)\vartheta_1 + c_2(x,t)\vartheta_2 + c_3(x,t)\vartheta_3 + c_4(x,t)\vartheta_4. \tag{6.25}$$

We are interested in the evolution of the first component of v, which is n. The first component is nonzero only for the vector ϑ_1, and consequently we have

$$n(x,t) = (\lambda_0 + 2\lambda_0^2 t_a)c_1(x,t).$$

Then (6.25) reads as follows:

$$v(x,t) = \frac{n(x,t)}{(\lambda_0 + 2\lambda_0^2 t_a)}\vartheta_1 + c_2(x,t)\vartheta_2 + c_3(x,t)\vartheta_3 + c_4(x,t)\vartheta_4. \tag{6.26}$$

The parameter ε is small compared with $2\lambda_0 + \frac{1}{t_a}$ (see (6.8)). Consequently, the major dynamical features will be given by the eigenvectors corresponding to the zero eigenvalue and the eigenvalues with lower absolute value. We have the inequalities

$$\lambda_4 < \lambda_2 < \lambda_1 = 0 \quad \text{and} \quad \lambda_4 < \lambda_3 < \lambda_1 = 0, \tag{6.27}$$

and therefore we consider the projection

$$v(x,t) = \frac{n(x,t)}{(\lambda_0 + 2\lambda_0^2 t_a)}\vartheta_1 + c_2(x,t)\vartheta_2 + c_3(x,t)\vartheta_3,$$

to obtain (from the fourth component of the vector v)

$$j_1(x,t) = -\frac{\varepsilon s^2 S'(x) t_a}{1 + 2\lambda_0 t_a} n(x,t). \tag{6.28}$$

This can be used to reduce the system (6.21) to the following system of two equations:

$$\frac{\partial n}{\partial t} + \varepsilon \frac{\partial j}{\partial x} = 0, \tag{6.29}$$

$$\frac{\partial j}{\partial t} + \varepsilon s^2 \frac{\partial n}{\partial x} = -2\lambda_0 j + 2b \frac{\varepsilon s^2 S'(x) t_a}{1 + 2\lambda_0 t_a} n. \tag{6.30}$$

The last step is to reduce these two equations to one equation for n. To this end, we differentiate (6.29) with respect to t and (6.30) with respect to x to obtain

$$\frac{\partial^2 n}{\partial t^2} + \varepsilon \frac{\partial^2 j}{\partial t \partial x} = 0, \qquad \frac{\partial^2 j}{\partial x \partial t} + \varepsilon s^2 \frac{\partial^2 n}{\partial x^2} = -2\lambda_0 \frac{\partial j}{\partial x} + 2b \frac{\partial}{\partial x} \frac{\varepsilon s^2 S'(x) t_a}{1 + 2\lambda_0 t_a} n. \tag{6.31}$$

Then, solving (6.31) for n, we obtain the hyperbolic version of the classical chemotaxis equation (compare with (1.5)):

$$\frac{\partial^2 n}{\partial t^2} + 2\lambda_0 \frac{\partial n}{\partial t} = \frac{\partial}{\partial x}\left(\varepsilon^2 s^2 \frac{\partial n}{\partial x} - \frac{2\, b\, \varepsilon^2 s^2 t_a}{1 + 2\lambda_0 t_a} S'(x) n \right). \tag{6.32}$$

Finally, let us note that s (or $\hat{s}$) given by (6.6) is the value of the speed of bacteria in units of s_0. On the other hand, if we give the values for characteristic time T and length L by (6.7), then the characteristic speed can be also considered as $L/T = 1$

mm/sec. In these units, the value of bacterial speed is simply given by $\overline{s} = \varepsilon s$. Using $\overline{s}$ instead of s, we can rewrite (6.32) in the following form:

$$\frac{\partial^2 n}{\partial t^2} + 2\lambda_0 \frac{\partial n}{\partial t} = \frac{\partial}{\partial x}\left(\overline{s}^2 \frac{\partial n}{\partial x} - \frac{2\, b\, \overline{s}^2\, t_a}{1 + 2\lambda_0 t_a} S'(x) n\right). \tag{6.33}$$

Here, the chemotactic sensitivity is given as a function of bacterial speed $\overline{s}$, adaptation time t_a, and turning parameters λ_0 and b, namely,

$$\chi = \frac{b\, \overline{s}^2\, t_a}{\lambda_0 + 2\lambda_0^2 t_a},$$

which we have already derived for the case in which the signal gradient is constant (see (5.7)). Equation (6.33) was derived for the simplified cartoon model (3.7), but a similar analysis can be done for the full model (2.2)–(2.3) with $t_e \neq 0$ and a general function g. This leads to the following hyperbolic chemotaxis equation (cf. (6.33)),

$$\frac{\partial^2 n}{\partial t^2} + 2\lambda_0 \frac{\partial n}{\partial t} = \frac{\partial}{\partial x}\left(\overline{s}^2 \frac{\partial n}{\partial x} - g'(S(x)) \frac{2\, b\, \overline{s}^2\, t_a}{(1 + 2\lambda_0 t_a)(1 + 2\lambda_0 t_e)} S'(x) n\right), \tag{6.34}$$

and the chemotactic sensitivity is now given by

$$\chi = g'(S(x)) \frac{b\, \overline{s}^2\, t_a}{\lambda_0 (1 + 2\lambda_0 t_a)(1 + 2\lambda_0 t_e)}. \tag{6.35}$$

Note that we can derive (6.33) as the limit $t_e \to 0$ of (6.34) for $g = \text{Identity}$. It should also be noted that in either case (and those that follow) the chemotactic sensitivity vanishes as $t_a \to 0$, which is to be expected since the system adapts instantaneously in this case. In this limit one sees, via (3.12) and Lemma 6.2, the interplay between the adaptation time and the allowable magnitude of the gradient: as $t_a \to 0$ we can allow the bound $\bar{K}$ on $|S'(x)|$ in Definition 6.1 to grow as long as $|\bar{K} t_a| \sim \mathcal{O}(1)$, and we obtain the same conclusions as in Lemma 6.2.

In contrast, one cannot extract from (6.35) the effect of letting $t_a \to \infty$, because the derivation of (6.34) or its simplified version (6.33) make use of the inequalities (6.27) to conclude that the projection of the solution onto the eigenvector ϑ_4 dies out faster than other modes. The spectral gap in (6.27) between λ_4 and other eigenvalues does not persist in the limit $t_a \to \infty$, and in addition the constant $\overline{C}$ tends to zero in the standing assumption (3.12). Consequently, we need a model for large t_a in which the turning rate is given by some nonlinear nonnegative function of the signal. Let us also note for later reference that the hyperbolic chemotaxis equation (6.34) gives the same parabolic limit as we will derive in section 7 from the full system (6.21).

6.1. A different random walk interpretation of the hyperbolic chemotaxis equation. As we observed in the discussion in the introduction and in [15], biasing the turning rates depending on the direction of travel will lead to a nonzero chemotactic velocity, and in this section we show that one can explicitly extract that bias from the chemotactic sensitivity derived from the cartoon internal dynamics. This leads back to a system of equations for left- and right-moving particles, but now without internal dynamics. One could obtain this by inverting the procedure that leads from the hyperbolic system without internal dynamics to the second-order scalar equation for the total density, but we proceed directly. The advantage of this system is that it provides a direct step to a microscopic model that bypasses the internal dynamics, and hence may be better suited for stochastic simulations. It also

suggests a model for chemotaxis in eukaryotic cells, which can measure gradients over their length.

Consider a random walk in which a particle moves along the x-axis at a constant speed $\overline{s}$ but at random instants of time reverses its direction according to a Poisson process with the turning frequency

$$\lambda = \lambda_0 \pm \frac{b\,\overline{s}\,t_a g'(S(x))}{(1+2\lambda_0 t_a)(1+2\lambda_0 t_e)} S'(x). \tag{6.36}$$

Here the sign depends on the direction of the particles: plus for particles moving to the left and minus for particles moving to the right. Let $u^{\pm}(x,t)$ be the density of particles at (x,t) that are moving to the right (plus) or left (minus) (note that here, and only in this section, there are no internal variables). Then $u^{\pm}(x,t)$ satisfy the equations

$$\begin{aligned} \frac{\partial u^+}{\partial t} + \overline{s}\frac{\partial u^+}{\partial x} = &-\left(\lambda_0 - \frac{b\,\overline{s}\,t_a g'(S(x))}{(1+2\lambda_0 t_a)(1+2\lambda_0 t_e)} S'(x)\right) u^+ \\ &+\left(\lambda_0 + \frac{b\,\overline{s}\,t_a g'(S(x))}{(1+2\lambda_0 t_a)(1+2\lambda_0 t_e)} S'(x)\right) u^-, \end{aligned} \tag{6.37}$$

$$\begin{aligned} \frac{\partial u^-}{\partial t} - \overline{s}\frac{\partial u^-}{\partial x} = &\left(\lambda_0 - \frac{b\,\overline{s}\,t_a g'(S(x))}{(1+2\lambda_0 t_a)(1+2\lambda_0 t_e)} S'(x)\right) u^+ \\ &-\left(\lambda_0 + \frac{b\,\overline{s}\,t_a g'(S(x))}{(1+2\lambda_0 t_a)(1+2\lambda_0 t_e)} S'(x)\right) u^-. \end{aligned} \tag{6.38}$$

The density of particles at (x,t) is given by the sum $n(x,t) = u^+(x,t) + u^-(x,t)$. Then, adding and subtracting (6.37) and (6.38), one can rewrite them as a system of two equations of the form (6.29)–(6.30) for the variables $n = u^+ + u^-$ and $j = u^+ - u^-$. Then one can follow the same procedure as before to show that the density n of these direction-sensing random walkers is described by (6.34). In this case, (6.34) is valid for all times for all biologically reasonable parameter regimes.

7. The parabolic scaling and derivation of the classical chemotaxis equation. The previous analysis used the scaling (6.7)–(6.8) and led to the hyperbolic chemotaxis equation (6.33), but the arguments are formal in several places. This equation formally reduces for large times to the classical chemotaxis equation (1.5), but to derive the latter rigorously we introduce a parabolic scaling that leads directly from the moment equations (4.7)–(4.10) to the classical chemotaxis equation. For this purpose we define a long time scale, as was done in [19], by setting

$$\hat{t} = \frac{t}{T_p}, \qquad \text{where} \qquad T_p = \frac{1}{\varepsilon^2} T, \tag{7.1}$$

where $T = 1$ sec is the time scale used in the hyperbolic scaling. All other parameters remain the same as in (6.1), (6.6), (6.7), and (6.8), and therefore the dimensionless

equations (6.2)–(6.5) take the form

$$\varepsilon^2 \frac{\partial \hat{n}}{\partial \hat{t}} + \varepsilon \frac{\partial \hat{j}}{\partial \hat{x}} = 0, \tag{7.2}$$

$$\varepsilon^2 \frac{\partial \hat{j}}{\partial \hat{t}} + \varepsilon \hat{s}^2 \frac{\partial \hat{n}}{\partial \hat{x}} = -2\hat{\lambda}_0 \, \hat{j} - 2\,\hat{b}\,\hat{j}_1, \tag{7.3}$$

$$\varepsilon^2 \frac{\partial \hat{n}_1}{\partial \hat{t}} + \varepsilon \frac{\partial \hat{j}_1}{\partial \hat{x}} = -\varepsilon \, \hat{S}'(\hat{x}) \, \hat{j} - \frac{1}{\hat{t}_a} \hat{n}_1, \tag{7.4}$$

$$\varepsilon^2 \frac{\partial \hat{j}_1}{\partial \hat{t}} + \varepsilon \hat{s}^2 \frac{\partial \hat{n}_1}{\partial \hat{x}} = -\varepsilon \hat{s}^2 \, \hat{S}'(\hat{x}) \, \hat{n} - \left(2\hat{\lambda}_0 + \frac{1}{\hat{t}_a} \right) \hat{j}_1 - 2\,\hat{b}\,\hat{j}_2. \tag{7.5}$$

For simplicity, we drop the hats in (7.2)–(7.5), and we consider the case of shallow gradients $S'(x) \sim \mathcal{O}(1)$ as before (see Definition 6.1). Therefore we can use the moment closure $j_2 = 0$ as before, and (7.2)–(7.5) can be written in the following matrix form

$$\varepsilon^2 \frac{\partial v}{\partial t} + \varepsilon \frac{\partial}{\partial x} (Av) = \varepsilon Q(x) v + Rv, \tag{7.6}$$

where v and A are given by (6.19) and

$$Q(x) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -S'(x) & 0 & 0 \\ -s^2 S'(x) & 0 & 0 & 0 \end{pmatrix}$$

$$\text{and} \quad R = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -2\lambda_0 & 0 & -2b \\ 0 & 0 & -\frac{1}{t_a} & 0 \\ 0 & 0 & 0 & -\left(2\lambda_0 + \frac{1}{t_a} \right) \end{pmatrix}.$$

Here all the entries of the matrices A, $Q(x)$, and R are $\mathcal{O}(1)$. Assuming the regular perturbation expansion

$$v = v^0 + \varepsilon v^1 + \varepsilon^2 v^2 + \cdots, \qquad \text{where} \quad v^0 = \begin{pmatrix} n^0 \\ j^0 \\ n_1^0 \\ j_1^0 \end{pmatrix} \quad \text{and} \quad v^1 = \begin{pmatrix} n^1 \\ j^1 \\ n_1^1 \\ j_1 \end{pmatrix};$$

substituting this into (7.6); and comparing terms of equal order in ε, we obtain

$$\varepsilon^0 \ : \qquad Rv^0 = 0, \tag{7.7}$$

$$\varepsilon^1 \ : \qquad \frac{\partial}{\partial x} \left(Av^0 \right) - Q(x) v^0 = Rv^1, \tag{7.8}$$

$$\varepsilon^2 \ : \qquad \frac{\partial v^0}{\partial t} + \frac{\partial}{\partial x} \left(Av^1 \right) - Q(x) v^1 = Rv^2. \tag{7.9}$$

The first equation, (7.7), implies that

$$v_0 = (\, n^0,\, 0\,,\, 0\,,\, 0\,)^T;$$

consequently, the second equation, (7.8), implies

$$j^1 = -\frac{s^2}{2\lambda_0}\frac{\partial n^0}{\partial x} + \frac{bs^2 t_a}{\lambda_0 + 2\lambda_0^2 t_a} S'(x) n^0. \tag{7.10}$$

Finally, (7.9) implies that the left-hand side

$$\frac{\partial v^0}{\partial t} + \frac{\partial}{\partial x}\left(Av^1\right) - Q(x)v^1$$

is in the range of the operator $\mathcal{R} : w \to Rw$. Consequently, using a Fredholm alternative, the left-hand side must be orthogonal to the vector $(1, 0, 0, 0)^T$. Hence,

$$\frac{\partial n^0}{\partial t} + \frac{\partial j^1}{\partial x} = 0.$$

Finally, using (7.10), we derive the classical chemotaxis equation in the following form:

$$\frac{\partial n^0}{\partial t} = \frac{\partial}{\partial x}\left(\frac{s^2}{2\lambda_0}\frac{\partial n^0}{\partial x} - \frac{bs^2 t_a}{\lambda_0 + 2\lambda_0^2 t_a} S'(x) n^0\right). \tag{7.11}$$

Equation (7.11) was derived for the simplified cartoon model (3.7), but a similar analysis can be done for the full cartoon model (2.2)–(2.3). This leads to the classical chemotaxis equation

$$\frac{\partial n}{\partial t} = \frac{\partial}{\partial x}\left(\frac{s^2}{2\lambda_0}\frac{\partial n}{\partial x} - g'(S(x))\frac{bs^2 t_a}{\lambda_0(1 + 2\lambda_0 t_a)(1 + 2\lambda_0 t_e)} S'(x) n\right). \tag{7.12}$$

This is the parabolic counterpart of the hyperbolic equation (6.34) and leads once again to the formula (6.35) for the chemotactic sensitivity. Rigorous estimates on how well the solution of the parabolic equation approximates the solution of the moment equations can be obtained using arguments analogous to those in [19].

8. Numerical examples. The macroscopic descriptions of chemotaxis embodied in either the modified classical chemotaxis equation (6.33) or the classical chemotaxis equation (7.11) are approximations of the original transport equation and the stochastic process describing movement that underlies it. In this section we present two numerical examples that illustrate how well the macroscopic descriptions approximate the solution of the microscopic process. We start by describing our numerical methods.

8.1. Numerical methods. The parameters in our computations are assumed to be dimensionless, and we choose $b = 1$, $t_a = 1$, $\lambda_0 = 1$, and $s = 0.1$; i.e., s is small compared to other parameters (compare with scaling (6.7)–(6.8)).

To solve the system (6.9)–(6.12), closed by (6.17), numerically, we first transform this system to the diagonal form

$$\frac{\partial v}{\partial t} + D_1 \frac{\partial}{\partial x} v = C_1(x) v.$$

Here D_1 is a diagonal 4×4 matrix. Then we use an explicit finite difference method with upwinding. To solve the hyperbolic modified chemotaxis equation (6.33) numerically, we first transform it to the system of two first-order equations in the diagonal form

$$\frac{\partial w}{\partial t} + D_2 \frac{\partial}{\partial x} w = C(x) w,$$

where D_2 is a diagonal 2×2 matrix. Again, we use an explicit finite difference method with upwinding.

To solve the classical chemotaxis equation (7.11) numerically, we use an implicit finite difference method (backward difference approximation in time and centered difference approximation for spatial derivatives).

Finally, to simulate the random walk of individuals, we consider an ensemble of 2000 or 8000 particles. Each particle is described by three variables—position x, velocity $\pm s$, and the internal state y. We use a small time step $dt = 0.01$ (i.e., the unbiased turning frequency divided by 100). During each time step the particle moves with speed s in the chosen direction, and we integrate the internal dynamics to find the change of y. At the end of each time step, a random number from $[0, 1]$ is generated and compared with the probability of the turn $\lambda(y)\mathrm{d}t$. If the turn occurs, the bacterium will move during the next time step in the opposite direction.[1]

8.2. Traveling bands. In this example we analyze the motion of the individuals in the interval $[0, 20]$ with the signal $S(x)$ given by

$$S(x) = 28 - 2|x - 14|. \tag{8.1}$$

The signal has a global maximum at the point 14, and its derivative is $S'(x) = -2\,\mathrm{sign}\,(x-14)$ for $x \neq 14$. We assume the same initial condition for all computations, namely,

$$n(x,0) = \begin{cases} 1 & \text{for } x \in [5,6], \\ 0 & \text{for } x \notin [5,6]; \end{cases} \tag{8.2}$$

we assume that all individuals are perfectly adapted at $t = 0$; and we use no-flux boundary conditions.

In Figure 8.1 we compare the results of the stochastic simulation of the random walk with the solutions of the macroscopic system (6.9)–(6.12) closed by (6.17). It happens that the solution of the modified chemotaxis equation (6.33) and the solution of the classical chemotaxis equation (7.11) are indistinguishable on the plots from the solution of (6.9)–(6.12). Thus the macroscopic results presented can be viewed as plots of the solution of any of these macroscopic equations.

In Figure 8.1 we see that the band travels to the right (i.e., toward the maximum of the signal), as expected, and then the individuals who arrive at the maximum first aggregate there. Eventually all individuals aggregate around the maximum of the signal. From the plots we also see that numerically the macroscopic equations give very good results in comparison with the Monte Carlo simulations. Finally, if we use the results for the time interval $[0, 600]$, i.e., under the influence of the constant gradient, we can compute the average speed of the bacteria in this interval and find that

$$V \doteq \frac{1}{150} = \frac{|b|s^2 S'(x) t_a}{\lambda_0 + 2\lambda_0^2 t_a},$$

which agrees with the result in (5.7).

[1] A Monte Carlo simulation that incorporates the internal dynamics used here, as well as a more detailed description of the motor behavior, is given in [39].

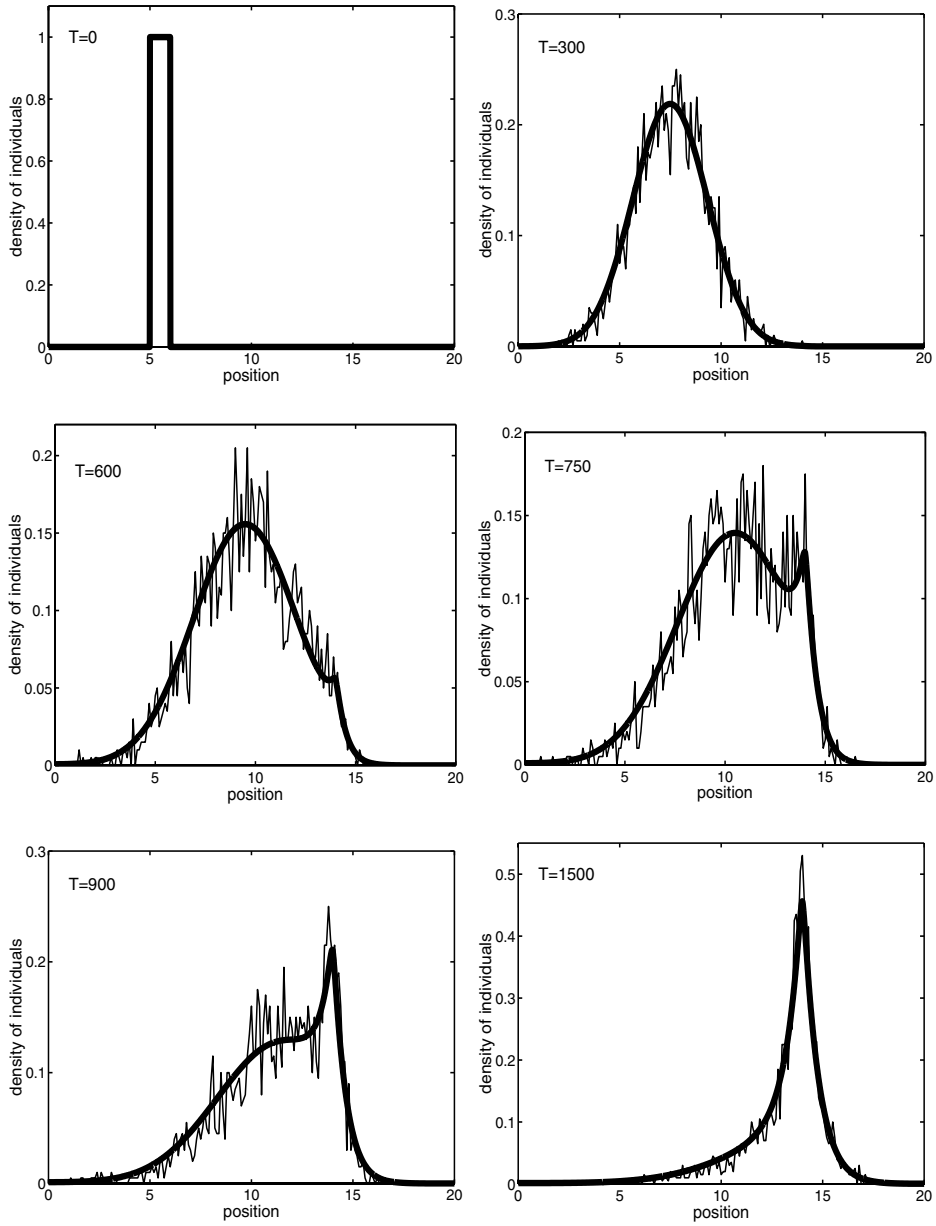


FIG. 8.1. *The graphs show the solutions of the macroscopic system* (6.9)–(6.12) *closed by* (6.17) *(thick smooth line) and the results of stochastic simulations of the velocity jump process with internal state variables (thin line). Moreover, the thick line can also be viewed as a solution of the modified chemotaxis equations* (6.33) *and the solution of the classical chemotaxis equation* (7.11)*, since the solutions of* (6.33)*,* (7.11)*, and* (6.9)–(6.12) *closed by* (6.17) *are indistinguishable on this scale. We used* 2000 *particles for the Monte Carlo simulations, and the parameters* $b = 1$, $t_a = 1$, $\lambda_0 = 1$, *and* $s = 0.1$.

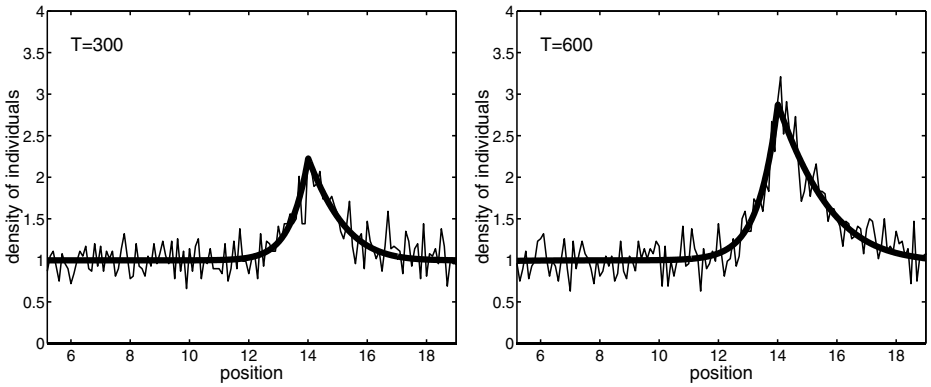


FIG. 8.2. *The graphs of the density of bacteria under the influence of the exponential signal function* (8.3) *with* $x_0 = 14$ *and* $C = 1$. *Initially the bacteria were uniformly distributed at a density equal to* 1. *The figure shows the density of bacteria at time* $T = 300$ *(left) and* $T = 600$ *(right). The bacteria aggregate around the point* $x = 14$, *as expected. In this figure we use the internal dynamics* (2.2)–(2.5) *with* $g(S) = 2\ln(S)$, $t_a = b = \lambda_0 = 1$, $t_e = 0$, *and* $s = 0.1$. *The solution of the classical chemotaxis equation* (7.11) *(thick line) and the Monte Carlo simulation (thin line) are shown.*

8.3. Exponential signal ramp. Various chemotaxis experiments have been done with exponential signal functions [14, 43]. The standard setup is that initially there is a uniform concentration of bacteria in the medium with an exponential signal ramp of the form

$$S(x) = \begin{cases} Ce^x & x \le x_0, \\ Ce^{x_0} & x \ge x_0. \end{cases} \tag{8.3}$$

After several minutes, bacteria aggregate at the top of the exponential ramp, i.e., around the point x_0.

As a second numerical example and test of the macroscopic equations, we reproduce these experiments with the exponential signal ramp. We again use here the internal dynamics (2.2)–(2.5) with a suitable choice of g. The exponential signal ramp was used experimentally because a cell swimming in one direction sees a constant *rate of increase of the signal*, and therefore the bias should remain approximately constant. To take this into account, we could choose

$$g(S) = C\frac{S(x)}{K_D + S(x)}. \tag{8.4}$$

However, as we only want to reproduce the experimental results qualitatively, we can approximate (8.4) by the logarithmic function. The numerical results for this are shown in Figure 8.2. We plot the solution of the classical chemotaxis equation (7.11), but the solutions of (6.9)–(6.12) and (6.33) again give the same results. Moreover, we also qualitatively reproduce the behavior observed in experiments (cf. [14, 43]).

9. Extensions of the analysis. In this section, we discuss two extensions of our analysis—the inclusion of a finite-duration tumbling state and the moment closure for arbitrary signal gradients.

9.1. Inclusion of a finite-duration tumbling state. As we mentioned earlier, the movement of *E.coli* consists of "running" smoothly with a speed s and "tumbling"

randomly. Tumbles cause the bacterium to reorient and swim in a new random direction. The duration of both runs and tumbles are exponentially distributed with means of 1 sec and 10^{-1} sec, respectively, in the absence of an extracellular signal. Thus cells spend 10 percent of the total time in the tumbling state. Earlier we neglected this time by assuming an instantaneous reversal of direction, but we now include it. We again restrict the analysis to one space dimension, since the generalization to higher dimensions only introduces some technical issues. We denote by

- $p^0(x, y, t)$ the number density of tumbling bacteria at time t and at point x with internal state y;
- $p^{\pm}(x, y, t)$ the number density of bacteria running to the *right* (resp., *left*) at time t and at point x with internal state y.

Suppose that a cell with internal state y moves along the x-axis at a constant speed s and at random instants of time stops with stopping time governed by a Poisson process of intensity $\alpha(y)$, and that a cell with internal state y tumbling at the point x to move at random instants of time starts according to a Poisson process with the intensity $\beta(y)$. Further, suppose that the direction of movement is unbiased, i.e., that the tumbling particle will go with probability 0.5 to the right and with probability 0.5 to the left, given that movement starts.

For simplicity we consider the simplified cartoon internal dynamics (3.7)–(3.8). Using the change of internal variables $y_2 = S(x) + z_2$, the movement of bacteria can be described by the following equations:

$$\frac{\partial p^+}{\partial t} + s\frac{\partial p^+}{\partial x} + \frac{\partial}{\partial z_2}\left[\left(-\frac{z_2}{t_a} - sS'(x)\right)p^+\right] = -\alpha(y)p^+ + \frac{1}{2}\beta(y)p^0, \tag{9.1}$$

$$\frac{\partial p^0}{\partial t} + \frac{\partial}{\partial z_2}\left[\left(-\frac{z_2}{t_a}\right)p^0\right] = \alpha(y)(p^+ + p^-) - \beta(y)p^0, \tag{9.2}$$

$$\frac{\partial p^-}{\partial t} - s\frac{\partial p^-}{\partial x} + \frac{\partial}{\partial z_2}\left[\left(-\frac{z_2}{t_a} + sS'(x)\right)p^-\right] = -\alpha(y)p^- + \frac{1}{2}\beta(y)p^0. \tag{9.3}$$

In order to compare this model with the previous one, we will specify $\alpha(y)$ and $\beta(y)$ as follows:

$$\alpha(y) = 2\lambda_0 + 2bz_2, \qquad \beta(y) = \beta_0 - \beta_1 z_2, \tag{9.4}$$

where

$$\lambda_0 > 0, \quad b > 0, \quad \beta_0 > 0, \quad \beta_1 \geq 0.$$

Then this model is equivalent to the model in section 4 in the limit $\beta_0 \to \infty$. One can show, using techniques similar to those used before, that the average position of the particles under the influence of a constant gradient $S'(x) = C$ is given by (cf. (5.7))

$$\langle x\rangle(t) = \frac{bs^2Ct_a}{\lambda_0 + 2\lambda_0^2 t_a}\left(\frac{\beta_0}{2\lambda_0 + \beta_0}\right)t. \tag{9.5}$$

Thus the tumbling state slows down the movement by the factor $\beta_0/(2\lambda_0 + \beta_0)$, and we recover (5.7) as $\beta_0 \to \infty$. Moreover, one can derive the following modified classical chemotaxis equation (cf. (6.33)):

$$\frac{\partial^2 n}{\partial t^2} + 2\lambda_0\frac{\partial n}{\partial t} = \frac{\partial}{\partial x}\left(s^2\frac{\partial n}{\partial x} - \frac{2s^2bt_a}{1 + 2\lambda_0 t_a}\left(\frac{\beta_0}{2\lambda_0 + \beta_0}\right)S'(x)n\right).$$

9.2. Moment closure for arbitrary signal gradients. Heretofore we have used the approximation (6.17), which is appropriate for shallow signal gradients. The question arises as to what can be done for large signal gradients, i.e., for signals that satisfy the standing assumption (3.12). Can we also find a moment closure of the form (6.13)?

To do that, we have to approximate the neglected term

$$2bj_2 = 2bs\int_{\mathbb{R}}(z_2)^2\big[p^+(x,z_2,t)-p^-(x,z_2,t)\big]\mathrm{d}z_2 = 2b\int_{\mathbb{R}}(z_2)^2\mathcal{J}(x,z_2,t)\mathrm{d}z_2. \tag{9.6}$$

Recall that the internal variable $z_2 \in \mathbb{R}$ evolves according to (3.10), i.e., according to the differential equation

$$\frac{\mathrm{d}z_2}{\mathrm{d}t} = -\frac{z_2}{t_a} \mp S'(x)s, \tag{9.7}$$

where the sign of the last term is determined by the sign of the velocity of the particle. Equation (9.7) suggests that we can assume $z_2 \approx \mp S'(x)st_a$. This is simply an assumption, but it leads to the following two naive moment closures:

$$\begin{aligned}
2bj_2 &= 2bs\int_{\mathbb{R}}(z_2)^2\big[p^+(x,z_2,t)-p^-(x,z_2,t)\big]\mathrm{d}z_2\\
&= 2bs\int_{\mathbb{R}}(z_2)(z_2)p^+(x,z_2,t)\mathrm{d}z_2 - 2bs\int_{\mathbb{R}}(z_2)(z_2)p^-(x,z_2,t)\mathrm{d}z_2\\
&\doteq 2bs\int_{\mathbb{R}}(-S'(x)st_a)(z_2)p^+(x,z_2,t)\mathrm{d}z_2 - 2bs\int_{\mathbb{R}}(S'(x)st_a)(z_2)p^-(x,z_2,t)\mathrm{d}z_2\\
&= -2bS'(x)t_as^2\int_{\mathbb{R}}z_2\big[p^+(x,z_2,t)+p^-(x,z_2,t)\big]\mathrm{d}z_2 = -2bS'(x)t_as^2n_1, \qquad (9.8)
\end{aligned}$$

$$2bj_2 = 2b\int_{\mathbb{R}}(z_2)^2\mathcal{J}(x,z_2,t)\mathrm{d}z_2 \doteq 2b\int_{\mathbb{R}}(\mp S'(x)st_a)^2\mathcal{J}(x,z_2,t)\mathrm{d}z_2 = 2b(S'(x))^2t_a^2s^2j. \tag{9.9}$$

These are both consistent with the moment closure (6.17) for shallow gradients of the signal, and consequently they lead to the same equations (6.33) and (7.11) in that case. On the other hand, they are much better than (6.17) for arbitrary signal gradients, which we illustrate here numerically.

To this end, suppose that the derivative of the signal function $S(x)$ is constant, i.e., $S'(x) = C$ and, as in section 5, we derive the corresponding average velocity of the individuals V (i.e., the average velocity which is approached asymptotically; cf. (5.7)). Surprisingly, the result is the same for both moment closures (9.8) and (9.9), namely,

$$V = \frac{bt_as^2S'(x)}{\lambda_0 + 2\lambda_0^2t_a - 2b^2t_a^3s^2(S'(x))^2}.$$

As in section 8, we set $b = \lambda_0 = t_a = 1$ and $s = 0.1$, and then have $\overline{C} = 10$ and

$$V = \frac{S'(x)}{300 - 2(S'(x))^2} \tag{9.10}$$

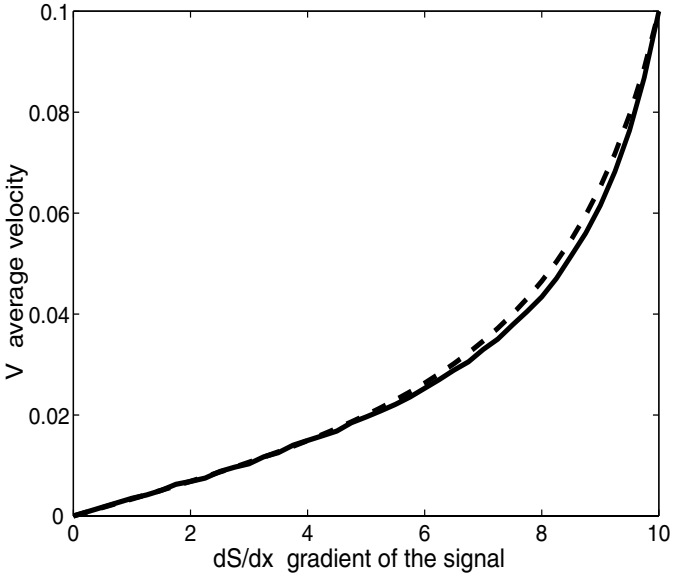


FIG. 9.1. *The dashed line shows the average velocity V given by the formula* (9.10) *as a function of the gradient of the signal. The solid line presents the average velocities of the bacteria obtained by the stochastic simulation.*

for $S'(x) \in [-10, 10]$. To verify this formula numerically, we make several stochastic simulations of the velocity jump process with internal variables, with the same parameters $b = \lambda_0 = t_a = 1$ and $s = 0.1$ for the constant gradients of the signal from the interval [0,10].

Figure 9.1 shows the graph of V (dashed line) as a function of gradient $S'(x)$. The solid line presents the velocities obtained by the stochastic simulation of the velocity jump process with internal variables. We see that, for the maximum possible signal function $S'(x) = 10$, all individuals move to the right with the speed $s = 0.1$, as expected. Moreover, we also see that the macroscopic equations obtained by the moment closure (9.8) or by moment closure (9.9) can give good macroscopic moment equations when used in (6.9)–(6.12).

10. Discussion and conclusions. We have shown how information about microscopic intracellular processes such as signal transduction and response can be translated into the macroscopic chemotactic sensitivity that appears in the macroscopic description of chemotaxis. This was done for a highly simplified description of intracellular dynamics, one which is based on linear dynamics for the response to an extracellular signal, but which nonetheless incorporates the two most important characteristics of any detailed signal transduction network, namely, excitation and adaptation. Linear dynamics and linear response may well be adequate for describing the type of signal changes a swimming bacterium normally sees, but that remains to be established. In addition, a great deal of further work is needed to identify the essential response modes in a general signal transduction network, even if a near-equilibrium assumption is used. A difficult part of that will be to determine how the extracellular signal feeds into the linearized response of the cell.

The moment approach used here leads firstly to a system of hyperbolic equations, and then via a hyperbolic (resp., parabolic) scaling of space and time to a single

hyperbolic (6.34) (resp., parabolic (7.12)) equation for the density of individuals. One can also use other scalings of space and time. For example, the parabolic scaling uses $T \sim \mathcal{O}(1/\varepsilon^2)$, but if one uses $T \sim \mathcal{O}(1/\varepsilon^3)$, the result is the elliptic equation for the steady states of (6.34).

The first systematic derivation of a chemotaxis equation from a velocity jump process is due to Patlak [36], who considers both internal and external biases in detail, but these biases are imposed. A basic assumption in [36] is that the run length is chosen and fixed whenever the particle turns, which is quite different from the stochastic process treated here. As was observed elsewhere [30], the particle motion between turns is deterministic, and thus, were the speed and run length constant, the process would be formally equivalent to a space jump process [29]. In general one can show that this process leads to a renewal equation that generalizes the renewal equation (15) derived in [29], from which a diffusion equation is obtained by suitable choice of the waiting time and jump distributions. Others have treated a process similar to the one treated here without the internal dynamics since Patlak's work, and the reader is referred to [30] for a review of the literature.

REFERENCES

[1] W. Alt, *Biased random walk model for chemotaxis and related diffusion approximation*, J. Math. Biol., 9 (1980), pp. 147–177.

[2] N. Barkai and S. Leibler, *Robustness in simple biochemical networks*, Nature, 387 (1997), pp. 913–917.

[3] J. Bartak, L. Herrmann, V. Lovicar, and O. Vejvoda, *Partial differential equations of evolution*, Ellis Horwood, Chichester, England, 1991.

[4] H. Berg, *How bacteria swim*, Scientific American, 233 (1975), pp. 36–44.

[5] H. Berg, *Random Walks in Biology*, University Press, Princeton, USA, 1983.

[6] H. C. Berg, *Bacterial microprocessing*, Cold Spring Harbor Symp. Quant. Biol., 55 (1990), pp. 539–545.

[7] H. C. Berg and D. A. Brown, *Chemotaxis in Escherichia coli analysed by three-dimensional tracking*, Nature, 239 (1972), pp. 500–504.

[8] S. M. Block, J. E. Segall, and H. C. Berg, *Impulse responses in bacterial chemotaxis*, Cell, 31 (1982), pp. 215–226.

[9] S. M. Block, J. E. Segall, and H. C. Berg, *Adaptation kinetics in bacterial chemotactics*, J. Bacteriol., 154 (1983), pp. 312–323.

[10] R. B. Bourret, K. A. Borkovich, and M. I. Simon, *Signal transduction pathways involving protein phosphorylation in prokaryotes*, Ann. Rev. Biochem., 60 (1991), pp. 401–441.

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

[12] K. C. Chen, R. M. Ford, and P. T. Cummings, *Cell balance equation for chemotactic bacteria with a biphasic tumbling frequency*, J Math. Biol., 47 (2003), pp. 518–546.

[13] P. Cluzel, M. Surette, and S. Leibler, *An ultrasensitive bacterial motor revealed by monitoring signaling proteins in single cells*, Science, 287 (2000), pp. 1652–1655.

[14] F. Dahlquist, P. Lovely, and K. D. E., *Quantitative analysis of bacterial migration in chemotaxis*, Nature New Biol., 236 (1972), pp. 120–123.

[15] R. Erban and H. G. Othmer, *From signal transduction to spatial pattern formation in* E. coli*: A paradigm for multi scale modeling in biology*, Multiscale Model. Simul., to appear.

[16] S. Goldstein, *On diffusion by discontinuous movements and the telegraph equation*, Quart. J. Mech. Appl. Math., 4 (1951), pp. 129–156.

[17] G. J. Habetler and B. Matkowsky, *Uniform asymptotic expansion in transport theory with small mean free paths, and the diffusion approximation*, J. Math. Phys., 4 (1975), pp. 846–854.

[18] R. Hersh, *Random evolutions: A survey of results and problems*, Rocky Mountain J. Math., 4 (1974), pp. 443–477.

[19] T. Hillen and H. G. Othmer, *The diffusion limit of transport equations derived from velocity-jump processes*, SIAM J. Appl. Math., 61 (2000), pp. 751–775.

[20] T. Hillen and A. Stevens, *Hyperbolic models for chemotaxis in* 1*-D*, Nonlinear Anal. Real World Appl., 1 (2000), pp. 409–433.

[21] D. Horstmann, *From* 1970 *until present: The Keller–Segel model in chemotaxis and its consequences* I, Jahresber. Deutsch. Math.-Verein., 105 (2003), pp. 103–165.
[22] M. Kac, *A stochastic model related to the telegrapher's equation*, Rocky Mountain J. Math., 3 (1974), pp. 497–509.
[23] E. Keller and L. Segel, *Initiation of slime mold aggregation viewed as an instability*, J. Theoret. Biol, 26 (1970), pp. 399–415.
[24] S. Khan, F. Castellano, J. Spudich, J. McCray, R. Goody, G. Reid, and D. Trentham, *Excitatory signaling in bacteria probed by caged chemoeffectors*, Biophys. J., 65 (1993), pp. 2368–2382.
[25] D. E. Koshland, *Bacterial Chemotaxis as a Model Behavior System*, Raven Press, New York, 1980.
[26] E. W. Larsen and J. B. Keller, *Asymptotic solution of neutron transport problems for small free mean paths*, J. Math. Phys., 15 (1974), pp. 75–81.
[27] H. McKean, *Chapman–Enskog–Hilbert expansions for a class of solutions of the telegraph equation*, J. Math. Phys., 75 (1967), pp. 1–10.
[28] C. J. Morton-Firth, T. S. Shimizu, and D. Bray, *A free-energy-based stochastic simulation of the Tar receptor complex*, J. Molec. Biol., 286 (1999), pp. 1059–1074.
[29] H. Othmer, S. Dunbar, and W. Alt, *Models of dispersal in biological systems*, J. Math. Biol., 26 (1988), pp. 263–298.
[30] H. G. Othmer and T. Hillen, *The diffusion limit of transport equations* II: *Chemotaxis equations*, SIAM J. Appl. Math., 62 (2002), pp. 1222–1250.
[31] H. G. Othmer and A. Stevens, *Aggregation, blowup, and collapse: The ABC's of taxis in reinforced random walks*, SIAM J. Appl. Math., 57 (1997), pp. 1044–1081.
[32] H. G. Othmer and P. Schaap, *Oscillatory cAMP signaling in the development of Dictyostelium discoideum*, Comments on Theoret. Biol., 5 (1998), pp. 175–282.
[33] K. J. Painter, P. K. Maini, and H. G. Othmer, *Development and applications of a model for cellular response to multiple chemotactic cues*, J. Math Biol., 41 (2000), pp. 285–314.
[34] G. C. Papanicolaou, *Asymptotic analysis of transport processes*, Bull. Amer. Math. Soc., 81 (1975), pp. 330–392.
[35] E. Pate and H. G. Othmer, *Differentiation, cell sorting and proportion regulation in the slug stage of Dictyostelium discoideum*, J. Theoret. Biol., 118 (1986), pp. 301–319.
[36] C. Patlak, *Random walk with persistence and external bias*, Bull. Math. Biophys., 15 (1953), pp. 311–338.
[37] E. M. Purcell, *Life at low Reynolds number*, Am. J. Phys., 45 (1977), pp. 3–11.
[38] J. E. Segall, S. M. Block, and H. C. Berg, *Temporal comparisons in bacterial chemotaxis*, Proc. Natl. Acad. Sci. USA, 83 (1986), pp. 8987–8991.
[39] S. Setayeshgar, C. W. Gear, H. G. Othmer, and I. G. Kevrekidis, *Application of coarse integration to bacterial chemotaxis*, Multiscale Model. Simul., to appear.
[40] P. Spiro, J. Parkinson, and H. Othmer, *A model of excitation and adaptation in bacterial chemotaxis*, Proc. Natl. Acad. Sci. USA, 94 (1997), pp. 7263–7268.
[41] R. C. Stewart and F. W. Dahlquist, *Molecular components of bacterial chemotaxis*, Chem. Rev., 87 (1987), pp. 997–1025.
[42] J. B. Stock and M. G. Surette, *Chemotaxis*, in Escherichia coli and Salmonella: Cellular and Molecular Biology, F. C. Neidhardt and R. Curtiss, eds., ASM Press, Washington, DC, 1996, Vol. I, pp. 1103–1129.
[43] R. M. Weis and D. E. Koshland, *Reversible receptor methylation is essential for normal chemotaxis of Escherichia coli in gradients of aspartic acid*, Proc. Natl. Acad. Sci. USA, 85 (1988), pp. 83–87.

SIAM J. APPL. MATH.
Vol. 65, No. 2, pp. 392–425

SYNCHRONY IN A POPULATION OF HYSTERESIS-BASED GENETIC OSCILLATORS*

ALEXEY KUZNETSOV†, MADS KÆRN‡, AND NANCY KOPELL†

Abstract. Oscillatory behavior has been found in different specialized genetic networks. Previous work has demonstrated nonsynchronous, erratic single-cell oscillations in a genetic network composed of nonspecialized regulatory components and based entirely on negative feedback. Here, we present the construction of a more robust, hysteresis-based genetic relaxation oscillator and provide a theoretical analysis of the conditions necessary for single-cell and population synchronized oscillations. The oscillator is constructed by coupling two subsystems that have previously been implemented experimentally. The first subsystem is the toggle switch, which consists of two mutually repressive genes and can display robust switching between bistable expression states and hysteresis. The second subsystem is an intercell communication system involved in quorum-sensing. This subsystem drives the toggle switch through a hysteresis loop in single cells and acts as a coupling between individual cellular oscillators in a cell population. We demonstrate the possibility of both population synchronization and suppression of oscillations (cluster formation), depending on diffusion strength and other parameters of the system. We also propose the optimal choice of the parameters and small variations in the architecture of the gene regulatory network that substantially expand the oscillatory region and improve the likelihood of observing oscillations experimentally.

Key words. coupled oscillators, relaxation oscillations, stability, intercell communication, gene networks

AMS subject classifications. 34C15, 34C26, 92D10

DOI. 10.1137/S0036139903436029

1. Introduction.

The variation in gene expression in response to internal or external signals is one of the most important means of cellular regulation. The rate at which a gene is transcribed into messenger RNA and subsequently translated into protein is influenced by many factors but is primarily controlled by how well the RNA polymerase complex can bind to and initiate transcription from a regulatory region of the DNA called the promoter. Signals that modulate transcription are often mediated through transcription factor proteins that bind to target sites within or near the promoter where they increase (activation) or decrease (repression) the probability of RNA polymerase complex binding and/or initiation of transcription. The manipulation of DNA sequence to create novel promoters containing customized transcription factor target sites and to mix and match such promoters with genes that encode the corresponding transcription factor proteins has allowed the construction of artificial gene regulatory networks with customizable functionality [1, 2, 3, 4, 5]. Such networks can be used to achieve complex and multifaceted control of cellular function and have promising scientific, medicinal, and biotechnological applications [6, 7, 8].

*Received by the editors October 8, 2003; accepted for publication (in revised form) May 12, 2004; published electronically December 16, 2004.
http://www.siam.org/journals/siap/65-2/43602.html

†Center for BioDynamics and Mathematical Department, Boston University, 111 Cummington St., Boston, MA 02215 (alexey@bu.edu, nk@bu.edu). The work of the first author was partially supported by NSF grant DMS-0109427. The work of the third author was partially supported by NSF grant DMS-0211505.

‡Center for BioDynamics and Biomedical Engineering Department, Boston University, 44 Cummington St., Boston, MA 02215 (mkaern@bu.edu). The work of this author was partially supported by NSF Bio-QuBIC Program grant EIA-0130331 and The Defense Advanced Research Projects Agency grant F30602-01-2-0579.

Mathematical modeling and analysis is becoming increasingly important as a tool to organize and to interpret vast amounts of experimental data and as a predictive tool in the construction of artificial gene networks [1, 2, 9]. The first step in the construction of an artificial gene network is to investigate if the proposed network architecture supports the desired functionality. In this paper, we use mathematical techniques to model and analyze an artificial gene network that is currently being implemented experimentally in the bacterium *Escherichia coli* [10]. The network is intended to regulate a population synchronous periodic oscillation in the levels of cellular protein in a constant density, well-stirred bio-reactor. The goals of the mathematical analysis are (1) to investigate if the network architecture supports single-cell and population synchronous oscillations, (2) to identify the parameter values and the experimental conditions where this behavior is supported, and (3) to suggest modifications to the network that optimize the robustness of single-cell and population oscillations.

The oscillator is to be constructed by combining two engineered gene networks that have previously been implemented experimentally in *E. coli*: the toggle switch [1] and an intercell communication system [11, 12, 13]. The engineered gene networks are carried on multicopy, self-replicating plasmids that interfere minimally with the host cell. As a result, the dynamics of the engineered networks may be considered independently of the dynamics of the cell's natural regulatory circuitry. The toggle switch is composed of two transcription factor proteins: the lac repressor, encoded by the gene *lacI*, and a temperature-sensitive variant of the λ *cI* repressor, encoded by the gene *cI857*. The synthesis of the two repressor proteins is regulated in such a way that expression of the *cI857* and *lacI* genes are mutually exclusive: The promoter P_{trc} that controls the expression of *cI857* is attenuated by the lac repressor while the promoter P_{L*} that controls the expression of *lacI* is attenuated by the λ repressor. Thus, a cell can be either in a state where λ repressor is abundant and lac repressor scarce (the *cI* on state) or in a state where lac repressor is abundant and λ repressor scarce (the *lacI* on state).

It has been demonstrated experimentally [1] that the state of cells harboring the toggle switch network can be changed permanently by transient inactivation of the dominant repressor protein. The λ repressor is inactivated at elevated temperature and the state can be changed from the *cI* ON to the *lacI* ON state by a transient increase in temperature. Conversely, the transition from the *lacI* ON to *cI* ON state ensues when the lac repressor protein is inactivated by the addition of sufficient amounts of the chemical isopropyl-β-D-thiogalactopyranoside (IPTG). At subcritical levels of IPTG, both states exist and are stable. Variation in IPTG has demonstrated hysteresis of these two steady states [1]. We exploit the presence of hysteresis in the toggle switch to construct an oscillator network by linking the toggle switch to a second network that autonomously drives cells through the hysteresis loop (Figure 1.1(A)).

The gene network intended to drive the oscillation involves components of the quorum-sensing system from *Vibrio fischeri* [14]. Quorum-sensing enables cells to sense population density through a transcription factor protein LuxR, which acts as a transcriptional activator of genes expressed from the P_{lux} promoter when a small organic molecular, the autoinducer (AI), binds to it. The AI is synthesized by the protein encoded by the gene *luxI*, and the AI can diffuse across the cell membrane causing the extracellular concentration of AI, as well as the AI concentration in individual cells, to depend on the density of AI-producing cells. These properties of the quorum-sensing system has been exploited experimentally to construct biosensors (see, e.g., [15, 16, 17]) to transfer information from one cell to another [11] and can,

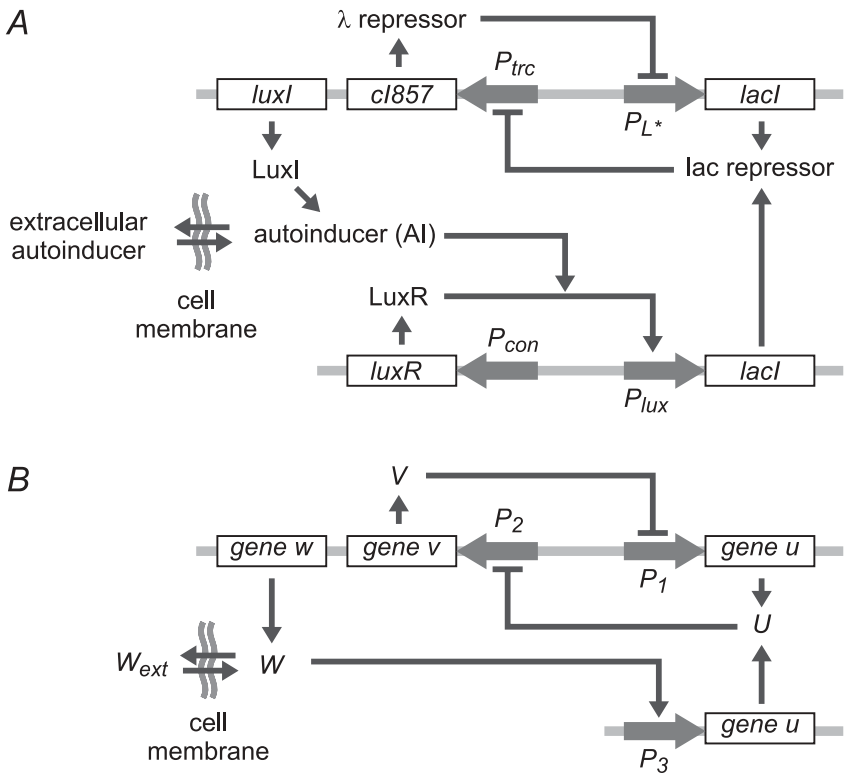


FIG. 1.1. *Schematic diagrams of the genetic oscillator network in isolated cells.* (A) *Full network composed of the toggle switch genes* cI857 *and* lacI *and their respective promoters* P_{trc} *(lac repressed) and* P_{L*} *(*λ *repressed) and the quorum-sensing genes* luxI *and* luxR*. Autoinducer* (AI) *is synthesized by the LuxI protein and activates expression of* lacI *from the* P_{lux} *promoter through binding to the LuxR transcription factor.* (B) *Minimal model. It is assumed that the activation of expression of the repressor gene u (*lacI*) occurs in a single step binding of W (AI) to the promoter* P_3 *(*P_{lux}*).*

in theory, be used to achieve synchronization across a cell population [18]. It was recently shown experimentally [10] that a variant of the network in Figure 1.1 lacking the *luxI* gene can respond to AI and be driven through a saddle-node bifurcation by increasing AI concentration.

This paper is organized as follows. In section 2, we discuss the structure of the genetic network and derive the equations that govern the dynamics of a minimal description of the network. In section 3, we investigate the dynamics of isolated cells and establish the condition for the organization of single-cell oscillations. In section 5, we consider the dynamics of an ensemble of cells and demonstrate the possibility of both population synchronization and suppression of oscillations, depending on diffusion strength and other parameters of the system. In section 4, we investigate the effect on the ability of isolated cells to oscillate when molecular details left out of the minimal model are taken into consideration. In particular, production of the autoinducer in two steps, rather than just one as assumed in the minimal model, improves the likelihood of observing oscillations experimentally. We also show that oscillatory behavior can be made much more robust by adding an additional connectivity to the network.

2. Minimal model. The molecular details of the genetic oscillator network that we wish to construct is illustrated schematically in Figure 1.1(A). It is a slight variant of a network constructed by Kobayashi et al. [10], where *luxI* is inserted downstream of *cI857* rather than downstream of *luxR*. The expression of the *cI857* is controlled by the promoter P_{trc}, and the expression of the *lacI* gene is controlled by the P_{L*} promoter. As a result, the AI is synthesized when the cell is in the *cI* ON/*lacI* OFF state. The AI binds to the LuxR protein, whose gene is expressed at a constant rate from the P_{con} promoter, and the LuxR-AI complex increases the rate of expression of the *lacI* gene by activation of the P_{lux} promoter. Hence, when cells are in the *lacI* OFF state, the AI will gradually accumulate and activate the production of λ repressor protein. The λ repressor eventually shuts down expression of *cI* and *luxI* from the P_{trc} promoter, causing a transition from the *cI* ON to the *lacI* ON state and a down-regulation of AI production. To complete the cycle, it is required that a cell returns to the *cI* ON state once AI production ceases in the *lacI* ON state. Therefore, oscillations require that the toggle switch component of the network is bistable at intermediate levels of AI and monostable when the level of AI is either high (*lacI* ON) or low (*cI* ON).

To ease the mathematical analysis, we initially employ a simplified model of the full system, illustrated in Figure 1.1(B). In this model, the promoters are renamed P_1 (P_{L*}), P_2 (P_{trc}), and P_3 (P_{lux}) and the transcription factors renamed U (lac repressor), V (λ repressor), and W (the LuxR-AI activator). The difference between the full (Figure 1.1(A)) and the simplified system (Figure 1.1(B)) lies in the regulation of the P_3 promoter. We assume for simplicity that the activator of P_3 is encoded by a single gene w rather than being the complex between LuxR and the AI. This assumption ignores a potential time delay introduced by the two-step synthesis of the LuxR-AI complex (i.e., LuxI $\rightarrow$ AI $\rightarrow$ LuxR-AI) and the titration and saturation of free LuxR by the AI. The effects of these assumptions on oscillations in single cells are investigated further in section 4.

2.1. Regulation of gene expression. The simplest model of gene expression involves only two steps: the transcription of a gene into mRNA and the translation of the mRNA into protein [19]. Consider the expression of a gene x that encodes the protein X and is regulated by the promoter P. When each cell harbors n_A active promoters from which the mRNA of gene x is transcribed at an average rate k, the approximation of the rate of mRNA change gives us the following differential equation:

$$\frac{dn_m}{dt} = n_A k - d_m n_m, \tag{2.1}$$

where d_m is the effective first-order rate constant associated with degradation of the mRNA within cells. This equation is, of course, only an approximation since it assumes that the number of mRNA molecules is continuous rather than discrete and since many additional steps are involved in both transcription and degradation of mRNA [21]. Messenger RNA molecules are usually degraded rapidly compared to other cellular processes, and it is often assumed that the concentration of mRNA rapidly reaches a pseudo-steady state where $n_m = n_A k/d_m$ such that dn_m/dt is zero. In some cases, the delay introduced by mRNA synthesis is important for oscillatory dynamics [2]. However, mRNA half-lives are difficult to manipulate experimentally, which makes it difficult to exploit these control parameters in vivo.

The mRNA is translated into a protein by ribosomes, and it is assumed that each x mRNA molecule gives rise to $b_x = k_{tl,x}/d_m$ copies of the protein X, where

$k_{tl,x}$ is the averaged translation rate. The parameter b_x is referred to as the burst parameter of the protein and depends on the efficiency of translation and the mRNA half-life [19]. The value of the translational efficiency depends, among several factors, on the nucleotide sequence of the ribosome binding sites (RBS) located within the upstream, noncoding part of the mRNA. The RBS is encoded by the DNA sequence immediately upstream of the start codon of the gene and is an independent regulatory element that can be manipulated experimentally. The sequence of the DNA encoding the RBS is one of the principal tools by which the parameters of an engineered gene network can be adjusted (see, e.g., [1]).

The equation that governs the evolution of the number of proteins, n_X, produced from n_m mRNA molecules is in the continuous approximation given by

$$\frac{dn_X}{dt} = k_{tl,x} n_m - k_X n_X, \tag{2.2}$$

where $k_{tl,x}$ is the averaged translation rate, introduced above and k_X is the effective first-order rate constant associated with the degradation of the protein within cells. When a pseudo-steady state approximation is invoked for mRNA ($n_m = n_A k/d_m$), it is obtained that $k_{tl,x} n_m = k_{tl,x} n_A k/d_m = b_x n_A k$. The equation for the number of proteins takes the form

$$\frac{dn_X}{dt} = b_x n_A k - k_X n_X. \tag{2.3}$$

The rate of protein decay, k_X, is a second experimental control parameter that can be altered by augmenting, or tagging, the protein with additional amino acids, which makes the protein a target of proteases that break down the protein into amino acids [2].

It is convenient to convert the equation for the evolution of the total number of proteins per cell into an equation for the evolution of cellular protein concentration, $[X](t) = n_X(t)/v(t)$, where $v(t)$ is the cell volume. Cells divide at regular time intervals T, and the cell volume is assumed to increase exponentially in accordance with the growth law $v(t) = v_0 \exp(k_g t)$, where v_0 is the cell volume immediately after division and $k_g = \ln(2)/T$. Cell division occurs when $v(t = T) = 2v_0$. The evolution equation for protein concentration is then obtained as

$$\frac{d[X]}{dt} = \frac{1}{v(t)} \left(\frac{dn_X}{dt} - [X] \frac{dv(t)}{dt} \right) = b_x k [A](t) - (k_X + k_g)[X], \tag{2.4}$$

where $[A](t)$ is the concentration of active promoters, $[A](t) = n_A(t)/v(t)$. It is noted that an exponential increase in cell volume is only an approximation of the quite complicated process of cell growth and division.

The concentration of active promoters, $[A](t)$, depends on the concentration of transcription factors that are bound to the promoter region at a given time. Consider the formation of a complex PE between the promoter, P, and a transcriptional effector E of that promoter through the cooperative binding of β effector molecules to the unoccupied promoter. This scheme can be represented by the reversible chemical reaction of the Hill type with the equilibrium constant K:

$$\beta E + P \leftrightarrows PE, \quad K = \frac{[PE]}{[E]^\beta [P]}, \tag{2.5}$$

where $[P]$, $[PE]$, and $[E]$ are the concentrations of unoccupied promoters, occupied promoters, and effector molecules, respectively. The parameter β is the Hill coefficient associated with the binding of the effector to the promoter.

The total concentration of promoters is proportional to the concentration $[P_{tot}]$ of the plasmid that carries the promoter. Plasmids are self-replicating, and the total number of plasmids (and, hence, of promoters) change as a cell progresses through the division cycle. The control of the plasmid copy number is quite elaborate [20] and must be balanced with the cell's growth and division. As a first approximation, it is assumed that the number of plasmids per cell scales proportionally with the cell volume such that the plasmid concentration remains fairly constant throughout the cell division cycle, i.e., that $[P_{tot}] = [P](t) + [PE](t)$ is constant. Combined with the equilibrium relation in (2.5), the conservation of plasmid concentration can be used to derive the concentration of active promoters $[A]$ used in (2.4). The effector can be either a transcriptional repressor or a transcriptional activator. In the case when the effector is the repressor, the unoccupied promoters are supposed to be active, and $[A]^R \equiv [P]$, where the superscript R stands for the repression case. Deriving concentration of the repressed promoters, $[PE]$, from (2.5) as a function of $[P]$, we have $[P_{tot}] = [P] + K[E]^\beta[P]$, or, taking into account the equivalence of $[P]$ and $[A]^R$, $[P_{tot}] = [A]^R + K[E]^\beta[A]^R$. Then, in the case of transcriptional repression, the concentration of active promoters is given by

$$[A]^R = \frac{[P_{tot}]}{1 + K[E]^\beta}. \tag{2.6}$$

In the case when the effector is the activator, the unoccupied promoters are assumed to be passive, and $[A]^A \equiv [PE]$, where the superscript A stands for the activation case. We derive concentration of unoccupied promoters from (2.5) as $[P] = [A]^A/K[E]^\beta$, then $[P_{tot}] = [A]^A/K[E]^\beta + [A]^A$. From this equation, the concentrations of active promoters is given by

$$[A]^A = \frac{[P_{tot}]K[E]^\beta}{1 + K[E]^\beta}. \tag{2.7}$$

Introducing the exponent a, we can write down the common formula for these two cases:

$$[A] = \frac{[P_{tot}]\left\{K[E]^\beta\right\}^a}{1 + K[E]^\beta}, \tag{2.8}$$

where the case of repression (R) corresponds to $a = 0$ and the case of activation (A) corresponds to $a = 1$.

Equation (2.4) can be generalized for the case of multiple promoters, controlling the production of the same protein. Suppose we have several protein effectors E_j, each of which influences production of the protein X, binding the corresponding promoter P_j ($j = 1, \ldots, M$). By summation of the contribution from each promoter P_j in the network, the evolution of the protein concentration $[X]$ can be written as

$$\frac{d[X]}{dt} = \sum_{j=1}^{M} \frac{b_{jx}k_j[P_{tot,j}][K_j[E_j](t)^{\beta_j}]^{a_j}}{1 + K_j[E_j](t)^{\beta_j}} - (k_X + k_g)[X]. \tag{2.9}$$

We also need to take into account that X may be able to penetrate the cell membrane by passive or active transport. An additional term for (2.9), which corresponds

to the passive transport, is $D_X([X] - [X_{ext}])$. Here, $[X_{ext}]$ is the extracellular concentration of X and D_X is an effective diffusion coefficient. The parameter D_X, in its simplest form, is defined by $D_X = S(t)p_X/v(t)$, where $S(t)$ is the cell surface area and p_X is the membrane permeability of X [22]. While D_X depends slightly on the stage of the cell division cycle, we will assume for simplicity that D_X is a constant. The resulting equation for a protein, which is synthesized from multiple promoters and can penetrate the cell membrane, is given by

$$\frac{d[X]}{dt} = \sum_{j=1}^{M} \frac{b_{jx}k_j[P_{tot,j}][K_j[E_j](t)^{\beta_j}]^{a_j}}{1 + K_j[E_j](t)^{\beta_j}} - (k_X + k_g)[X] - D_X([X] - [X_{ext}]). \tag{2.10}$$

Most proteins within the cell are unable to penetrate the cell membrane, and the diffusive term is in the present case only relevant for the AI.

2.2. The genetic oscillator model. The network diagram in Figure 1.1(B) can be converted into a system of evolution equations by using (2.10) for each of the three proteins U, V, and W synthesized from the three promoters. We use $[U]_i$, $[V]_i$, and $[W]_i$ to denote the concentrations of U, V, and W in cell i and $[W_{ext}]$ to denote the extracellular concentration of the AI:

$$\begin{aligned}
\frac{d[U]_i}{dt} &= \frac{b_{1u}k_1[P_{tot,1}]}{1 + K_1[V]_i^{\beta}} + \frac{b_{3u}k_3[P_{tot,3}]K_3[W]_i^{\eta}}{1 + K_3[W]_i^{\eta}} - (k_U + k_g)[U]_i, \\
\frac{d[V]_i}{dt} &= \frac{b_{2v}k_2[P_{tot,2}]}{1 + K_2[U]_i^{\gamma}} - (k_V + k_g)[V]_i, \\
\frac{d[W]_i}{dt} &= \frac{b_{2w}k_2[P_{tot,2}]}{1 + K_2[U]_i^{\gamma}} - (k_W + k_g)[W]_i - D_W([W]_i - [W_{ext}]),
\end{aligned} \tag{2.11}$$

where β, γ, and η denote the Hill coefficients of the P_1, P_2, and P_3 promoter, respectively.

Since the AI is able to penetrate the cell membrane, it is necessary to consider how the production of AI in an ensemble of N cells changes the extracellular AI concentration. The flux ϕ_i (in number/time unit) of W across the membrane of an individual cell is $\phi_i = S(t)p_W([W]_i - [W_{ext}])$ [22], and the evolution of the extracellular autoinducer concentration is given by

$$\frac{d[W_{ext}]}{dt} = \frac{v_c D_W}{v_{ext}} \frac{1}{N} \sum_{i=1}^{N} ([W]_i - [W_{ext}]) - k_0[W_{ext}], \tag{2.12}$$

where v_{ext} is the volume of the extracellular space, v_c is the total volume of N cells, and k_0 is the effective first-order constant of removal of AI from the extracellular medium. We assume that the experiments are carried out in a continuously stirred, constant volume flow reactor where the extracellular medium is homogeneous and the number of cells is kept constant by continuous dilution of the cell culture by a steady inflow of fresh growth medium and outflow of extracellular medium and cells. It is the rate of this dilution that determines the value of the parameter k_0.

To reduce the number of parameters in the system, we assume that U and V have identical half-lives, $k_d = k_U = k_V$. This assumption is based on the fact that the protein decay rate can be controlled in experiments. The identical half-lives are determined by identical protease tags added to these proteins. However, this assumption is not a constraint for design of the network but just a simplification for

our analysis. To normalize the equations, we introduce the following dimensionless variables:

$$u_i = \sqrt[\gamma]{K_2}[U]_i, \quad v_i = \sqrt[\beta]{K_1}[V]_i, \quad w_i = \sqrt[\eta]{K_3}[W]_i,$$
$$w_e = \sqrt[\eta]{K_3}[W_{ext}], \quad \tau = (k_d + k_g)t. \tag{2.13}$$

With these assumptions, the system is governed by the dimensionless system:

$$\begin{aligned}
\frac{du_i}{d\tau} &= \alpha_1 f(v_i) + \alpha_3 h(w_i) - u_i,\\
\frac{dv_i}{d\tau} &= \alpha_2 g(u_i) - v_i,\\
\frac{dw_i}{d\tau} &= \bar{\alpha}_2 g(u_i) - \delta w_i - D(w_i - w_e),\\
\frac{dw_e}{d\tau} &= \frac{D_e}{N}\sum_{i=1}^{N}(w_i - w_e) - \delta_e w_e,
\end{aligned} \tag{2.14}$$

where the functions are defined by

$$f(v) = \frac{1}{1+v^\beta}, \quad g(u) = \frac{1}{1+u^\gamma}, \quad h(w) = \frac{w^\eta}{1+w^\eta}, \tag{2.15}$$

and the dimensionless parameters are defined by

$$\begin{aligned}
\alpha_1 &= \frac{\sqrt[\gamma]{K_2} b_{1u} k_1 [P_{tot,1}]}{k_d + k_g}, \quad \alpha_2 = \frac{\sqrt[\beta]{K_1} b_{2v} k_2 [P_{tot,2}]}{k_d + k_g},\\
\bar{\alpha}_2 &= \frac{\sqrt[\eta]{K_3} b_{2w} k_2 [P_{tot,2}]}{k_d + k_g}, \quad \alpha_3 = \frac{\sqrt[\gamma]{K_2} b_{3u} k_3 [P_{tot,3}]}{k_d + k_g},\\
\delta &= \frac{k_W + k_g}{k_d + k_g}, \quad \delta_e = \frac{k_0}{k_d + k_g}, \quad D = \frac{D_W}{(k_d + k_g)}, \quad D_e = \frac{v_c D_W}{v_{ext}(k_d + k_g)}.
\end{aligned} \tag{2.16}$$

3. Isolated element. We first establish the conditions for oscillations in isolated cells. Cells can be considered as isolated elements in the limit $D_e \ll \delta_e$, corresponding to a vanishing cell density, where the contribution from cellular autoinducer production to the extracellular autoinducer concentration is vanishing and $w_e \to 0$. The evolution of protein content in an isolated cell is thus determined by

$$\begin{aligned}
&\frac{du}{d\tau} = \alpha_1 f(v) + \alpha_3 h(w) - u, \quad \frac{dv}{d\tau} = \alpha_2 g(u) - v,\\
&\frac{dw}{d\tau} = \bar{\alpha}_2 g(u) - (D+\delta)w = \varepsilon\left(\alpha_4 g(u) - w\right),
\end{aligned} \tag{3.1}$$

where $\varepsilon = D + \delta = (D_W + k_g)/(k_d + D_W)$ and $\alpha_4 = \bar{\alpha}_2(k_d + D_W)/(D_W + k_g)$. We suppose also that $\bar{\alpha}_2$ is of the same order as $(D+\delta)$, i.e., $\alpha_4 = O(1)$ because otherwise dynamics becomes trivial (the only stationary state).

When the parameter ε is small ($\varepsilon \ll 1$), the evolution of the system splits into two well-separated time-scales. In the fast time-scale, changes of the coordinates per unit of time τ are of order 1. Here, we can assume w to be stationary, since $dw/d\tau \sim \varepsilon \ll 1$. The fast motion ceases in the vicinity of the curve, where $du/d\tau = 0$ and $dv/d\tau = 0$, which is called the manifold of slow motion. On the manifold, changes of the coordinates per unit of time τ are of order ε, and we can introduce the slow time $\tau_1 = \varepsilon\tau$, where the changes are of order 1.

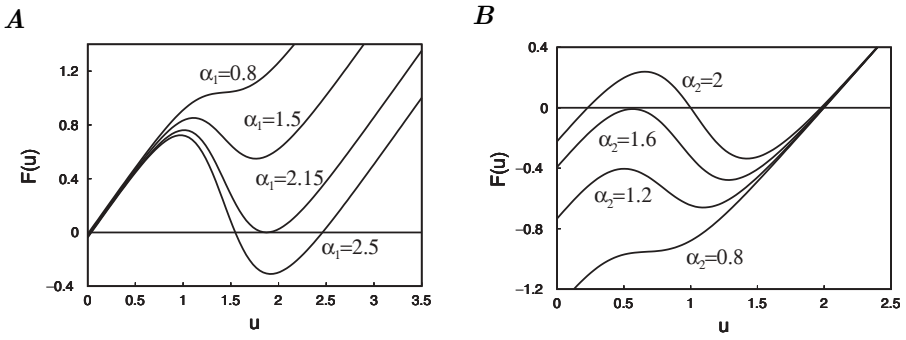


FIG. 3.1. *Geometrical investigation of equilibrium states in the fast subsystem with* $\alpha_w = 0$. *Equilibrium states are located where* $F(u) = 0$. (A) *Increasing the value of* α_1 *causes a transition from one to three equilibrium states, with a new equilibrium with high* u *being created through a saddle-node bifurcation. Parameter values are* $\alpha_2 = 4$, $\beta = \gamma = 3$. (B) *Creation of an equilibrium state with low* u *through a saddle-node bifurcation by an increase in* α_2. *Parameter values are* $\alpha_1 = 2$, $\beta = \gamma = 3$.

3.1. The fast subsystem. The first step in our analysis is to establish the conditions where the fast subsystem can be driven through a bistability region by varying the autoinducer concentration. Two conditions must be satisfied by the fast subsystem. (1) Two saddle-node bifurcations that define a region of bistability must exist and (2) the bifurcations must occur as the autoinducer concentration is varied. To establish the analytical conditions, we look for equilibria on the fast time-scale for $\varepsilon \to 0$, where the full system reduces to the toggle switch equations [1] augmented with a constant production term α_w arising from a constant concentration of autoinducer:

$$\begin{aligned} \frac{du}{d\tau} &= \alpha_1 f(v) + \alpha_w - u = P(u,v), \\ \frac{dv}{d\tau} &= \alpha_2 g(u) - v = Q(u,v), \end{aligned} \tag{3.2}$$

where $\alpha_w = \alpha_3 h(w)$. These equations correspond to those used by Kobayashi et al. to guide the construction of a toggle-based AI biosensor [10]. By Bendixson's criterion [23], the system in (3.2) has no closed orbit since the divergence of the vector field $P_u' + Q_v' = -2$ does not change sign.

3.1.1. Absence of autoinducer. In the absence of the autoinducer ($\alpha_w = 0$), the equilibrium states (u_0, v_0) of the system in (3.2) can be found by setting $v = \alpha_2 g(u)$ as the zeros of the function $F(u)$ given by

$$F(u) = u - \alpha_1 f\left(\alpha_2 g(u)\right) = u - \frac{\alpha_1 (1+u^\gamma)^\beta}{\alpha_2^\beta + (1+u^\gamma)^\beta}. \tag{3.3}$$

Since $F(u) \to -\alpha_1/(\alpha_2^\beta + 1) < 0$ for $u \to 0$ and $F(u) \to u - \alpha_1 > 0$ for $u \to \infty$, the existence of at least one steady state is guaranteed.

A necessary, but not sufficient, condition for the existence of multiple equilibrium states is that $F(u)$ be an N-shaped function such that there exist local extrema (where $F'(u) = 0$). Figure 3.1 illustrates the transition from monostability when α_1 is varied for $\alpha_2 = 4$, $\beta = \gamma = 3$. At very low values of α_1, the function $F(u)$ is monotonically increasing and there exists a single equilibrium state where u_0 is low and v_0 is high.

When α_1 increases, a local maximum and a local minimum emerge, but there is still only a single equilibrium state. As α_1 increases further, the local minimum of $F(u)$ is shifted downward, and it coincides with $F(u) = 0$ when α_1 reaches a critical value α_1^c (at approximately $\alpha_1 = 2.14925$ in Figure 3.1). This critical point corresponds to a saddle-node bifurcation where the two conditions $F(u) = 0$ and $F'(u) = 0$ are simultaneously fulfilled and a new equilibrium state is created. For α_1 higher than the critical value, the function $F(u)$ has three zeros corresponding to three equilibrium states. Two of these states are destroyed when α_1 is very high (greater than approximately 22.9767 for $\alpha_2 = 4$, $\beta = \gamma = 3$) where the local maximum is shifted to negative values of $F(u)$ (not shown). The system is again monostable, this time with an equilibrium state where u_0 is high and v_0 is low. As illustrated in Figure 3.1(B), a similar bifurcation scenario is observed when α_2 is varied.

The characteristic polynomial that determines stability of the equilibrium states of the fast subsystem is given by

$$\begin{aligned} \lambda^2 + 2\lambda + F'(r) &= 0, \\ F'(r) &= 1 - \alpha_1\alpha_2 f_v'(v_0(r))g_u'(r), \end{aligned} \tag{3.4}$$

where we have introduced the parameter r to represent the equilibrium state (u_0, v_0). This parameter is obtained from (3.1) by setting $du/d\tau = 0$ and $dv/d\tau = 0$:

$$r = u_0, \quad v_0 = \alpha_2 g(r). \tag{3.5}$$

It can be shown that when a single equilibrium state exists, it is always stable (monostability), and when three equilibrium states exist, one of them is unstable and the remaining two are stable (bistability).

As described above, the transition from monostability to bistability occurs through saddle-node bifurcations. Their location can be predicted from (3.4) by finding solutions where $\lambda = 0$, i.e., from $F'(r) = 0$. This equation can be written in a parametric form (Appendix A) to obtain sets of critical parameter values $(\alpha_1^c(r), \alpha_2^c(r))$ that determine the location of the saddle-node bifurcations in the α_1, α_2 phase plane:

$$\begin{aligned} \alpha_1^c(r) &= \tfrac{\beta\gamma r^{\gamma+1}}{1+r^\gamma} \Big/ \left(\tfrac{\beta\gamma r^\gamma}{1+r^\gamma} - 1\right), \\ \alpha_2^c(r) &= (1+r^\gamma)\left(\tfrac{\beta\gamma r^\gamma}{1+r^\gamma} \Big/ \left(\tfrac{\beta\gamma r^\gamma}{1+r^\gamma} - 1\right) - 1\right)^{1/\beta}. \end{aligned} \tag{3.6}$$

In these equations, r is in the range $(r_l : \infty)$ with r_l defined by $r_l^\gamma = (\beta\gamma - 1)^{-1}$ (implying that $\beta\gamma > 1$ since r_l must be positive). Note that $\alpha_1^c(r) < 0$ when $r < r_l$, which violates the condition that all the parameters must be positive reals.

3.1.2. Presence of autoinducer. In the presence of autoinducer ($\alpha_w > 0$), the equilibrium states are obtained as the solution of $F(u) = \alpha_w$ rather than $F(u) = 0$. In order to use variation in α_w to drive the fast subsystem through a bistability region, it is essential that an increase (or decrease) in α_w causes the fast subsystem to pass through the two saddle-node bifurcations. Therefore, the system must be monostable when α_w is lower than a critical value $\alpha_w^+ > 0$, having the only equilibrium with low u (denoted r^-), monostable when α_w is greater than the second critical value $\alpha_w^- > \alpha_w^+$, having the only equilibrium with high u (denoted r^+), and bistable when $\alpha_w^- > \alpha_w > \alpha_w^+$. This scenario is depicted in Figure 3.2(A), where one, two, or three equilibrium states arises as α_w is varied. The critical values in α_w, where the fast

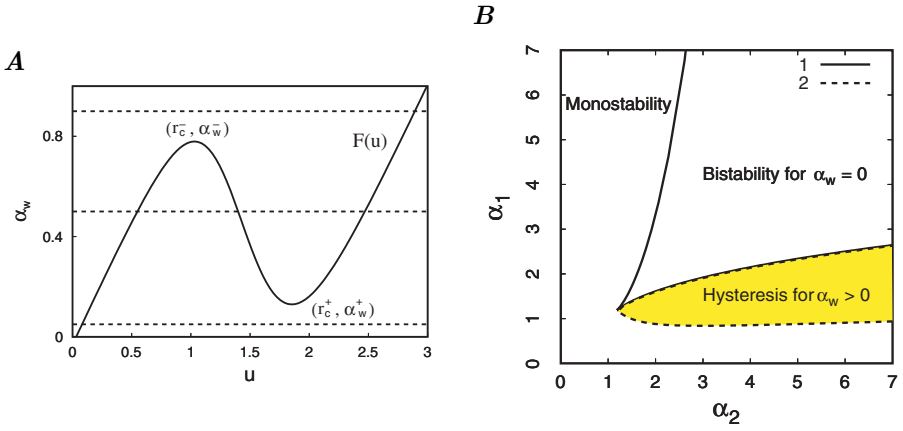


FIG. 3.2. (A) *Changes in the number of equilibrium states by variation in* α_w. *At* $\alpha_w = 0.05$ *there is a single equilibrium state at low* u. *There are three equilibrium states at* $\alpha_w = 0.5$ *and a single equilibrium state for* $\alpha_w = 0.9$. *The points labeled* (r_c^-, α_w^-) *and* (r_c^+, α_w^+) *correspond to the values of* α_w *at the saddle-node bifurcations. Parameter values are* $\alpha_1 = 2, \alpha_2 = 4$, $\beta = \gamma = 3$. (B) *Different regions of the* α_1, α_2 *phase plane showing different behavior for* $\beta = \gamma = 3$. *The solid curve encloses a region where the system is bistable in the absence of autoinducer. The solid and the dashed curve enclose a region where bistability can occur in the presence of autoinducer, i.e., a hysteresis loop for* $\alpha_w > 0$.

subsystem has two equilibrium states (a stable node and a saddle-node), satisfy the conditions

$$\alpha_w^- = F(r_c^-), \quad \alpha_w^+ = F(r_c^+), \quad F'(r_c^\pm) = 0, \tag{3.7}$$

where $r_c^\pm$ are the values of u corresponding to the extrema of $F(u)$.

To achieve hysteresis when α_w is varied, $F(u)$ must have two extrema and they must be located in the positive quadrant, i.e., $F(r_c^\pm) > 0$. The section of parameter plane where the fast subsystem satisfies the required conditions are thus bounded by two curves: one where bistability ceases to exist, corresponding to the merger of extrema of $F(u)$, and one where the minimum of $F(u)$ crosses into negative values. As derived in Appendix B, the merging extrema of the function $F(u)$ defines a curve $(\alpha_1^m(r), \alpha_2^m(r))$ in the α_1, α_2 phase plane given by

$$\alpha_1^m(r) = \frac{(1+r^\gamma)[1+R_1(r)]^2}{\gamma\beta R_1(r) r^{\gamma-1}}, \quad \alpha_2^m(r) = (1+r^\gamma) R_1(r)^{1/\beta}, \tag{3.8}$$

where

$$R_1(r) = -\frac{(\gamma-1) - r^\gamma(1+\beta\gamma)}{(\gamma-1) - r^\gamma(1-\beta\gamma)}. \tag{3.9}$$

The curve $(\alpha_1^m(r), \alpha_2^m(r))$ is in addition subject to the condition that the system is monostable in the absence of autoinducer. In other words, $F(u) = \alpha_w$ must have a single solution for $\alpha_w = 0$ and the saddle-node bifurcations must therefore occur at values of $\alpha_w = F(u) > 0$. The boundary of this condition coincides with that of emergence of bistability in the unperturbed toggle switch, which is determined by the curves of saddle-node bifurcations in (3.6). Figure 3.2(B) illustrates the regions of different dynamics in the α_1, α_2 parameter space. The solid curve is obtained

from (3.6), and the dashed curve shows the merging of extrema (3.8). The area enclosed by the solid curves corresponds to the region in parameter space where the fast subsystem shows bistability in the absence of autoinducer, and the shaded area corresponds to the region of parameter space where there exists a bistable region for $0 < \alpha_w^+ < \alpha_w < \alpha_w^-$, where α_w^+ and α_w^-, as previously defined in this section, are the critical values of autoinducer at the saddle-node bifurcations (see Figure 3.2(A)).

3.2. The slow subsystem. Given sufficient time-scale separation, the fast subsystem reaches a point on the manifold of slow motion, where the dynamics is governed by

$$\frac{dw}{d\tau} = \varepsilon(\alpha_4 g(u) - w). \tag{3.10}$$

Here u satisfies the condition $F(u) = \alpha_3 h(w)$ obtained from (3.2). When the parameters of the fast subsystem are such that there exist two extrema of $F(u)$ at $u = r_c^-$ (the local maximum) and $u = r_c^+$ (the local minimum), the slow subsystem can drive the fast subsystem through a bistability region if $\alpha_w = \alpha_3 h(w)$ can assume values on either side of the interval $[\alpha_w^-, \alpha_w^+]$ where $\alpha_w^\pm = F(r_c^\pm)$, as was illustrated in Figure 3.2(A).

The equilibrium states of the whole system are given by the intersection in u, α_w space between $F(u)$ and the curve

$$\alpha_w(w(u)) \equiv \alpha_3 h(w), \quad w(u) = \alpha_4 g(u). \tag{3.11}$$

The curve $\alpha_w(w(u))$ is a monotonically decreasing function of u since $w(u)$ is a monotonically decreasing function of u and h is monotonic. In order for the slow subsystem to meet the above conditions, it is required that

$$\alpha_w(w(r_c^-)) > F(r_c^-), \quad \alpha_w(w(r_c^+)) < F(r_c^+). \tag{3.12}$$

This condition implies that $\alpha_w(w(u))$ and $F(u)$ must intersect for values of u where $F'(u) < 0$. Figure 3.3(A) illustrates the different scenarios that are possible for different values of the parameters of the slow subsystem. When the parameters are appropriately adjusted, the curves $\alpha_w(w(u))$ and $F(u)$ intersect once in the region where $F'(u) < 0$ and the conditions in (3.12) are satisfied. For other parameter values, there may be one, two, or three intersections of the curves, which violates one of the conditions in (3.12).

The limits of the inequalities in (3.12) can be used to obtain the regions in the parameter space, where the slow subsystem satisfies the required conditions. In particular, equation $\alpha_w(w(r_c^-)) = F(r_c^-)$ requires that $\alpha_w(w(u))$ and $F(u)$ intersect in the point where $F'(u) = 0$, i.e., in the maximum of the function $F(u)$. Hence, the critical values of the parameters where $\alpha_w(w(u))$ intersects an extremum of $F(u)$ satisfies the following condition:

$$\alpha_w(w(r_c^\pm)) = F(r_c^\pm), \quad F'(r_c^\pm) = 0. \tag{3.13}$$

We apply this condition to obtain the region in the (α_3, α_4) parameter plane for different values of η where the slow subsystem satisfies the requirements for relaxation oscillations. These curves are plotted in Figure 3.3(B) and show an increase of the oscillatory region (filled) with increasing η.

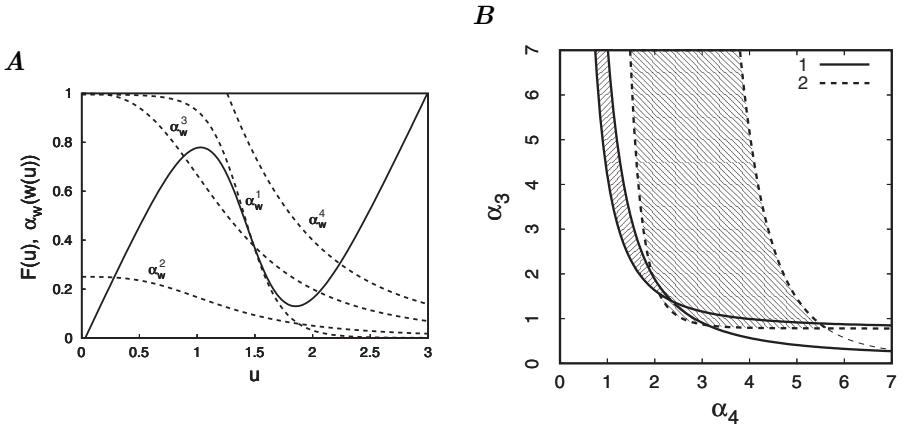


FIG. 3.3. *Geometrical analysis of equilibrium states as the parameters of the slow subsystem are varied.* (A) *Oscillations are possible* (3.12) *when the curves* $F(u)$ *and* $\alpha_w(w(u))$ *intersect only in the region where* $F'(u) < 0$ *(curve* α_w^1*,* $\eta = 4$*,* $\alpha_3 = 1$*,* $\alpha_4 = 3.8$*). The curves* α_w^2*,* α_w^3*, and* α_w^4 *are obtained for* $\eta = 1$*,* $\alpha_4 = 1$*, and* $\alpha_3 = 0.5$*, 2, 4, respectively. They illustrate monostability (*α_w^2 *and* α_w^4*) and multistability (*α_w^3*) in the system. Other parameters are* $\alpha_1 = 2$*,* $\alpha_2 = 4$*,* $\beta = \gamma = 3$*.* (B) *Regions in the* α_3*,* α_4 *parameter space where the conditions* (3.12) *are satisfied for different values of* η*:* (1) $\eta = 2$*;* (2) $\eta = 6$*.*

The condition in (3.13) is solved with respect to α_1 and α_2 (Appendix C) to give a set of bifurcation points (α_1^H, α_2^H) in the limit $\varepsilon = 0$:

$$\alpha_1^H(r) = \frac{1 + R_2(r)}{r - R_3(r)}, \quad \alpha_2^H(r) = (1 + r^\gamma)\,(R_2(r))^{1/\beta}, \tag{3.14}$$

where

$$\begin{aligned} R_2(r) =& \frac{\beta\gamma r^{\gamma-1}\,(r - R_3(r))}{(\beta\gamma r^\gamma - r^\gamma - \beta\gamma r^{\gamma-1} R_3(r) - 1)} - 1, \\ R_3(r) =& \alpha_3 \left(\frac{\alpha_4}{1 + r^\gamma}\right)^\eta \Bigg/ \left(1 + \left(\frac{\alpha_4}{1 + r^\gamma}\right)^\eta\right). \end{aligned} \tag{3.15}$$

As illustrated in Figure 3.4(A), the bifurcation curve has a loop structure and defines two distinct regions of parameter space. The region R is the set of α_1, α_2 values where system (3.1) can display oscillations for sufficiently low values of ε . The region labeled M defines a set of α_1, α_2 values where system (3.1) displays multistability.

3.3. Bifurcation analysis. The positions of equilibrium states $S = (u_0, v_0, w_0)$ in the full system in (3.1) are determined by the equations

$$\alpha_1 f(v_0) - u_0 + \alpha_3 h(w_0) = 0, \quad \alpha_2 g(u_0) - v_0 = 0, \quad \alpha_4 g(u_0) - w_0 = 0. \tag{3.16}$$

The stability of the equilibrium states are obtained from the Jacobian matrix,

$$J = \begin{pmatrix} -1 & \alpha_1 f_v'(v_0) & \alpha_3 h_w'(w_0) \\ \alpha_2 g_u'(u_0) & -1 & 0 \\ \varepsilon\alpha_4 g_u'(u_0) & 0 & -\varepsilon \end{pmatrix}, \tag{3.17}$$

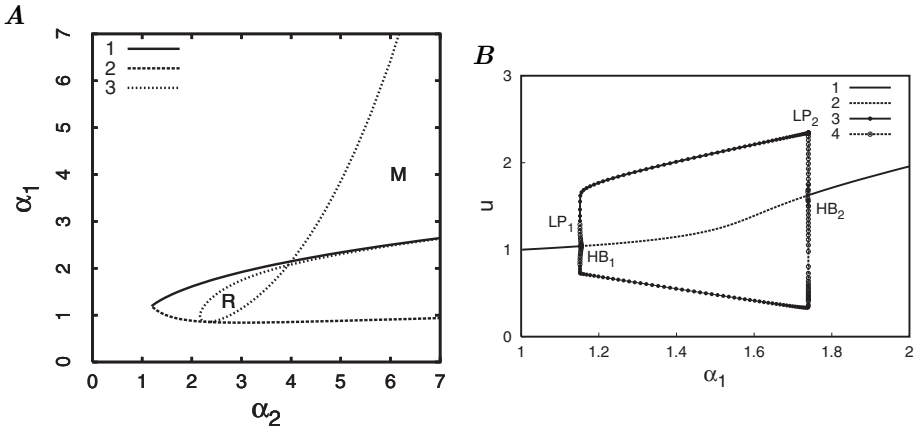


FIG. 3.4. *Bifurcation analysis of the full system.* (A) *Regions of different dynamic behavior in the* α_1, α_2 *parameter plane for* $\alpha_3 = 1, \alpha_4 = 3, \beta = \gamma = \eta = 3$. *Oscillations can occur in the region labeled R for sufficiently low* ε. *The system is bistable in the region labeled M and monostable everywhere else.* (B) *An example of bifurcation diagram obtained by variation in* α_1 *for a fixed value of* α_2 *(*$\alpha_2 = 3$*).*

by evaluation of the characteristic equation given by

$$\lambda^3 + \lambda^2(2+\varepsilon) + \lambda\left(1 - \alpha_1\alpha_2 f_v'(v_0)g_u'(u_0) + 2\varepsilon - \varepsilon\alpha_3\alpha_4 h_w'(w_0)g_u'(u_0)\right) + \varepsilon - \varepsilon\alpha_1\alpha_2 f_v'(v_0)g_u'(u_0) - \varepsilon\alpha_3\alpha_4 h_w'(w_0)g_u'(u_0) = 0. \tag{3.18}$$

The Andronov–Hopf bifurcation, which gives birth to a limit cycle, occurs when a pair of complex conjugate eigenvalues crosses the imaginary axis. If we write down the characteristic equation in the form $\lambda^3 + a\lambda^2 + b\lambda + c = 0$, then the condition for the Andronov–Hopf bifurcation takes the form $ab - c = 0$. From (3.18), this implies that the bifurcation occurs when the following condition is fulfilled:

$$1 - \alpha_1\alpha_2 f_v'(v_0)g_u'(u_0) + \varepsilon\left(2 - \frac{\alpha_3\alpha_4}{2} g_u'(u_0)h_w'(w_0)\right) + \varepsilon^2\left(1 - \frac{\alpha_3\alpha_4}{2} g_u'(u_0)h_w'(w_0)\right) = 0. \tag{3.19}$$

In the limit $\varepsilon \to 0$, we recover condition (3.13) for $\alpha_w(w(u))$ intersecting an extremum of $F(u)$. This is because a solution of the system (3.16), u_0, fits the equation $\alpha_w(w(u_0)) = F(u_0)$, and (3.19) for $\varepsilon = 0$ takes the form $1 - \alpha_1\alpha_2 f_v'(v_0)g_u'(u_0) = 0$, which is equivalent to $F'(u_0) = 0$. In other words, oscillations are constrained to be in the region where the conditions imposed by the slow subsystem (3.13) are satisfied, which, in turn, lies inside the region of hysteresis of the fast subsystem (Figure 3.2(B)).

Figure 3.4(B) illustrates in more detail the bifurcation structure of the full system when α_1 is varied at constant values of α_2. Here, Andronov–Hopf bifurcations, which correspond to entering and exiting from the oscillatory region, are labeled as HB_1 and HB_2. These bifurcations are subcritical and accompanied by saddle-node bifurcations of limit cycles LP_1 and LP_2. The points of Andronov–Hopf bifurcations agree well with the points of intersection of curve 3 of Figure 3.4(A) with the line that corresponds to the given value of α_2. This agreement shows that the region R in Figure 3.4(A) gives a good approximation for the oscillatory region of the full system if ε is small.

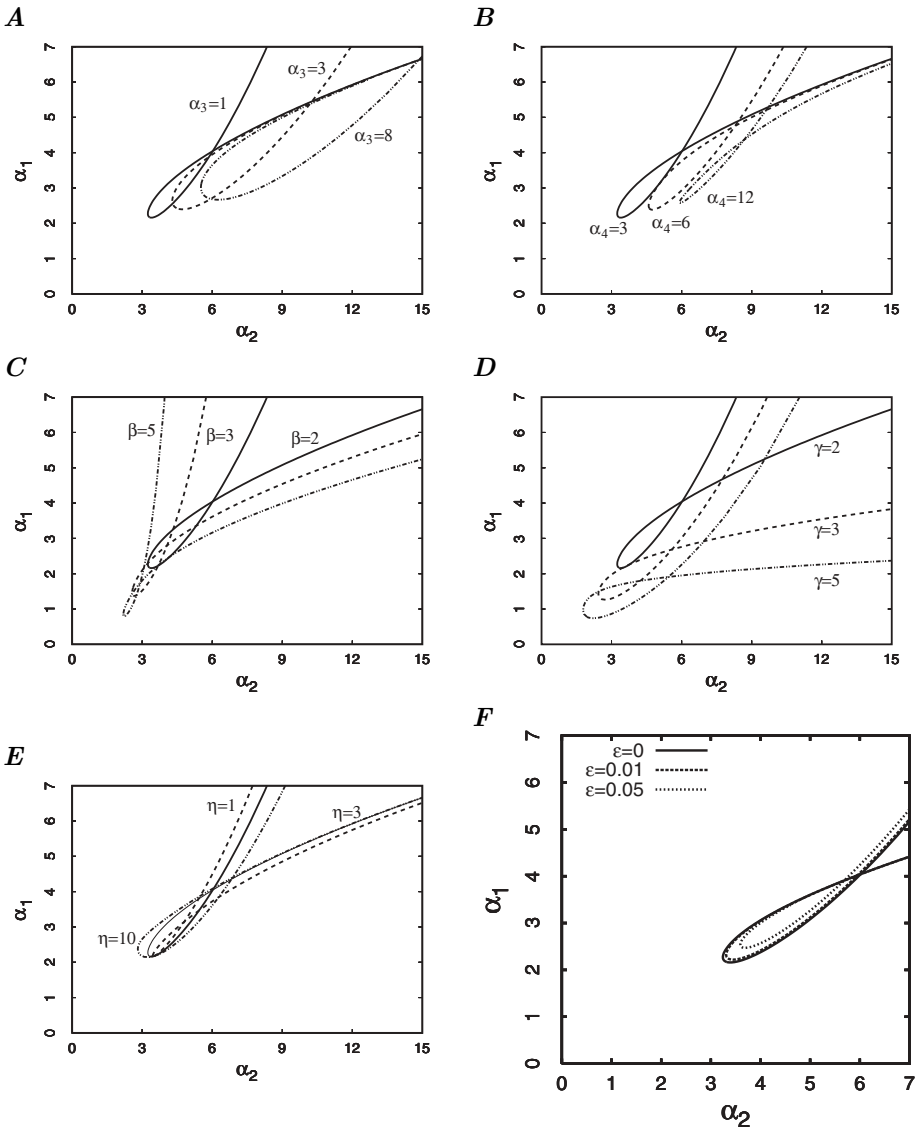


FIG. 3.5. *Increasing the oscillatory region in the α_1, α_2 parameter plane. All plots show the bifurcation curve $(\alpha_1^H(r), \alpha_2^H(r))$ for the reference parameters $\alpha_3 = 1$, $\alpha_4 = 3$, $\beta = \gamma = 2, \eta = 3$, $\varepsilon = 0$ in full. The six plots show the effect of variation in* (A) α_3, (B) α_4, (C) β, (D) γ, (E) η, *and* (F) ε *relative to the reference parameters.*

3.4. Parameter dependence. In this section we are optimizing conditions for oscillations by variation of all of the model parameters. In Figure 3.5(A)–(E), we plot the bifurcation curve $(\alpha_1^H(r), \alpha_2^H(r))$ defined in (3.14), i.e., for $\varepsilon = 0$. Increasing ε decreases the region in the α_1, α_2 parameter plane where oscillations are observed (Figure 3.5(F)). Comparing different curves in Figure 3.5(A), the range of both α_1 and α_2 values where oscillation can occur is seen to expand as α_3 is increased, indicating that larger values of α_3 increase the likelihood of oscillations. In Figure 3.5(B), it is seen that the region of oscillations is maximized at intermediate values of α_4. Therefore, the rate of AI synthesis must be carefully chosen to observe oscillations.

This can be done experimentally by manipulating the *luxI* RBS. Interestingly, Figure 3.5(C) shows a counterintuitive result, namely that the region of oscillations expands as β is decreased, i.e., when the degree of nonlinearity is decreased. Figure 3.5(D) and (E) shows the opposite effect for different nonlinearity exponents, namely that the oscillatory region shrinks when γ and η are decreased.

The exponents β and η have opposite influence because these two parameters change slopes of the function $F(u)$ and $a_w(u)$ independently. In the case where $a_w(u)$ coincides with the middle (decreasing) branch of $F(u)$, the system is very sensitive to changing other parameters. This is because very small variations of a parameter may cause an intersection outside the middle branch of $F(u)$, which corresponds to a stable equilibrium state. When the slope of $a_w(u)$ is less than of the middle branch of $F(u)$, relaxation oscillations cannot occur (see Figure 3.3, curve a_w^3). Thus, the larger the η, the larger the slope of $a_w(u)$, and the larger the tolerance of other parameters for oscillatory dynamics. By contrast, the larger the β, the larger the slope of $F(u)$, and the smaller the region of relaxation oscillations for given η. Parameter γ changes both $F(u)$ and $a_w(u)$, which results in an increase of the oscillatory region with increase of this parameter.

4. Oscillations in more detailed models. In this section, we consider how details left out during the derivation of the minimal model affect the ability of the single cells to display oscillatory behavior. We consider three important assumptions: (1) titration and saturation of the LuxR transcription factor by the AI, (2) two-step synthesis of the AI, and (3) the effect of "leaky" promoters. We also consider how oscillatory behavior can be made more robust by adding an additional connectivity to the network.

4.1. Taking LuxI synthesis into account increases the oscillatory region. As mentioned in the Introduction, the AI is not a gene product, but a small molecule synthesized by the protein encoded by the *luxI* gene (see Figure 1.1A). A more realistic description of the network would therefore involve production of AI in two steps, synthesis of the LuxI protein by the transcription and the translation of *luxI* and subsequent synthesis of the AI by the LuxI protein. This can be accounted for by introducing a new dimensionless variable, x, for the concentration of the LuxI protein and a rate of AI production that is proportional to x. The minimal model (3.1) is recovered when x is assumed to be in a quasi-steady state, $dx/d\tau = 0$. This assumption, however, is not justified since LuxI is a stable protein whose evolution occurs on the same time-scale as the slow variable w. Assuming the time-scales are the same, we take degradation rates of LuxI and AI to be equal and denote both of them as δ. When δ is small, the location of the oscillatory region is slightly shifted in the parameter space (data not shown), indicating that the model where LuxI synthesis is ignored, i.e., (3.1), is a reasonable approximation for this case. On the other hand, when δ is not small, the effect of time lag introduced by the two-step synthesis of the AI is significant. In the minimal model (3.1), oscillations are suppressed when the value of δ exceeds roughly 0.08 for all value of α_4. In the model that incorporates LuxI, oscillations cease when δ exceeds 0.3. This a major improvement for the likelihood of observing oscillations experimentally since smallness of the parameter δ is a major experimental challenge. It requires that the protein half-life, which typically is roughly 30 min or longer, is roughly 20 times shorter than the cell division time. Fortunately, the production of AI in two steps allows for a significant increase in the value of δ where oscillations can be observed. If other parameters are adjusted appropriately, it is possible to get oscillations for δ as high as 0.3.

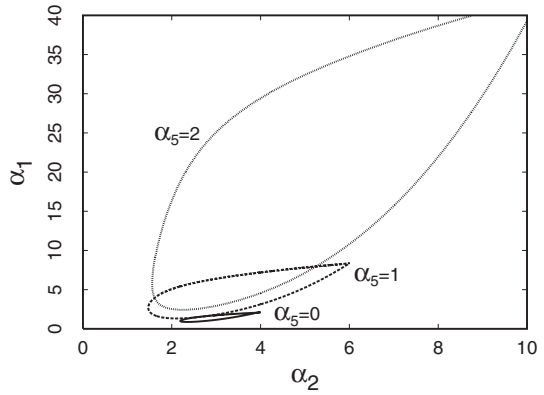


FIG. 4.1. *Increasing robustness of oscillations. The boundary of the oscillatory region in the α_1, α_2 parameter plane for different values of α_5. Parameter values: $\alpha_3 = 1$, $\alpha_4 = 0.03$, $\beta = \gamma = \eta = \zeta = 3$, $\delta = 0.01$, $D = 0$.*

4.2. Adding connectivity to the network increases the oscillatory region. The previous sections have demonstrated that the organization of oscillations in isolated cells requires that most of the parameters are precisely adjusted to a fairly narrow region of parameter space. We investigated if small changes to the system may enhance the region of parameter space where oscillations can be observed. One change that has a dramatic effect on the system is to express the gene coding for the V repressor from a promoter, denoted P_{W2}, that is repressed by AI.

$$\begin{aligned} \frac{du}{d\tau} &= \alpha_1 f(v) + \alpha_3 h(w) - u, \\ \frac{dv}{d\tau} &= \alpha_2 g(u) + \alpha_5 j(w) - v, \\ \frac{dw}{d\tau} &= \bar{\alpha}_2 g(u) - (\delta + D)w, \end{aligned} \tag{4.1}$$

where $\alpha_5 j(w) = \alpha_5/(1 + w^\zeta)$ represents expression of the protein V via the promoter P_{W2}. The model in (3.1) is recovered in the limit $\alpha_5 = 0$.

The addition of an AI-repressed promoter synthesizing the V repressor has a significant impact on the ability of isolated cells to oscillate since it favors the V high state in the absence of autoinducer without making the U high state harder to achieve in the presence of autoinducer. As a result, as α_5 increases, there is an increase in the region of parameter space where the fast subsystem has no stable equilibrium states and, thus, is able to oscillate. In Figure 4.1, we compare the region in the α_1, α_2 parameter plane where oscillations are observed for different values of α_5. It is evident that increased α_5 causes the region of oscillations to expand considerably, thus making oscillatory behavior in isolated cells more robust.

4.3. LuxR synthesis. As mentioned in the Introduction, the transcription factor that activates expression from the P_{lux} promoter is not the AI, as was assumed in the minimal model, but a complex comprised of LuxR and AI (Figure 1.1(A)). This complex is formed in a bimolecular reaction:

$$LuxR + AI \leftrightarrows LuxR - AI, \quad K_4 = \frac{k_{4f}}{k_{4b}}, \tag{4.2}$$

where K_4 is the equilibrium constant and k_{4b} and k_{4f} are the rate constants for the dissociation and association reaction, respectively. The *luxR* gene is assumed to be expressed at a constant rate from a plasmid-borne, constitutive promoter P_{con}, such that the LuxR protein is synthesized at a constant rate v_R. To obtain the minimal model (3.1), we need to assume here that the concentration of free autoinducer [AI] is negligible. That is, we assume that LuxR synthesis rate is large ($v_R \gg 1$) to provide LuxR for binding with AI, and the association reaction (4.2) is fast ($k_{4f} \gg k_{4b}$ and $k_{4f} \gg 1$). Our simulations reveal (data not shown) that, for smaller v_R, the oscillatory region shrinks and shifts to smaller α_1 and α_2. Violation of the other inequalities (e.g., $k_{4f} \ll k_{4b}$) makes the changes more significant. Hence, the details of formation of this effector complex may make oscillations more difficult to obtain.

4.4. Promoter leakage. In all of the models investigated, we have assumed that the promoters are fully repressible or fully silenced meaning that there is no expression from the promoter when repressor concentration is high or when activator is absent. In reality, many bacterial promoters are "leaky," and expression occurs at a basal level even under conditions where repressor is present in excess or activator is completely absent from the system.

To evaluate the effect of promoter leakage on the ability of isolated cells to oscillate, we introduced a constant synthesis term in each of the variables u and v that is proportional to maximal synthesis rate α_j. For simplicity, we use the same proportionality factor μ, corresponding to identical relative basal synthesis rates for all promoters. For large Hill coefficients $\eta = \zeta = 3$, the oscillations were observed at a fairly high value of leakage $\mu = 0.1$, i.e., 10% of the maximal synthesis rate for all promoters (data not shown). Decreasing the values of η and ζ causes the oscillatory region to be confined to lower values of μ. This indicates that organization of oscillations in isolated elements does not require the very tightly regulated promoters.

5. Ensemble of cells. We now study collective dynamics of the cell population. Introduction of coupling between elements of an ensemble can lead to qualitative changes of their dynamics. We are interested in providing synchronous oscillations, which would correspond to macroscopic oscillations of a protein concentration over the whole population. We demonstrate the possibility of both population synchronization and suppression of oscillations, depending on coupling strength and other parameters of the system.

First we make a transformation of the coordinates and parameters to combine intra- and extracellular degradation of the autoinducer into single term, thereby decreasing the number of parameters in the system. Then the system (2.14), describing the population of $i = 1, \dots, N$ cells, takes the form

$$
\begin{aligned}
\frac{du_i}{dt} &= \alpha_1 f(v_i) - u_i + \alpha_3 h(w_i),\\
\frac{dv_i}{dt} &= \alpha_2 g(u_i) - v_i,\\
\frac{dw_i}{dt} &= \bar{\varepsilon}\,(\bar{\alpha}_4 g(u_i) - w_i) + 2d(\bar{w}_e - w_i),\\
\frac{d\bar{w}_e}{dt} &= \frac{d_e}{N}\sum_{i=1}^{N}(w_i - \bar{w}_e).
\end{aligned}
\tag{5.1}
$$

Here, $\bar{w}_e = w_e\,(1 + \delta_e/D_e)$, $\bar{\varepsilon} = D + \delta - \frac{D}{(1+\delta_e/D_e)}$, $d = \frac{D}{2(1+\delta_e/D_e)}$, $d_e = D_e + \delta_e$, and $\bar{\alpha}_4 = \bar{\alpha}_2/\bar{\varepsilon}$.

Let us consider the simplest synchronous solution, i.e., identical synchronization of all elements of the ensemble: $u_i = u(t), v_i = v(t), w_i = w(t)$, $i = \overline{1,N}$. These equalities give the manifold of identity of corresponding coordinates: $M\{u_i, v_i, w_i : u_i = u_j, v_i = v_j w_i = w_j \forall i = \overline{1,N}, j = \overline{1,N}\}$. Now we study two matters: (1) dynamics on this manifold and (2) its stability. We show that if the isolated element displays relaxation oscillations, then the ensemble has the solution of identical synchronization for both small and large coupling strength. However, for the latter, the synchronous state may not be stable.

5.1. Identical synchronization. Dynamics on the manifold of identity of corresponding coordinates, M, is given by the following system:

$$\begin{aligned}
\frac{du}{dt} &= \alpha_1 f(v) - u + \alpha_3 h(w),\\
\frac{dv}{dt} &= \alpha_2 g(u) - v,\\
\frac{dw}{dt} &= \bar{\varepsilon}\,(\bar{\alpha}_4 g(u) - w) + 2d(\bar{w}_e - w),\\
\frac{d\bar{w}_e}{dt} &= d_e(w - \bar{w}_e).
\end{aligned} \tag{5.2}$$

Suppose that we have relaxation oscillations in each isolated element (for which $D_e \ll \delta_e$), i.e., we have the oscillations in this system with $d \to 0$ and $\bar{\varepsilon} \to \varepsilon = D + \delta$. We also assumed $\varepsilon \ll 1$ to obtain the oscillations.

We show first that the oscillations persist for small nonzero coupling strength $0 < d \ll 1$. We suppose also that the extracellular coupling coefficient is not small: $d_e \sim 1$. Then the system can be divided into fast and slow parts. The fast subsystem

$$\frac{du}{dt} = \alpha_1 f(v) - u + \alpha_3 h(w), \tag{5.3}$$

$$\frac{dv}{dt} = \alpha_2 g(u) - v, \tag{5.4}$$

$$\frac{d\bar{w}_e}{dt} = d_e(w - \bar{w}_e) \tag{5.5}$$

gives dynamics of three variables in the fast time-scale, where w is a constant. The u, v equations and the $\bar{w}_e$ equation do not depend on one another, so the fast subsystem splits into two independent parts. The u, v part is identical to the fast subsystem of the isolated element, in which all trajectories on the (u, v) plane converge to one of the equilibria. Trajectories of the $\bar{w}_e$ equation converge to the equilibrium state $\bar{w}_e = w$.

The slow subsystem is determined on the manifold of slow motion, i.e., in the intersection of all nullclines of the fast subsystem. This implies that we need to consider the equation for w on the manifold $\{\bar{w}_e = w, F(u) = \alpha_3 h(w)\}$. Substitution of the first constraint in the third equation of the system (5.2) gives

$$\frac{dw}{dt} = \bar{\varepsilon}(\bar{\alpha}_4 g(u) - w), \tag{5.6}$$

where u is a function of w, taken from the second constraint ($u = F^{-1}(\alpha_3 h(w))$). This equation has the same form as the slow equation obtained for the isolated element

(3.10), with parameters $\bar{\varepsilon}$ and $\bar{\alpha}_4$ representing other combinations of the initial parameters. Thus, for a given set of initial parameters D, δ, D_e, and δ_e, the slow dynamics of the system (5.2) with weak coupling strength d differs from the slow dynamics of the isolated element, but the parametric portrait of the isolated element with respect to parameters ε, α_4 coincides with the portrait for the system (5.2) with respect to parameters $\bar{\varepsilon}$ and $\bar{\alpha}_4$. If we have a solution for an isolated element with some values of ε and α_4, we can obtain the same solution in the system (5.2) with weak coupling strength d by tuning the parameters D, δ, D_e, and δ_e so that $\bar{\varepsilon}$ and $\bar{\alpha}_4$ take values ε and α_4. Thus, if a solution exists for the isolated element, then the same solution exists for the ensemble on the manifold of identical synchronization. Thus, we have shown, in particular, that there exists a regime of relaxation oscillations for a nonzero but weak coupling strength ($0 < d \ll 1$).

Next we consider the existence of a relaxation oscillation solution for large coupling strength $d \gg 1$. A shift in the frequency of the oscillation is obtained below for this case. The analysis can be performed analogously to that in [24]. There, the authors have proved that, for large coupling strength, the coupling term remains $O(1)$. Analogously, in our case, the coupling term $d(\bar{w}_e - w)$ is $O(\bar{\varepsilon})$ for large d, because the remaining part of the equation for w_i in system (5.2) is of that order (this follows from our analysis below). As $d \to \infty$, $w \to \bar{w}_e$, so $d(\bar{w}_e - w)$ is essentially a function of either one of the coordinates which enter the term. (This was proved rigorously in [24] for a related set of equations.) Using this, as in [24], we introduce $c(w) = d(\bar{w}_e - w)$.

We derive $\bar{w}_e$ from the definition of $c(w)$:

$$\bar{w}_e = w + \frac{1}{d} c(w). \tag{5.7}$$

Taking the derivative of this equation, we get

$$\frac{d\bar{w}_e}{dt} = \frac{dw}{dt}\left(1 + \frac{1}{d}\frac{dc(w)}{dw}\right). \tag{5.8}$$

Substituting this derivative and $c(w)$ into the third and fourth equation of system (5.2), we can rewrite them in the form

$$\frac{dw}{dt} = \bar{\varepsilon}(\bar{\alpha}_4 g(u) - w) + 2c(w), \tag{5.9}$$

$$\frac{dw}{dt}\left(1 + \frac{1}{d}\frac{dc(w)}{dw}\right) = -\frac{d_e}{d} c(w). \tag{5.10}$$

Excluding dw/dt from these equations, we get

$$[\bar{\varepsilon}(\bar{\alpha}_4 g(u) - w) + 2c(w)]\left(1 + \frac{1}{d}\frac{dc(w)}{dw}\right) = -\frac{d_e}{d} c(w). \tag{5.11}$$

The left-hand side of this equation is $O(\bar{\varepsilon})$. For a nonzero result in the leading order, $O(\bar{\varepsilon})$, we suppose that $\frac{d_e}{d} \sim 1$ and obtain

$$c_0(w) = -\frac{\bar{\varepsilon}(\bar{\alpha}_4 g(u) - w)}{2 + d_e/d}. \tag{5.12}$$

Note that we have not assumed $\bar{\varepsilon}$ small. The above result is valid for $\bar{\varepsilon} \sim 1$ whenever $d \gg \bar{\varepsilon}$.

Substituting $c_0(w)$ for $d(\bar{w}_e - w)$ in (5.2), we have the following three-dimensional system for synchronous oscillations in the limit of large coupling:

$$\begin{aligned} \frac{du}{dt} &= \alpha_1 f(v) - u + \alpha_3 h(w), \\ \frac{dv}{dt} &= \alpha_2 g(u) - v, \\ \frac{dw}{dt} &= \bar{\varepsilon} \frac{d_e}{2d + d_e} \left(\bar{\alpha}_4 g(u) - w \right). \end{aligned} \tag{5.13}$$

Hence, increasing the coupling strength d perturbs the parameter in front of the slow equation, changing the rate of change of the autoinducer, i.e., the slow timescale of the system. It follows from (5.11) that this perturbation is negligible if $d_e \gg d$ (to leading order $c(w) = 0$). But in the intermediate case $d_e \sim d$, the perturbation slows down the oscillations. In the limiting case $d_e \ll d$, (5.11) gives $c_0(w) = -\bar{\varepsilon}(\bar{\alpha}_4 g(u) - w)/2$, and, substituting $d(\bar{w}_e - w)$ in (5.2) by this formula, we obtain $dw/dt = o(\bar{\varepsilon})$. Thus, in the case $d_e \ll d$, the rate of change of the slow variable, w, is decreased by an order of magnitude, and so is the frequency of oscillations.

5.2. Stability of the synchronous solution. Now we examine stability of the solution obtained above with respect to small perturbations of the equalities of corresponding coordinates of the elements. We show that the synchrony may become unstable for large coupling strength. Let us define the perturbations in the following way: $u_i = u + \xi, u_j = u - \xi, v_i = v + \nu, v_j = v - \nu, w_i = w + \zeta, w_j = w - \zeta$, where i and j are any two numbers from 1 to N. Thus, we are perturbing any two elements of the ensemble in such a way that the perturbation does not affect the remaining elements. These perturbations are called transversal (or evaporational [25]) and test stability of the manifold of identical coordinates of the elements. The linearized equations for these perturbations are

$$\begin{aligned} \frac{d\xi}{dt} &= \alpha_1 f'(v)\nu - \xi + \alpha_3 h'(w)\zeta, \\ \frac{d\nu}{dt} &= \alpha_2 g'(u)\xi - \nu, \\ \frac{d\zeta}{dt} &= \bar{\varepsilon} \left(\bar{\alpha}_4 g'(u)\xi - \zeta \right) - 2d\zeta, \end{aligned} \tag{5.14}$$

where u, v, and w are taken in identity manifold with dynamics, governed by system (5.2). We solve this system numerically, calculating its Lyapunov exponents. They reveal stability of the synchronous solution with respect to the transversal perturbations and are therefore called transversal Lyapunov exponents. Figure 5.1 presents curves of the maximal transversal Lyapunov exponent vs. the coupling strength d for several sets of the other parameters. A negative value of the exponent implies transversal stability. The first curve corresponds to a set of parameters for which synchrony remains stable for any coupling strength. The second curve shows the maximal transversal Lyapunov exponent when only the parameter $\bar{\varepsilon}$ is changed. For this case, a fivefold increase in $\bar{\varepsilon}$ causes loss of stability with increasing coupling strength. The third curve illustrates the influence of another parameter on stability: changing α_1 shifts the manifold of slow motion so that the intersection with the nullcline of slow motion is shifted far apart from the extrema of the manifold. This shift makes the limit cycle more symmetric (see Figure 5.2) and leads to stability of this solution for any coupling strength even with a higher value of $\bar{\varepsilon}$, for which the oscillations are not

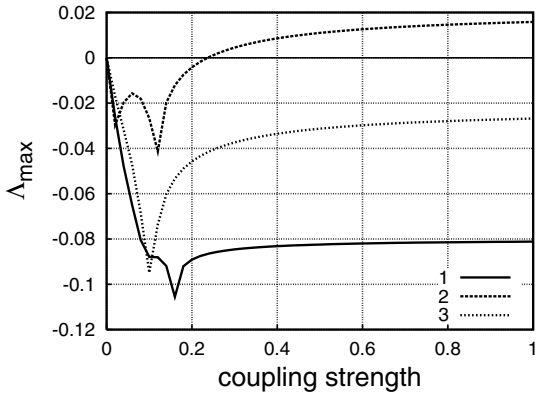


FIG. 5.1. *The largest transversal Lyapunov exponent vs. coupling strength for different sets of the parameters. Curve* (1) *corresponds to* $\bar{\varepsilon} = 0.01, \alpha_1 = 3$. *Curve* (2) *corresponds to* $\bar{\varepsilon} = 0.05, \alpha_1 = 3$, *so the nullclines are the same (see Figure* 5.2(A)*), and shows instability for large coupling strength. Curve* (3) *shows that oscillations can be stable even for such a high value of* $\bar{\varepsilon}$ *(*$\bar{\varepsilon} = 0.05$*) if the limit cycle is more symmetric (*$\alpha_1 = 3.2$ *as in Figure* 5.2(B)*).*

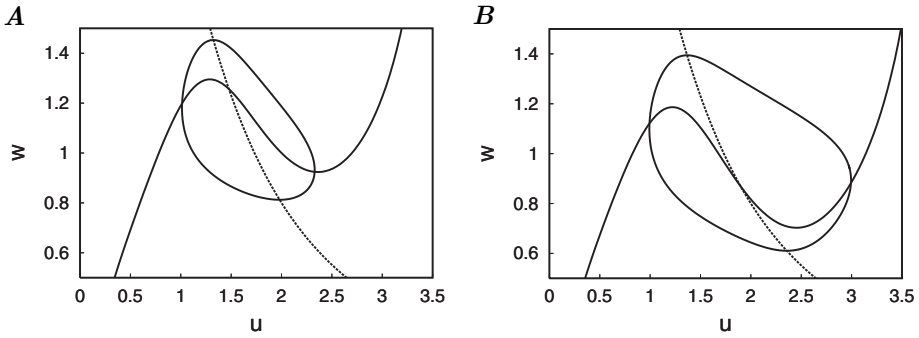


FIG. 5.2. *Position of the nullclines and form of limit cycles for* (A) *unstable and* (B) *stable synchrony. The only different parameter for these two cases is* α_1, *which is equal to* 3 *in the case* (A) *and* 3.2 *in the case* (B). *The other parameters are* $\alpha_2 = 5, \alpha_3 = 1, \bar{\alpha}_4 = 4, \beta = \gamma = \eta = 2$.

of relaxation type ($\bar{\varepsilon} = 0.05$). The illustrations of time series for the ensemble of 20 elements in the cases of stable and unstable identical synchronization solutions are presented in Figure 5.3. Thus, depending on the parameters of the element, we can keep the synchronous solution stable for any positive coupling strength or destabilize it for a large coupling strength.

The dependence of stability of a synchronous solution on parameters of the element can be explained qualitatively. The manifold of identical synchronization, M, has stable and unstable regions. Stability of a trajectory on this manifold is determined by the Lyapunov exponents, which measure whether perturbations decay or grow. As can be seen from computer simulations of this system, the perturbations grow during the fast motion, i.e., in the region, where u corresponds to the negative slope of the manifold of slow motion, $F'(u) < 0$ (see, e.g., Figure 5.2). By contrast, the perturbations decrease during the slow motion. If the time-scales are well separated, then the interval of time on the slow motion is much longer and contraction wins. Increasing $\bar{\varepsilon}$ leads to faster dynamics of the autoinducer (see (5.2)) and decreases intervals of time with slow motion. Thus, a synchronous solution may lose stability with

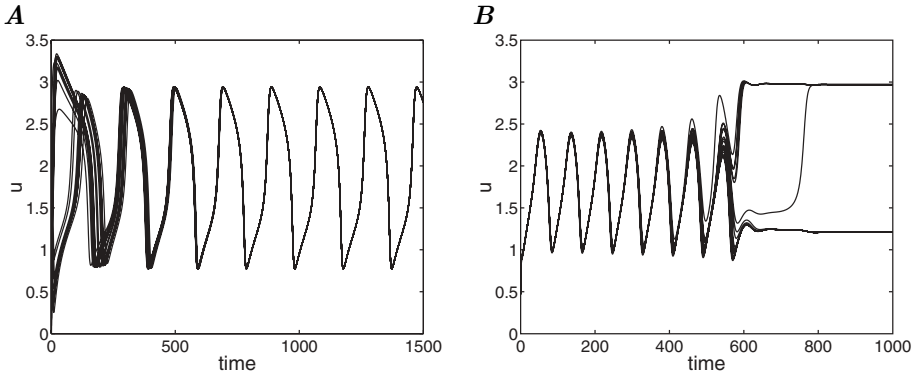


FIG. 5.3. *Examples of time series for the ensemble of* 20 *elements in the cases of* (A) *stable and* (B) *unstable synchronous solution.* $\alpha_1 = 3, \alpha_2 = 5, \alpha_3 = 1, \bar{\alpha}_4 = 4, \beta = \gamma = \eta = 2$*;* (A) $\bar{\varepsilon} = 0.01, d = 0.005$*;* (B) $\bar{\varepsilon} = 0.05, d = 0.3$.

respect to the transversal perturbations when dynamics of the autoinducer becomes faster.

The same argument can be applied to explain dependence of stability on the form of the limit cycle. Given the same time separation ($\bar{\varepsilon}$) for both trajectories in Figure 5.2, in the case (A), the major part of the trajectory lies in the middle region of u, where $F'(u) < 0$. Here, the transversal perturbations grow, giving divergence of the close trajectories from the limit cycle. In the case (B), the limit cycle has much larger parts outside the middle region, which contributes to the decrease of the perturbations and causes convergence in average along the limit cycle.

5.3. Stable equilibria for large coupling strength. In this section we show that large diffusion may cause emerging steady states of the protein concentrations and ceasing of the oscillations in the population. We are going to show existence and stability of new equilibria in the phase space of the ensemble (5.1) for large coupling strength. Taking into account our result on synchronization of this population, the new equilibria may coexist with the stable synchronous periodic solution, dividing the phase space into basins of attraction.

We conduct the analysis analogous to [26] and [27]. Consider for simplicity a pair of the elements

$$
\begin{aligned}
\frac{du_1}{dt} &= \alpha_1 f(v_1) - u_1 + \alpha_3 h(w_1),\\
\frac{dv_1}{dt} &= \alpha_2 g(u_1) - v_1,\\
\frac{dw_1}{dt} &= \bar{\varepsilon}\,(\bar{\alpha}_4 g(u_1) - w_1) + 2d(\bar{w}_e - w_1),\\
\frac{du_2}{dt} &= \alpha_1 f(v_2) - u_2 + \alpha_3 h(w_2),\\
\frac{dv_2}{dt} &= \alpha_2 g(u_2) - v_2,\\
\frac{dw_2}{dt} &= \bar{\varepsilon}\,(\bar{\alpha}_4 g(u_2) - w_2) + 2d(\bar{w}_e - w_2),\\
\frac{d\bar{w}_e}{dt} &= \frac{d_e}{2}(w_1 + w_2 - 2\bar{w}_e).
\end{aligned}
\tag{5.15}
$$

Equilibrium states of the system are given by

$$\begin{aligned}
\alpha_1 f(v_1) - u_1 + \alpha_3 h(w_1) &= 0,\\
\alpha_2 g(u_1) - v_1 &= 0,\\
\bar{\varepsilon}(\bar{\alpha}_4 g(u_1) - w_1) + 2d(\bar{w}_e - w_1) &= 0,\\
\alpha_1 f(v_2) - u_2 + \alpha_3 h(w_2) &= 0,\\
\alpha_2 g(u_2) - v_2 &= 0,\\
\bar{\varepsilon}(\bar{\alpha}_4 g(u_2) - w_2) + 2d(\bar{w}_e - w_2) &= 0,\\
w_1 + w_2 - 2\bar{w}_e &= 0.
\end{aligned} \tag{5.16}$$

Again, we derive v_i from these equations as $v_i = \alpha_2 g(u_i)$, and substituting them into the remaining equations, we can write u_i as a function of w_i: $u_i = F^{-1}(\alpha_3 h(w_i))$, where, as before, $F(u) = u - \alpha_1 f(\alpha_2 g(u))$. The extracellular autoinducer concentration can also be presented as a function of w_i: $\bar{w}_e = (w_1 + w_2)/2$. Since u_i, v_i, and $\bar{w}_e$ are determined by w_i, the equilibria of the system can be found from a two-dimensional system presented in the following form:

$$\begin{aligned}
w_2 &= w_1 - \frac{\bar{\varepsilon}}{d} R(w_1),\\
w_1 &= w_2 - \frac{\bar{\varepsilon}}{d} R(w_2),
\end{aligned} \tag{5.17}$$

where

$$R(w) = \bar{\alpha}_4 g\left(F^{-1}(\alpha_3 h(w))\right) - w. \tag{5.18}$$

This system gives two curves in the (w_1, w_2) plane, intersections of which correspond to equilibria of the pair of elements.

Consider the case where each isolated element displays relaxation oscillations. In particular, let us take the same parameters of the element as in Figure 5.2(A). The curves given by system (5.17) are shown in Figure 5.4 for two different values of the coupling parameter d. Here, with increasing coupling strength, two new intersections of these curves emerge. The intersections correspond to equilibria, the stability of which is shown below.

To explain the emergence of the new equilibria, we divide the dynamics of the system into fast and slow motion, taking $\bar{\varepsilon} \sim d \ll 1$. Consider first the fast subsystem of system (5.15):

$$\begin{aligned}
\frac{du_1}{dt} &= \alpha_1 f(v_1) - u_1 + \alpha_3 h(w_1),\\
\frac{dv_1}{dt} &= \alpha_2 g(u_1) - v_1,\\
\frac{du_2}{dt} &= \alpha_1 f(v_2) - u_2 + \alpha_3 h(w_2),\\
\frac{dv_2}{dt} &= \alpha_2 g(u_2) - v_2,\\
\frac{d\bar{w}_e}{dt} &= \frac{d_e}{2}(w_1 + w_2 - 2\bar{w}_e),
\end{aligned} \tag{5.19}$$

where w_1 and w_2 can be taken to be constant ($\dot{w}_1 = \dot{w}_2 = 0$) and equal to their initial values. This system has three independent parts: for (u_1, v_1), (u_2, v_2), and for $\bar{w}_e$

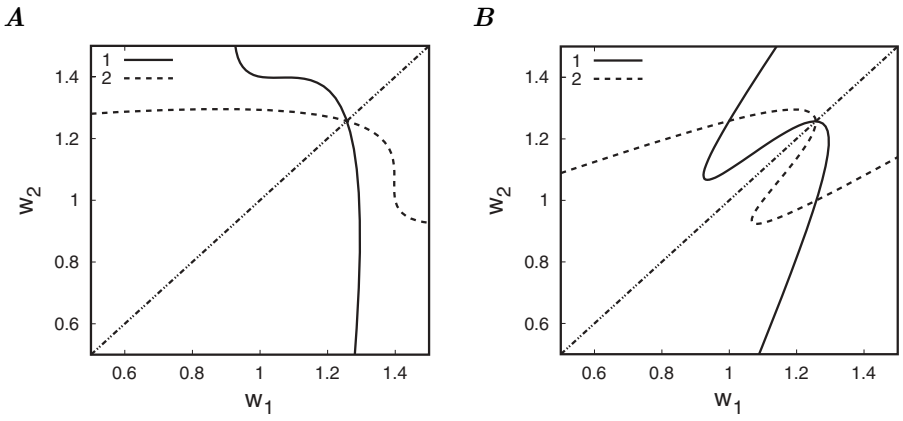


FIG. 5.4. *Equilibrium states of the pair of elements are in the points of intersection of the two curves given by system* (5.17)*, which are plotted for* (A) $\bar{\varepsilon}/d = 2$*;* (B) $\bar{\varepsilon}/d = 0.5$*. Curves* 1 *and* 2 *correspond to the first and the second equations in* (5.17)*. The dashed diagonal line is the manifold of identity of the w coordinates. In case* (B) *we have two intersections of the nullclines outside the diagonal, which are stable equilibria.*

(each part does not include variables from other parts). The equation for $\bar{w}_e$ has the equilibrium $\bar{w}_e = (w_1 + w_2)/2$, which is stable. The remaining two systems for (u_i, v_i) coincide with the fast subsystems for the isolated element (3.2), i.e., the elements are effectively uncoupled with respect to fast motion. Hence, these systems cannot have closed orbits. The only trajectories which attract or repel all others nearby are equilibrium states, defined, as before, by

$$-F(u_i) + \alpha_3 h(w_i) = 0, \qquad v_i = \alpha_2 g(u_i), \quad i = 1, 2. \tag{5.20}$$

The position of the equilibria, depending on w_i, constitutes the manifold of slow motion for the whole system (5.15), where the fast equations do not contribute to the motion, and the motion is governed entirely by the slow subsystem. The manifold for each of the elements of the coupled system (5.15) is given by the same curve (Figure 5.5), which is identical to the one obtained for the isolated element (3.2). Figure 5.5 shows trajectories for the two elements from their initial conditions $(u_1, \alpha_{w,1})$ and $(u_2, \alpha_{w,2})$, where $\alpha_{w,i} = \alpha_3 h(w_i)$. Once the elements come to their manifolds of slow motion, fast motion ceases and the trajectory moves along the manifold, governed by the slow subsystem

$$\begin{aligned} \frac{dw_1}{dt} &= \bar{\varepsilon}\,(\bar{\alpha}_4 g(u_1) - w_1) + d(w_2 - w_1), \\ \frac{dw_2}{dt} &= \bar{\varepsilon}\,(\bar{\alpha}_4 g(u_2) - w_2) + d(w_1 - w_2), \end{aligned} \tag{5.21}$$

where $u_i = F^{-1}(\bar{\alpha}_4 h(w_i)), i = 1, 2$. For $d = 0$, w_i increases along the left-hand branch of $F(u)$ and decreases on the right-hand branch. In the case plotted in Figure 5.5, by providing attraction of the w_i coordinates to each other, the coupling term speeds up motion along the manifold until $w_1 = w_2$. After that, the coupling slows down the motion. If the coupling strength is high enough, it can stop the motion along the manifold, compensating for the slow dynamics of the individual elements, as plotted in Figure 5.5.

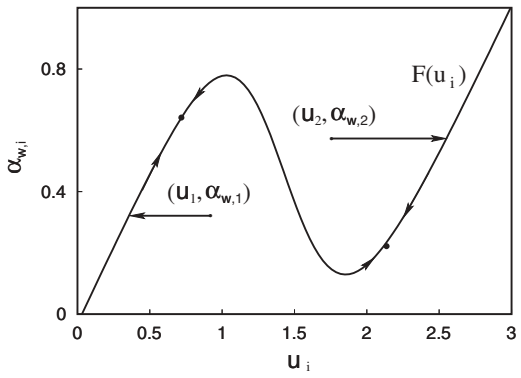


FIG. 5.5. *Fast motion in the pair of elements. The manifolds of slow motion for these two elements coincide with each other. The fast motion attracts the trajectory to one of the outer branches of the manifold. The middle branch is unstable with respect to fast motion.*

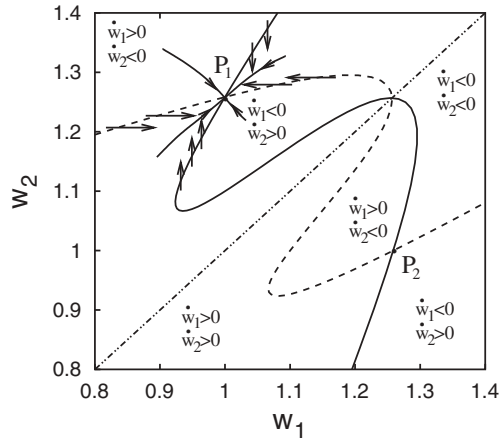


FIG. 5.6. *Phase diagram of slow system* (5.22) *near equilibria* P_1 *and* P_2.

5.4. Stability of the new equilibria. We are going to show that the equilibrium states are stable with respect to both fast and slow motion. Consider first the slow subsystem (5.21), which, using the function $R(w)$ defined in (5.18), can be presented as

$$\begin{aligned} \frac{dw_1}{dt} &= \bar{\varepsilon} R(w_1) + d(w_2 - w_1), \\ \frac{dw_2}{dt} &= \bar{\varepsilon} R(w_2) + d(w_1 - w_2). \end{aligned} \tag{5.22}$$

The curves (5.17), plotted in Figure 5.4, are nullclines of this system. Determining the sign of $\dot{w}_1$ and $\dot{w}_2$ from (5.22) in different regions of the (w_1, w_2) plane, we can qualitatively plot directions of trajectories for the slow subsystem. This gives us the picture in Figure 5.6, which shows stability with respect to slow motion of the equilibrium states P_1 and P_2.

Next consider stability with respect to the fast motion. We know that, in the singular limit $\bar{\varepsilon} \to 0$, the middle branch of the manifold of slow motion for the isolated element is unstable (Figure 5.5). Even in the case when there is an equilibrium in the middle branch and the slow motion converges to it, the fast motion provides divergence (see Figure 3.2(A) curve α_{w3}). Thus, we can determine stability of an equilibrium state with respect to the fast motion based only upon its position on the manifold of slow motion. In particular, if the equilibrium state is situated in the middle branch of the manifold, then it is unstable.

This rule helps us to determine stability of the equilibria in the intersections of the curves in the (w_1, w_2) plane (Figure 5.4). The curves have three monotonic branches, which correspond to the branches of the manifold of slow motion, given by $F(u)$. The middle branch of the manifold, which is unstable with respect to fast motion, corresponds to the middle branch of the curves $w_1(w_2)$ and $w_2(w_1)$. Hence, the equilibrium state in the diagonal in Figure 5.4(A) and (B) is unstable, but the pair of new equilibria in Figure 5.4(B) is stable with respect to fast motion. This suggests the new equilibrium states are stable for large coupling strength.

Stability of the equilibria can be shown rigorously by the standard characteristic equation method. The characteristic equation for the pair of interacting elements (5.15) can be written in the following form:

$$\Omega_1\Omega_2(-d_e-\lambda) - \Omega_1\Delta_2 d_e d - \Omega_2\Delta_1 d_e d = 0. \tag{5.23}$$

Here, $\Delta_i = (1+\lambda)^2 - \alpha_2 g'(u_i)\alpha_1 f'(v_i)$, $\Omega_i = \Delta_i(-\bar{\varepsilon}-2d-\lambda)+(1+\lambda)\alpha_3 h'(w_i)\bar{\varepsilon}\alpha_4 g'(u_i)$. This is a seventh order equation with respect to λ.

To carry out stability analysis for the case considered above in a more rigorous way, we suppose $d \sim \bar{\varepsilon} \ll 1$. We are looking first for eigenvalues of the order one, $\lambda \sim 1$. To leading order of magnitude, (5.23) gives

$$\Delta_1\Delta_2\lambda^2(-d_e-\lambda) = 0.$$

The first root of this equation is $\lambda_1 = -d_e$; the other four roots, $\lambda_{2,3}$ and $\lambda_{4,5}$, come from the quadratic equations $\Delta_1 = 0$ and $\Delta_2 = 0$, respectively. All of them are negative if $1 - \alpha_2 g'(u_i)\alpha_1 f'(v_i) > 0$, which is equivalent to $F'(u_i) > 0$ (see section 3.3). The remaining two eigenvalues, $\lambda_{6,7}$, obtained from this equation are zero. We need to consider a lower order of magnitude of equation (5.23) to determine their signs. Thus next we suppose $\lambda \sim \bar{\varepsilon}$. Then the leading order of (5.23) is $O(\bar{\varepsilon}^2)$:

$$\begin{aligned}
&\Big[F'(u_1)(-\bar{\varepsilon}-2d-\lambda) + \bar{\varepsilon}\alpha_3 h'(w_1)\alpha_4 g'(u_1)\Big] \\
&\quad\times\Big[F'(u_2)(-\bar{\varepsilon}-2d-\lambda) + \bar{\varepsilon}\alpha_3 h'(w_2)\alpha_4 g'(u_2)\Big](-d_e) \\
&\quad-\Big[F'(u_1)(-\bar{\varepsilon}-2d-\lambda) + \bar{\varepsilon}\alpha_3 h'(w_1)\alpha_4 g'(u_1)\Big]F'(u_2)d_e d \\
&\quad-\Big[F'(u_2)(-\bar{\varepsilon}-2d-\lambda) + \bar{\varepsilon}\alpha_3 h'(w_2)\alpha_4 g'(u_2)\Big]F'(u_1)d_e d = 0,
\end{aligned} \tag{5.24}$$

where we take only $O(1)$ terms in Δ_i and $O(\bar{\varepsilon})$ terms in Ω_i. This is a quadratic equation, which can be written in the form $a\lambda^2 + b\lambda + c = 0$ with the coefficients $a = -F'(u_1)F'(u_2)$, $b = -2(\bar{\varepsilon}+d)F'(u_1)F'(u_2)+\bar{\varepsilon}\alpha_3\alpha_4[F'(u_1)h'(w_2)g'(u_2)+F'(u_2)\times h'(w_1)g'(u_1)]$, $c = -(\bar{\varepsilon}+2d)\bar{\varepsilon}F'(u_1)F'(u_2) + (\bar{\varepsilon}+d)\bar{\varepsilon}\alpha_3\alpha_4[F'(u_1)h'(w_2)g'(u_2)+F'(u_2)h'(w_1)g'(u_1)] - \bar{\varepsilon}^2\alpha_3^2\alpha_4^2 h'(w_1)h'(w_2)g'(u_1)g'(u_2)$. We need to show that $\lambda_{6,7} = (-b \pm \sqrt{b^2-4ac})/2a$ are negative or, equivalently, that $-b \pm \sqrt{b^2-4ac} > 0$. Consider the region where all the eigenvalues of the leading order, $O(1)$, are negative, i.e.,

$F'(u_i) > 0$. Then it is obvious that the first coefficient, a, is negative. All three terms of c coefficient are negative because $h'(w_i) \geq 0$ and $g'(u_i) \leq 0$ everywhere. Thus, $ac > 0$, $b^2 - 4ac < b^2$, or $\sqrt{b^2 - 4ac} < |b|$. The coefficient b is also negative; hence the latter inequality gives $\sqrt{b^2 - 4ac} < -b$, or $-b - \sqrt{b^2 - 4ac} > 0$. Thus, all eigenvalues are negative under conditions $F'(u_i) > 0$, showing stability of the equilibria in the outer branches of the function $F(u)$.

Analogously, one can show stability of the equilibria for the case of large or moderate coupling strength d.

5.5. Impact of the collective dynamics. We now discuss how requirements for obtaining oscillations are changed in the case of a population of interacting cells. One of the key experimental problems is to achieve a sufficiently slow dynamics of the autoinducer. The time-scale of autoinducer dynamics is determined by the coefficient in front of the right part of the equation for autoinducer concentration—a rate constant. For the isolated element (3.1), this coefficient is ε; for the system on the manifold of identical synchronization (5.2), it is $\bar{\varepsilon}$. We have shown that the lower the rate constant (the slower dynamics of autoinducer) the larger the oscillatory region (see Figure 3.5(F)). For population dynamics, the rate constant depends on the coupling strength (see, e.g., (5.13)). It follows from (5.11) that the autoinducer dynamics is much slower than in isolated element if $d_e \ll d$ and $d \gg \bar{\varepsilon}$. Hence, we can slow down dynamics of the autoinducer using properties of the collective dynamics. The assumption that $d_e \ll d$ ($d = \frac{D}{2(1+\delta_e/D_e)}$, $d_e = D_e + \delta_e$) is quite plausible for the experiment because D_e is presented usually as $D_e = \frac{D\rho}{1-\rho}$, where ρ is cell density, and $\rho \ll 1$. Then $D_e \ll D$, and if we suppose that $\delta_e \sim D_e$, we come to the inequality $d_e \ll d$. The other assumption ($d \gg \bar{\varepsilon}$) is also plausible because the permeability of the cell membrane to the autoinducer molecules is expected to be relatively high ($D \gg 1$).

Stability of the synchronous solution also depends on the time-scale of the autoinducer dynamics. We have shown that synchrony may be stable for a small rate constant $\bar{\varepsilon}$ and unstable for larger $\bar{\varepsilon}$ (see Figure 5.1). Those computations were made for $d_e \gg d$ so that the period of oscillations was not changed significantly with increasing coupling strength. Now, if $d_e \ll d$, increasing coupling strength also slows down the oscillations (5.13), causing stabilization of the synchronous solution (data not shown).

On the other hand, increasing coupling strength also causes emergence of the new equilibria. The higher the coupling strength, the larger the domains of attraction of the new equilibria and the larger the probability of obtaining a steady state instead of synchronous oscillations in the experiment. The effect is very strong because the coupling strength sufficient for the formation of the equilibria is of the same order of magnitude as $\bar{\varepsilon}$. To avoid formation of the equilibria, we need to keep $\bar{\varepsilon}$ larger than d, which contradicts the conditions used above. Thus, our theoretical study of the simplified model (5.1) predicts a potential problem for the experimental implementation of the synchronous oscillatory dynamics.

6. Solving experimental problems (discussion). In this section we summarize the results of our investigations and discuss the conditions required for population synchronous oscillations in the light of constraints imposed by experimental considerations. Our study of a simplified model shows that population synchronous oscillations are theoretically possible. However, there may be some difficulties in achieving population synchronous oscillations experimentally. First of all, a strong

interaction between cells (e.g., high permeability of the membrane to the autoinducer) may result in the suppression of synchronous oscillations and a transition to a stable heterogeneous population state where individual cells are locked in different stable equilibrium states. On the other hand, if the cell-to-cell interactions are too weak, individual cells may oscillate but will be unable to achieve synchrony. Therefore, the parameters that determine the coupling strength between cells must be finely adjusted. Cell-to-cell variations in parameter values and initial conditions increase the likelihood that the population synchronous oscillation will be suppressed. As our study has shown, mere difference in initial conditions for different cells may be sufficient to obtain a stationary population state and suppression of individual cellular oscillators, even though the synchronous solution is asymptotically stable.

Fortunately, the system provides a possibility for attracting a very broad distribution of initial conditions to the synchronous solution. This requires the dynamics of the extracellular autoinducer to be much slower than the intracellular ones, which corresponds to the condition $d_e \ll d$ discussed in the end of the previous section. In this case (and for large coupling strength d), the motion from the initial conditions starts with relaxation of the concentrations of the intracellular autoinducer in different cells toward the state where the concentrations are equal to the extracellular concentration. If the extracellular autoinducer is washed out at the onset of the experiment, then the initial state will have the extracellular concentration that is close to zero. Thus, the concentrations inside the cells approach a low value. Once the low concentration has been achieved, the concentrations of the repressor proteins inside all cells approach the same state: low u and high v. This state is in the domain of attraction of the synchronous solution, and synchrony can be achieved even when other attractors exist.

Moreover, if the parameters of individual cells are different, the identical synchronization solution does not exist at all. Our computational study of the model of the population shows that the introduction of inhomogeneity in parameters increases the probability of obtaining a stationary state starting from random initial conditions. Generally speaking, a larger variation in the parameter values of individual elements (e.g., a larger difference in the individual cellular oscillators and high cell-to-cell variability) will decrease the likelihood of observing a population synchronized oscillatory state. This is particularly important since many engineered gene regulatory networks, including the toggle switch and the Lux-based cell-to-cell communication system, are carried on self-replicating plasmids. The number of plasmids per cell (and hence the number of genes and promoters they carry) is know to vary quite dramatically [20], and it will probably be necessary to minimize this source of cell-cell variability in order to achieve population synchronous oscillations. This could be achieved by using plasmids with more elaborate mechanisms of copy-number and partitioning control or by integrating the engineered network into the bacterial chromosome.

In addition to parameter differences caused by fluctuations in the number of genes per cell, there are other sources of cell-to-cell variability that cannot easily be minimized. These include variation in growth rates, differences in the concentration of polymerases and of ribosomes, and others. Moreover, genetic manipulations usually introduce rather coarse changes, and it is generally difficult to fine-tune the parameter values that govern the dynamics of individual cells. Therefore, to increase the likelihood of population synchronous oscillations, it is important to have a rather large parameter region where oscillatory dynamics for the isolated element is observed.

We have shown that this can be achieved by increasing the maximal rate (α_3)

of synthesis of one of the repressors (u) from the promoter that is regulated by the autoinducer. The expansion of the oscillatory region also requires an increase of the maximal rate (α_2) of synthesis of the second repressor (v). Our investigation indicates that an increased nonlinearity, i.e., the parameters for promoter cooperativity β, γ, and η, does not significantly expand the region of parameter space where oscillations in isolated elements can be observed. Increasing the value of β (cooperativity for promoter that synthesizes u repressor) actually causes a contraction of the oscillatory region. In addition, to achieve the largest oscillatory region possible, the rate of decay of the autoinducer ε must be as small as possible (of a smaller order of magnitude, $\varepsilon \ll 1$). This may pose an experimental challenge because it implies that the rates of autoinducer synthesis ($\varepsilon \times \alpha_4$) and decay ($\varepsilon$) are at least an order of magnitude lower than the rates of the repressors synthesis (α_1 and α_2) and decay (which is equal to 1). While the maximal rate of autoinducer synthesis could be made low by genetic manipulations, the rate of decay is primarily determined by dilution due to the cell growth. This dilution of course affects all cellular components identically. Moreover, the autoinducer is able to penetrate the cell membrane, and the value of the rate parameter ε depends linearly on the diffusion coefficient D of the autoinducer, which may be large.

For mathematical tractability, our simplified model ignores the fact that the synthesis of the autoinducer is a two-step process that requires the synthesis of the protein LuxI. Taking this step into consideration partially solves the problem with the smallness of ε. First, the additional step introduces a delay in production of the autoinducer. Second, the degradation rate of LuxI is the same as that of the repressors since the LuxI protein is unable to penetrate the cell membrane. As shown by numerical simulations, it is possible to achieve oscillations when the decay rate of the autoinducer is only three times smaller than the decay rates of the repressor proteins. This is in contrast to the simplified model, where an order of magnitude difference was required. In other words, the simplified model may actually underestimate the likelihood of achieving oscillatory dynamics. Further improvement of the gene network toward the ability to oscillate was achieved by the introduction of a promoter that is repressible by the autoinducer and synthesizes protein v, i.e., a negative feedback from the autoinducer to one of the repressors (see section 4.2). Such a promoter has previously been described in the literature [28]. Our simulations demonstrate that the oscillatory region can be substantially increased when this negative feedback is added. It is our belief that this slight variation in the architecture of gene regulatory network would be very useful experimentally as it increases the robustness of single-cell oscillations quite dramatically.

The problem of new equilibria also can be overcome in the experiment if we take into account the LuxI synthesis step and/or introduce the additional network connectivity. In the simplified model, the degradation rate constant of the autoinducer $\bar{\varepsilon}$ must be larger than the coupling coefficient d to avoid the presence of equilibria. On the other hand, $\bar{\varepsilon}$ must be small to provide relaxation type of oscillations, and d must be large enough to synchronize the ensemble. This contradiction is resolved if the second slow process, LuxI synthesis, is taken into account. This introduces an additional delay in production of the autoinducer and allows the rate constants of the autoinducer to take higher values without losing the oscillations. Moreover, these rate constants are expected to take much higher values than the rate constants of the LuxI synthesis because the effective rate of degradation for the autoinducer includes the diffusion coefficient D ($\bar{\varepsilon}_w = D + \delta - \frac{D}{(1+\delta_e/D_e)}$), but the rate of degradation

for LuxI does not ($\bar{\varepsilon}_x = \delta$). Hence, the LuxI dynamics has the function of slowing down the oscillations; its rate constant must be several times smaller than those of the other proteins. The rate constants of the autoinducer can be chosen large enough so that the degradation rate constant is larger than the coupling coefficient: $\bar{\varepsilon}_w > d$. Thus, it is theoretically possible to simultaneously achieve absence of the equilibria and strongly enough attracting synchronous oscillations.

Appendix. Derivation of parametric formula for approximations of bifurcation boundaries.

Appendix A. The saddle-node bifurcation for the toggle switch. We are solving the system which gives the curve of the saddle-node bifurcation for the fast subsystem in the absence of the autoinducer:

$$F(u) = 0, \qquad F'(u) = 0.$$

Substituting the function $F(u)$, defined as (3.3), we can rewrite it in the form

$$\frac{\alpha_1}{1+\left(\frac{\alpha_2}{1+u^\gamma}\right)^\beta} = u, \qquad \left(1+\left(\frac{\alpha_2}{1+u^\gamma}\right)^\beta\right)^2 = \left(\frac{\alpha_2}{1+u^\gamma}\right)^\beta \frac{\alpha_1\beta\gamma u^{\gamma-1}}{1+u^\gamma}. \tag{A.1}$$

We define $R = (\frac{\alpha_2}{1+u^\gamma})^\beta$. From the first equation of the latter system, we have $R = \frac{\alpha_1}{u} - 1$. Using this combination in the second equation of system (A.1), we obtain

$$\left(\frac{\alpha_1}{u}\right)^2 = \left(\frac{\alpha_1}{u} - 1\right)\frac{\alpha_1\beta\gamma u^{\gamma-1}}{1+u^\gamma}, \tag{A.2}$$

or

$$\alpha_1 = \frac{\beta\gamma u^{\gamma+1}}{\beta\gamma u^\gamma - (1+u^\gamma)}. \tag{A.3}$$

R has appeared to be a function of u and depends on γ and β, but it does not depend on α_1 and α_2:

$$R(u) = \frac{\beta\gamma u^\gamma}{\beta\gamma u^\gamma - (1+u^\gamma)} - 1. \tag{A.4}$$

Then, from the definition of R, we have

$$\alpha_2 = (R(u))^{1/\beta}(1+u^\gamma). \tag{A.5}$$

We introduce a parameter $r > 0$, replacing u in the obtained formulas, to show that we have obtained a bifurcation boundary in the space (α_1, α_2) parametrized by an independent parameter. The resulting formulas of the saddle-node bifurcation curve for the fast subsystem in absence of the autoinducer are

$$\begin{aligned} \alpha_1^c &= \frac{\beta\gamma r^{\gamma+1}}{1+r^\gamma} \Big/ \left(\frac{\beta\gamma r^\gamma}{1+r^\gamma} - 1\right), \\ \alpha_2^c &= (1+r^\gamma)\left(\frac{\beta\gamma r^\gamma}{1+r^\gamma} \Big/ \left(\frac{\beta\gamma r^\gamma}{1+r^\gamma} - 1\right) - 1\right)^{1/\beta}. \end{aligned} \tag{A.6}$$

Appendix B. Merging extrema of the function $F(u)$. We have the following condition for merging extrema of a function:

$$F^{'}(u) = 0, \quad F^{''}(u) = 0,$$

which, in our case, takes the form

$$\left(1 + \left(\frac{\alpha_2}{1+u^\gamma}\right)^\beta\right) = \left(\frac{\alpha_2}{1+u^\gamma}\right)^\beta \frac{\alpha_1\beta\gamma u^{\gamma-1}}{1+u^\gamma},$$

$$\left(\frac{\alpha_2}{1+u^\gamma}\right)^\beta = -\frac{(\gamma-1) - u^\gamma(1+\beta\gamma)}{(\gamma-1) + u^\gamma(\beta\gamma-1)}. \tag{B.1}$$

We define $R = (\frac{\alpha_2}{1+u^\gamma})^\beta$. From the second equation of the latter system, we have

$$R = -\frac{(\gamma-1) - u^\gamma(1+\beta\gamma)}{(\gamma-1) + u^\gamma(\beta\gamma-1)}. \tag{B.2}$$

This formula shows again that R has appeared to be a function of u, which depends on γ and β, but does not depend on α_1 and α_2. Performing analogous calculations as in the previous case and introducing an independent parameter r instead of u, we obtain the following parametric representation of the boundary:

$$\alpha_1^m = \left(1 + R_1(r)\right)^2 \left(1 + r^\gamma\right) \Big/ \left(R_1(r)\gamma\beta r^{\gamma-1}\right),$$

$$\alpha_2^m = \left(R_1(r)\right)^{1/\beta} \left(1 + r^\gamma\right), \tag{B.3}$$

where r is the parameter and

$$R_1(r) = -\frac{(\gamma-1) - r^\gamma(1+\beta\gamma)}{(\gamma-1) - r^\gamma(1-\beta\gamma)}. \tag{B.4}$$

Appendix C. The approximation for the Andronov–Hopf bifurcation. The Andronov–Hopf bifurcation curve for vanishing ε is approximated by the curve in the parameter space that corresponds to the intersection of the nullcline of slow motion with the manifold of slow motion in an extremum of the latter:

$$-F(u) + \alpha_3 h(\alpha_4 g(u)) = 0, \quad F'(u) = 0.$$

Let us solve this condition with respect to the parameters α_1 and α_2. This case is very similar to the condition of the boundary for the saddle-node bifurcation for the fast subsystem in the absence of the autoinducer (see Appendix A) because the only difference is the additive term $\alpha_3 h(\alpha_4 g(u))$. This term depends only on u and parameters α_3 and α_4, which gives us a possibility to apply the same steps as in the previous case, defining $R_3(u) = \alpha_3 h(\alpha_4 g(u))$. These calculations give the following curve in (α_1, α_2) parameter plane:

$$\alpha_1^H = \left(1 + R_2(r)\right) / \left(r - R_3(r)\right), \quad \alpha_2^H = \left(R_2(r)\right)^{1/\beta} \left(1 + r^\gamma\right), \tag{C.1}$$

where r is an independent parameter and

$$R_2(r) = \frac{\beta\gamma r^{\gamma-1}\left(r - R_3(r)\right)}{\left(\beta\gamma r^\gamma - r^\gamma - \beta\gamma r^{\gamma-1} R_3(r) - 1\right)} - 1,$$

$$R_3(r) = \alpha_3 \left(\frac{\alpha_4}{1+r^\gamma}\right)^\eta \Big/ \left(1 + \left(\frac{\alpha_4}{1+r^\gamma}\right)^\eta\right). \tag{C.2}$$

Acknowledgments. The authors thank James Collins, Tasso Kaper, and Horacio Rotsteine for useful discussions.

REFERENCES

[1] T. S. Gardner, C. R. Cantor, and J. J. Collins, *Construction of a genetic toggle switch in Escherichia coli*, Nature, 403 (2000), pp. 339–342.
[2] M. B. Elowitz and S. Leibler, *A synthetic oscillatory network of transcriptional regulators*, Nature, 403 (2000), pp. 335–338.
[3] C. C. Guet, M. B. Elowitz, W. Hsing, and S. Leibler, *Combinatorial synthesis of genetic networks*, Science, 296 (2002), pp. 1466–1470.
[4] Y. Yokobayashi, R. Weiss, and F. H. Arnold, *Directed evolution of a genetic circuit*, Proc. Natl. Acad. Sci. USA, 99 (2002), pp. 16587–16591.
[5] M. R. Atkinson, M. A. Savageau, J. T. Myers, and A. J. Ninfa, *Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in Escherichia coli*, Cell. 113 (2003), pp. 597–607.
[6] W. Weber and M. Fussenegger, *Artificial mammalian gene regulation networks—novel approaches for gene therapy and bioengineering*, J. Biotechnol., 98 (2002) pp. 161–187.
[7] J. Hasty, D. McMillen, and J. J. Collins, *Engineered gene circuits*, Nature, 420 (2002), pp. 224–230.
[8] M. L. Simpson, G. S. Sayler, J. T. Fleming, and B. Applegate, *Whole-cell biocomputing*, Trends Biotechnol., 19 (2001), pp. 317–323.
[9] M. Kærn, W. Blake, and J. J. Collins, *The engineering of gene regulatory networks*, Ann. Rev. Biomed. Eng., 5 (2003), pp. 179–206.
[10] H. Kobayashi, M. Kærn, M. Araki, K. Chung, T. S. Gardner, C. R. Cantor, and J. J. Collins, *Programmable cells: Interfacing natural and engineered gene networks*, Proc. Natl. Acad. Sci. USA, 101 (2004), pp. 8414–8419.
[11] R. Weiss and T. Knight, *Engineered communications for microbial robotics*, in DNA6: Sixth International Meeting on DNA Based Computers, DNA 2000, Boston, MA, Springer-Verlag, New York, 2000, pp. 1–16.
[12] L. You, R. S. Cox 3rd, R. Weiss, and F. H. Arnold, *Programmed population control by cell-cell communication and regulated killing*, Nature, 428 (2004), pp. 868–871.
[13] S. Basu, R. Mehreja, S. Thiberge, M. T. Chen, and R. Weiss, *Spatiotemporal control of gene expression with pulse-generating networks*, Proc. Natl. Acad. Sci. USA, 101 (2004), pp. 6355–6360.
[14] C. Fuqua and P. E. Greenberg, *Listening in on bacteria: Acyl-homoserine lactone signaling*, Nat. Rev. Mol. Cell Biol., 3 (2002), pp. 685–695.
[15] M. K. Winson, S. Swift, L. Fish, J. P. Throup, F. Jorgensen, S. R. Chhabra, B. W. Bycroft, P. Williams, and G. S. Stewart, *Construction and analysis of luxCDABE-based plasmid sensors for investigating N-acyl homoserine lactone-mediated quorum sensing*, FEMS Microbiol. Lett., 163 (1998), pp. 185–192.
[16] M. Burmolle, L. H. Hansen, G. Oregaard, and S. J. Sorensen, *Presence of N-acyl homoserine lactones in soil detected by a whole-cell biosensor and flow cytometry*, Microb. Ecol., 45 (2003), pp. 226–236.
[17] J. B. Andersen, A. Heydorn, M. Hentzer, L. Eberl, O. Geisenberger, B. B. Christensen, S. Molin, and M. Givskov, *Gfp-based N-acyl homoserine-lactone sensor systems for detection of bacterial communication*, Appl. Environ. Microbiol., 67 (2001), pp. 575–585.
[18] D. McMillen, N. Kopell, J. Hasty, and J. J. Collins, *Synchronizing genetic relaxation oscillators by intercell signaling*, Proc. Natl. Acad. Sci. USA, 99 (2002), pp. 679–684.
[19] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden, *Regulation of noise in the expression of a single gene*, Nat. Gene., 31 (2002), pp. 69–73.
[20] J. Paulsson and M. Ehrenberg, *Noise in a minimal regulatory network: Plasmid copy number control*, Q. Rev. Biophys., 34 (2001), pp. 1–59.
[21] P. S. Swain, M. B. Elowitz, and E. D. Siggia, *Intrinsic and extrinsic contributions to stochasticity in gene expression*, Proc. Natl. Acad. Sci. USA, 99 (2002), pp. 12795–12800.
[22] T. F. Weiss,*Cellular Biophysics*, Vol. 1, MIT Press, Cambridge, MA, 1986.
[23] L. Perko, *Differential Equations and Dynamical Systems*, Springer-Verlag, New York, 1991.
[24] G. S. Medvedev and N. Kopell, *Synchronization and transient dynamics in the chains of electrically coupled FitzHugh–Nagumo oscillators*, SIAM J. Appl. Math., 61 (2001), pp. 1762–1801.

[25] A. Pikovsky, O. Popovuch, and Yu. Maistrenko, *Resolving clusters in chaotic ensembles of globally coupled identical oscillators*, Phys. Rev. Lett., 87 (2001), pp. 044102-1–044102-4.

[26] V. N. Belykh, I. V. Belykh, and M. Hasler, *Hierarchy and stability of partially synchronous oscillations of diffusively coupled dynamical systems*, Phys. Rev. E (3), (2000), pp. 6332–6345.

[27] N. Kopell, L. F. Abbott, and C. Soto-Trevino, *On the behaviour of a neural oscillator electrically coupled to a bistable element*, Phys. D, 121 (1998), pp. 367–395.

[28] K. A. Egland and E. P. Greenberg, *Conversion of the vibrio fischeri transcriptional activator,* LuxR, *to a repressor*, J. Bacteriol., 182 (2000), pp. 805–811.

SIAM J. APPL. MATH.
Vol. 65, No. 2, pp. 426–457

THE RIEMANN PROBLEM FOR REVERSIBLE REACTIVE FLOWS WITH METASTABILITY*

ANDREA CORLI† AND HAITAO FAN‡

Abstract. A hyperbolic model for dynamic phase transitions is studied. The model involves three phases: liquid, vapor, and a mixture of them. Metastable regions are present both in the liquid and in the vapor phase.

Results on the behavior of traveling wave profiles of the model, involving viscosity, species diffusion and relaxation, are obtained. These behaviors are consistent with physical intuitions. Admissibility criteria (kinetic relations) that mimic the behavior of traveling wave profiles are then proposed. Admissible basic waves of the model are liquefaction, evaporation, and isobaric waves, in addition to Lax shock and rarefaction waves. Based on these waves, solutions of the Riemann problem for the model are constructed for general Riemann initial data. Most of the physical phenomena are embodied in the solver. For some Riemann initial data solutions are expected to be nonunique. Which solutions actually appear depends on whether nucleation already occurred or not. The model admits both solutions, as it should.

The model also has two other types of waves, collapsing and explosion waves. More complicated solutions involving these two waves are also proposed and discussed.

Key words. Riemann problem, phase transitions, hyperbolic conservation laws

AMS subject classifications. 35L65, 35L67, 76T30

DOI. 10.1137/S0036139903429671

1. Introduction.

In the study of liquid-vapor phase transitions in fluids, one often encounters an intermediate regime where the pure phases are mixed together. The pure phase regions contain metastable subregions. The aim of this paper is to study a simple hyperbolic model of phase transitions that involves all three configurations above as well as metastability. The model in one space dimension for isothermal flow in Lagrangian coordinates is

$$\begin{cases} v_t - u_x = 0, \\ u_t + p(v,\lambda)_x = \epsilon u_{xx}, \\ \lambda_t = \dfrac{a}{\epsilon}(p - p_e)\lambda(\lambda - 1) + b\epsilon\lambda_{xx}. \end{cases} \tag{1.1}$$

Here v denotes the specific volume, u the velocity, $p = p(v,\lambda)$ the pressure, and p_e a fixed equilibrium pressure. The number ϵ denotes a positive viscosity coefficient, and a and b denote two real positive parameters; the ratio ϵ/a is the typical reaction time. The quantity $\lambda \in [0,1]$ is the mass density fraction of vapor in the fluid; therefore $\lambda = 0$ identifies a liquid regime, $\lambda = 1$ a vapor regime, and $\lambda \in (0,1)$ a mixture of two pure phases. The system (1.1) is shown, in [9, 10], to exhibit all major wave patterns observed in shock tube experiments on retrograde fluids [5, 15].

We assume that the pressure p is defined in $(0,+\infty) \times [0,1]$ and satisfies the following essential assumptions:

$$p_v < 0, \qquad p_{vv} > 0, \tag{1.2}$$

*Received by the editors June 3, 2003; accepted for publication (in revised form) February 19, 2004; published electronically December 16, 2004.
http://www.siam.org/journals/siap/65-2/42967.html
†Department of Mathematics, University of Ferrara, Via Machiavelli 35, 44100 Ferrara, Italy (crl@unife.it).
‡Department of Mathematics, Georgetown University, Washington, DC 20057-0996 (fan@math.georgetown.edu).

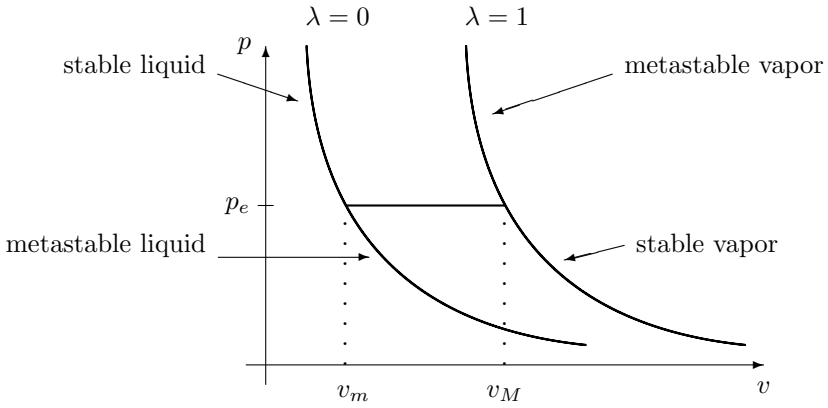


FIG. 1. *Pressure curves.*

$$p_\lambda > 0. \tag{1.3}$$

However, in order to simplify the analysis below, we also require

$$p > 0, \quad \lim_{v \to 0+} p(v, \lambda) = +\infty, \tag{1.4}$$

$$\int^{+\infty} \sqrt{-p_v(v, \lambda)}\, dv = +\infty. \tag{1.5}$$

Both assumptions (1.4) and (1.5) could be dropped by modifying slightly the following. In particular (1.5) prevents the formation of vacuum. An example of pressure satisfying (1.2)–(1.5) is

$$p(v, \lambda) = \frac{1 + \lambda}{v}. \tag{1.6}$$

We define on the v axis the points v_m, respectively, v_M, as the abscissas of the intersections of the curves $p = p(v, 0)$, $p = p_e$ and $p = p(v, 1)$, $p = p_e$; see Figure 1. The part of the graph of the function $p = p(v, 0)$ lying below the line $p = p_e$ is called the *metastable liquid region*, while that of function $p = p(v, 1)$ lying above the line $p = p_e$ is called the *metastable vapor region*. For example, the liquid in metastable liquid region tends to vaporize. But the evaporation will not start until seeds for evaporation, typically tiny vapor bubbles or particles of impurities, are present. The remaining parts of the curves are the *stable liquid* and *stable vapor region*.

Consider then the following initial value problem for system (1.1):

$$\begin{cases} v_t - u_x = 0, \\ u_t + p(v, \lambda)_x = \epsilon u_{xx}, \\ \lambda_t = \dfrac{a}{\epsilon}(p - p_e)\lambda(\lambda - 1) + b\epsilon\lambda_{xx}, \\ (u, v, \lambda)(x, 0) = (u_0, v_0, \lambda_0)(x) = \begin{cases} (u_-, v_-, \lambda_-) & \text{if } x < 0, \\ (u_+, v_+, \lambda_+) & \text{if } x > 0. \end{cases} \end{cases} \tag{1.7}$$

Assume that the problem (1.7) has a solution $(u^\epsilon, v^\epsilon, \lambda^\epsilon)(x, t)$; assume moreover that there is a sequence ϵ_n, $n = 1, 2, \ldots$, with $\epsilon_n \to 0+$ as $n \to \infty$, such that the limit

$$(u, v, \lambda)(x, t) := \lim_{n \to \infty} (u^{\epsilon_n}, v^{\epsilon_n}, \lambda^{\epsilon_n})(x, t)$$

exists in some strong sense, such as pointwise. Then the limit satisfies the Riemann problem

$$\begin{cases} v_t - u_x = 0, \\ u_t + p(v,\lambda)_x = 0, \\ (p(v,\lambda) - p_e)\,\lambda(\lambda-1) = 0, \\ (u,v,\lambda)(x,0) = (u_0,v_0,\lambda_0)(x). \end{cases} \tag{1.8}$$

Remark that under assumption (1.2), the first two equations of (1.8) form a strictly hyperbolic system of conservation laws with eigenvalues $-\sqrt{-p_v(v,\lambda)}$, $\sqrt{-p_v(v,\lambda)}$. The third equation imposes some constraints on the states under consideration. The Riemann problem (1.8) is the object of our investigations. Remark that analogously to [2, 3] the first three equations in (1.8) can be understood by eliminating λ as

$$\begin{cases} v_t - u_x = 0, \\ u_t + p(v,0)_x = 0, \\ (v,u) \in \Omega_0, \end{cases} \qquad \begin{cases} v_t - u_x = 0, \\ u_t = 0, \\ (v,u) \in \Omega_*, \end{cases} \qquad \begin{cases} v_t - u_x = 0, \\ u_t + p(v,1)_x = 0, \\ (v,u) \in \Omega_1, \end{cases} \tag{1.9}$$

where $\Omega_0 = \Omega_1 = (0,+\infty)\times(-\infty,+\infty)$, $\Omega_* = [v_m, v_M]\times(-\infty,+\infty)$.

The structure of solutions of (1.8) is easier to obtain than that of (1.7). Nevertheless, we want to keep track as much as possible of the derivation of the latter from the former system, and we shall focus on solutions of (1.8) that are some type of strong limits of solutions of (1.7) as $\epsilon_n \to 0$ for some sequence $\{\epsilon_n\}$.

From the third equation in (1.8) it is clear that piecewise smooth solutions of (1.8) consist of smooth pieces satisfying either

$$\lambda = 0 \quad \text{or} \quad \lambda = 1 \tag{1.10}$$

or

$$p = p_e. \tag{1.11}$$

In order to have solutions to the Riemann problem (1.8) which mimic as far as possible solutions of (1.7), we choose waves of (1.8) which have traveling wave profiles of (1.7). This choice is a way to impose kinetic relations [1, 12], which imposes admissibility restrictions on waves of conservation laws. However, the results about traveling waves at our disposal (see [7, 9]) are not sufficient to guarantee the existence of a solution for every set of initial data. Nevertheless, the behavior of traveling waves, carrying physical meanings, can serve as a guide for kinetic relations for (1.8). To guarantee the existence of solutions for general Riemann initial data, we shall supplement (1.8) with kinetic relations whose behaviors are consistent with the relevant behavior of traveling waves of (1.7).

One should not expect one kinetic relation that is independent of time to be valid for all time $t > 0$. Consider Shearer's type of nonunique solutions [13]: The Riemann initial data represents a tube of metastable vapor, i.e., $\lambda = 1$, $p > p_e$, flowing at a given speed towards the center. For this type of data, the system (1.8) has two Riemann solutions, one without phase boundary and the other with two phase boundaries moving apart from the center. The center is occupied by liquid. There are studies utilizing different kinetic relations that favor one of them over the other. We want to say that both solutions are good, but good at different time periods: As time t starts to increase from 0, there is no seed, such as a liquid drop, to start

the condensation; the vapor will stay as vapor for a sizable period of time. Thus, at the early stage, the solution without phase boundary is appropriate. As $t \to \infty$, liquid drops will form through nucleation and the condensation will eventually happen, resulting in phase boundaries, and the fluid will eventually settle down in the stable phase. This clearly shows that the solution without phase boundary is valid only up to a certain finite time, so is the kinetic relation that picks this solution. The process as t increases from 0 is as follows: After some time, a liquid drop is formed though a random fluctuation (or nucleation) process. This liquid drop serves as a seed for condensation. Since the surrounding vapor is in the metastable region, further condensation will occur around this liquid drop, resulting in the growth of the liquid or liquid/vapor mixture region. For this time period, the solution with two phase boundaries moving out, with the liquid drop in the center and metastable vapor on outer sides, is appropriate. This example illustrates that for a different time period, different Riemann solvers should be employed. Thus, any admissibility criterion that favors one of the solutions over the other is applicable only for one of time intervals:

$$[0, T_{\text{initiation of the 1st drop}}),$$

$$[T_{\text{initiation of the 1st drop}}, T_{\text{initiation of the 2nd drop}}), \dots$$

The nucleation is a random process, and hence $T_{\text{initiation of the 1st drop}}$ is a random variable. Since the model does not include the nucleation mechanism directly in the equations, the model should allow both solutions, with additional admissibility criterion stating when and which one of them is admissible.

In this paper, we proposed the admissibility criterion applicable in the first time interval. Under this kinetic condition, we can finally prove the existence of a global Riemann solver which singles out solutions uniquely.

Riemann solutions involve many kinds of elementary waves. First of all, liquefaction and evaporation waves may arise in the connection of two different pure phases; they can be either subsonic or sonic. In the latter case, our solver admits rarefaction waves attached to them. This pattern appears also in solution of systems where genuinely nonlinearity fails [12] as well as in combustion theory [11, 2]. A selection criterion is imposed to both liquefaction and evaporation waves in order to find a unique solution to the Riemann problem; if profiles exist, the criterion amounts to choosing the wave traveling with the slowest speed (in absolute value). Moreover it is required that, if we fix a state in back of the phase transition, then the speed of the phase boundary increases if $|p(v_-, \lambda_-) - p_e|$ increases; here (v_-, λ_-) is the state in the back. This requirement has a physical ground and is satisfied by the traveling wave profile having the slowest speed. We refer to [3] for a Riemann solver that applies to deflagration waves having both rarefaction waves attached to a sonic phase boundary and a superimposed kinetic condition. Liquefaction and evaporation waves may arise also in the connection of a pure phase with a mixture phase; also in this case they can be either subsonic or sonic. No selection condition is needed for this case.

Another kind of waves are the isobaric waves, along which the pressure equals the equilibrium pressure p_e; they are stationary and require no kinetic condition. Lax shocks and rarefaction waves in the pure phases complete the set of elementary waves.

This paper is organized as follows. In section 2 we list all possible elementary waves of (1.8). They are reacting or nonreacting shock waves, nonreacting rarefaction waves, and isobaric waves. The existence of traveling wave profiles to some reacting shocks, namely liquefaction and evaporation waves, was established in [7, 9]; in

section 3 we continue the analysis started with those papers. We show first that liquefaction (and evaporation) profiles are monotone; moreover, if the state v_- in the back is fixed, then among these profiles there is a slowest one. Our main result in this section is that the speed of the slowest profile is a decreasing (respectively, increasing) function of v_-. Kinetic conditions are given in section 4. In section 5, using the elementary waves and the related kinetic conditions, we construct solutions to the Riemann problem (1.8) for arbitrary Riemann data; comments about our Riemann solver are gathered at the end of this section. Numerical works in [10] indicated the existence of another two kinds of reacting shocks. They are explosion waves and collapsing waves. In section 6 we show through some examples how these waves may be used to obtain a different Riemann solver.

2. Basic admissible waves. In this section we introduce the elementary waves to be used in the following. We state as well some results proved elsewhere about the existence of traveling wave profiles. If a wave has such a profile, we call it *admissible*.

We review now some simple facts about solutions to (1.8), (1.7). Recall that we consider only states (v, u, λ) satisfying either (1.10) or (1.11).

Let us fix two states (v_-, u_-, λ_-) and (v_+, u_+, λ_+). If they are connected by a wave having a jump discontinuity with speed c, then the Rankine–Hugoniot conditions must hold:

$$\begin{cases} -c[v] - [u] = 0, \\ -c[u] + [p] = 0. \end{cases} \tag{2.1}$$

From these equations it follows that

$$c^2 = -\frac{p(v_+, \lambda_+) - p(v_-, \lambda_-)}{v_+ - v_-}. \tag{2.2}$$

Remark then that a necessary condition for a jump to take place is that in the plane (v, p) the line joining $(v_-, p(v_-))$ and $(v_+, p(v_+))$ has negative slope.

Traveling waves of (1.7) are solutions of

$$\begin{cases} -cv' - u' = 0, \\ -cu' + p(v, \lambda)' = u'', \\ -c\lambda' = aw(v, \lambda) + b\lambda'', \\ (v, u, \lambda)(\pm\infty) = (v_\pm, u_\pm, \lambda_\pm), \quad (v', u', \lambda')(\pm\infty) = (0, 0, 0), \end{cases} \tag{2.3}$$

where $w(v, \lambda) = (p - p_e)\lambda(\lambda - 1)$ and prime ($'$) denotes $d/d\xi$ with $\xi = (x - ct)/\epsilon$. Here the data $(v_\pm, u_\pm, \lambda_\pm)$ satisfy the Rankine–Hugoniot conditions (2.1). Remark that (2.3) implies

$$\begin{cases} -cv' = c^2(v - v_-) + p - p_-, \\ -c\lambda' = a(p - p_e)\lambda(\lambda - 1) + b\lambda''. \end{cases} \tag{2.4}$$

For simplicity in the rest of this section we refer only to the case of phase boundaries with *nonnegative* wave speeds. Then the speed of a phase boundary joining (v_-, u_-, λ_-) and (v_+, u_+, λ_+) is

$$c = \sqrt{-\frac{p(v_+, \lambda_+) - p(v_-, \lambda_-)}{v_+ - v_-}}\,. \tag{2.5}$$

Phase boundaries with negative speeds can be obtained by interchanging the left and right side of phase boundaries with positive speeds.

(i) *Liquefaction waves.* After a liquefaction wave passes, the state of fluid changes from vapor or vapor/liquid mixture ($v_+, \lambda_+ \neq 0$) to liquid ($v_-, \lambda_- = 0$); see Figure 2(a)–(b). Further requirements on the data are
(a) either

$$\lambda_- = 0,\ \lambda_+ = 1, \qquad v_- < v_+, \qquad p(v_-, \lambda_-) > p(v_+, \lambda_+) \geq p_e \tag{2.6}$$

(b) or

$$\lambda_- = 0,\ 0 < \lambda_+ < 1, \qquad v_- < v_+, \qquad p(v_-, \lambda_-) > p(v_+, \lambda_+) = p_e. \tag{2.7}$$

Under either (2.6) or (2.7), it is shown in [7] and [9] that admissible waves exist if

$$c \geq 2\sqrt{ab|p(v_-, \lambda_-) - p_e|}, \tag{2.8}$$

$$c^2 + p_v\,(v_\pm, \lambda_\pm) < 0, \tag{2.9}$$

and there is no other equilibrium point of (2.3), with v value between v_- and v_+. On the other hand, if the speeds satisfy

$$c < 2\sqrt{ab|p(v_+, \lambda_+) - p_e|}, \tag{2.10}$$

then there are no admissible liquefaction waves.
Condition (2.9) means that a liquefaction wave is subsonic or sonic with respect to both side states. The absolute value in (2.8) and (2.10) is not necessary in this case, but it has been put in order to have a unique condition valid also for the next case. Note that in case (i)(b), the condition (2.10) does not impose any restriction.
We further observe through numerical computation that for case (i)(b), there is always a traveling wave as long as $\lambda_+ \in (0, 1)$, even if (2.8) is not satisfied.

(ii) *Evaporation waves.* The state in front of an evaporation wave is metastable liquid: $(v_+, \lambda_+ = 0)$ with $p_+ < p_e$. In the back of the evaporation shock is vapor: $(v_-, \lambda_- = 1)$ with $v_- > v_+$ and $p_- < p_+ < p_e$; see Figure 2(c)–(d). We require moreover that
(a) either

$$\lambda_- = 1,\ \lambda_+ = 0,\ v_- > v_+,\ p(v_-, \lambda_-) < p(v_+, \lambda_+) \leq p_e \tag{2.11}$$

(b) or

$$\lambda_- = 1,\ 0 < \lambda_+ < 1,\ v_- > v_+,\ p(v_-, \lambda_-) < p(v_+, \lambda_+) = p_e\ .$$

Results about the admissibility of evaporation waves are given in [7], [9]: if conditions (2.8), (2.9) hold and there is no other equilibrium point of (2.3) with v value between v_- and v_+, then the evaporation wave is admissible. Similarly, under the condition (2.10), there is no admissible evaporation wave. Our numerical tests show that there is a traveling wave for case (ii)(b) as long as $\lambda_+ \in (0, 1)$, even if (2.8) is not satisfied.

(iii) *Isobaric waves.* Smooth isobaric waves are solutions of (1.8) with

$$p\,(v(x,t), \lambda(x,t)) = p_e \tag{2.12}$$

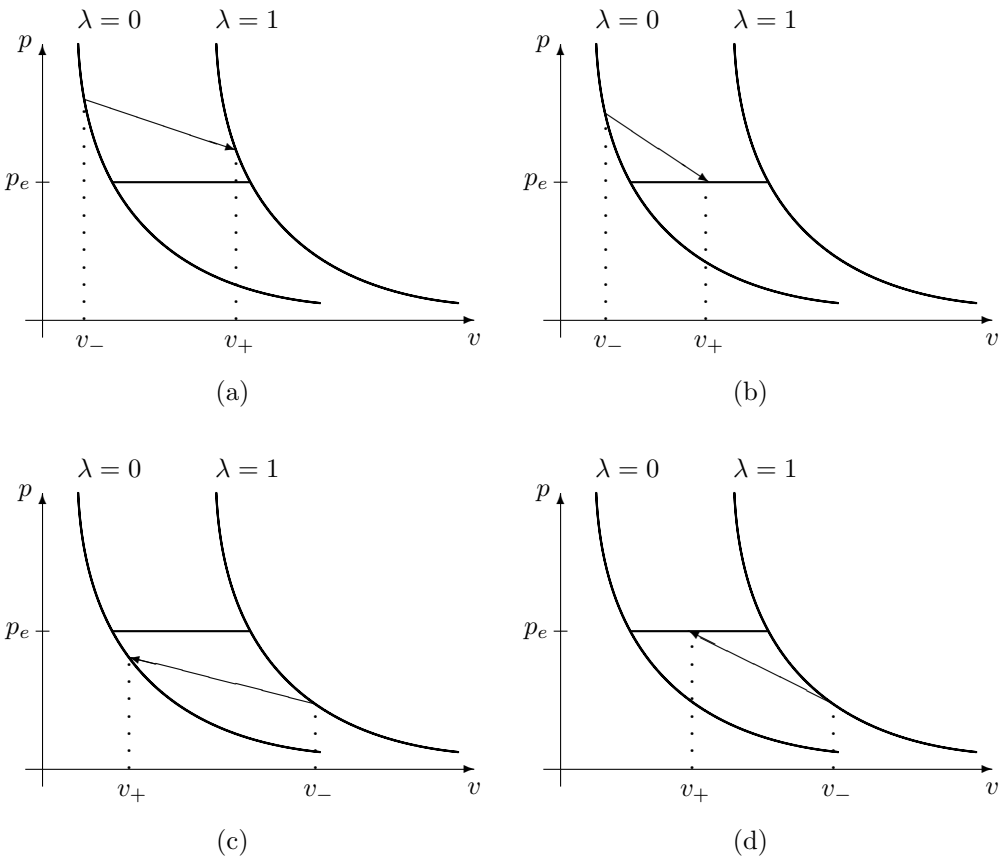


FIG. 2. *Admissible waves.* (a) *and* (b), *liquefaction shocks;* (c) *and* (d), *evaporation shocks.*

for (x,t) in some region D called isobaric region. For convenience, suppose D contains a rectangle $[a,b]\times[t_0,t_1]$. Then in this rectangle we have

$$\begin{cases} u = u(x,t_0), \\ v = u_x(x,t_0)(t-t_0) + v(x,t_0), \\ \lambda = \lambda_e(v(x,t)), \end{cases} \tag{2.13}$$

where the last equation is the explicit form of (2.12), which exists due to $p_\lambda > 0$. Two constant states (v_-,u_-,λ_-), (v_+,u_+,λ_+) are connected by an isobaric wave only if $p(v_-,\lambda_-) = p(v_+,\lambda_+) = p_e$; as a consequence the wave is stationary and $u_+ = u_-$.

(iv) *Nonreacting compressive shock waves.* Under the assumption (1.2) the nonreacting shocks must satisfy $\lambda_- = \lambda_+ = 0$ or 1 and the Lax admissibility conditions

(a) either (1-shocks)

$$-\sqrt{-p_v(v_-,\lambda_-)} > c > -\sqrt{-p_v(v_+,\lambda_+)}$$

(b) or (2 shocks)

$$\sqrt{-p_v(v_-,\lambda_-)} > c > \sqrt{-p_v(v_+,\lambda_+)},$$

where c is the speed of the wave determined by the Rankine–Hugoniot conditions. Due to the assumption $p_{vv} > 0$, we have $\text{sign}\partial_{v_\pm} c = -\text{sign}c$ for both Lax shocks. We shall use these inequalities when solving Riemann problems.

(v) *Nonreacting rarefaction waves.* These are continuous solutions of (1.8) of the form $(u, v, \lambda)(x, t) = (u, v, \lambda)(x/t)$ with $\lambda =$ constant. By the third equation in (1.8), we have $\lambda(x, t) \equiv \lambda_\pm = 0$ or 1, and the system (1.8) is reduced to the nonreacting p-system for gas dynamics. Thus, the nonreacting rarefaction waves are the same as the classical rarefaction waves in gas dynamics:

(a) first family of rarefaction waves,

$$u_+ - u_- = \int_{v_-}^{v_+} \sqrt{-p_v(v,\lambda_-)}dv, \quad v_- < v_+; \tag{2.14}$$

(b) second family of rarefaction waves,

$$u_+ - u_- = -\int_{v_-}^{v_+} \sqrt{-p_v(v,\lambda_-)}dv, \quad v_- > v_+. \tag{2.15}$$

Remark 2.1. We consider again the case of a right-moving liquefaction wave; see (i) above. For a *fixed* v_- we denote the speed of the liquefaction wave connecting v_- with v_+ by $c(v_+)$. It is easy to check that the function $c(v_+)$ is increasing when (2.9) holds. We define v^* by the relation $c(v^*) = 2\sqrt{ab(p(v_-, 0) - p_e)}$. Moreover, consider the line joining $(v_-, p(v_-, 0))$ with $(v_+, p(v_+, 1))$ and let $(v^T, p(v^T, 1))$ be the point in which this line is tangent to the curve $p(v, 1)$. As a consequence of the conditions (2.8) and (2.9), we see that the existence of liquefaction waves connecting v_- with v_+ is ensured if $v_+ \in [v^*, \min\{v^T, v_M\}]$. In the case that the pressure has the form (1.6) then a simple calculation shows that the interval $[v^*, \min\{v^T, v_M\}]$ is never empty for any $v_- \le v_m$ if ab is sufficiently small. In other words, for these values of ab for every v_- we have, in the state space, a whole curve of admissible liquefaction waves.

For the evaporation shocks, the situation is somewhat different: For any fixed $ab > 0$ and as $v_- \to \infty$, the $c(v^*) = 2\sqrt{ab(p(v_-, 1) - p_e)}$ will become bigger than the slope of the tangent of $p(v, 1)$, destroying the possibility of a direct traveling wave link from (v_-, λ_-) to (v_+, λ_+). Thus, for any fixed $ab > 0$, there is an upper bound for v_- so that there is no evaporation wave with v_- above this bound.

The class of admissible waves could be enlarged by admitting the collapsing and explosion waves introduced below. One side of these waves is an equilibrium mixture of liquid and vapor, while the other side is metastable liquid or vapor. Such waves are numerically verified to exist by computing some initial value problems of (1.7) for t not small. The numerical results revealed the following: For these kinds of waves to appear, the pressure of the metastable side must differ from equilibrium by a certain amount. Such waves cannot appear in a Riemann solver when the Riemann initial data consist of a piece of liquid and a piece of vapor unless the initial data is set such that the pressure of the metastable side differs sufficiently from equilibrium. The reason for this phenomenon is left for future investigation. It is clear that if these waves appear in Riemann solvers for such Riemann data, a severe loss of uniqueness will occur. For these reasons we consider the following waves separately.

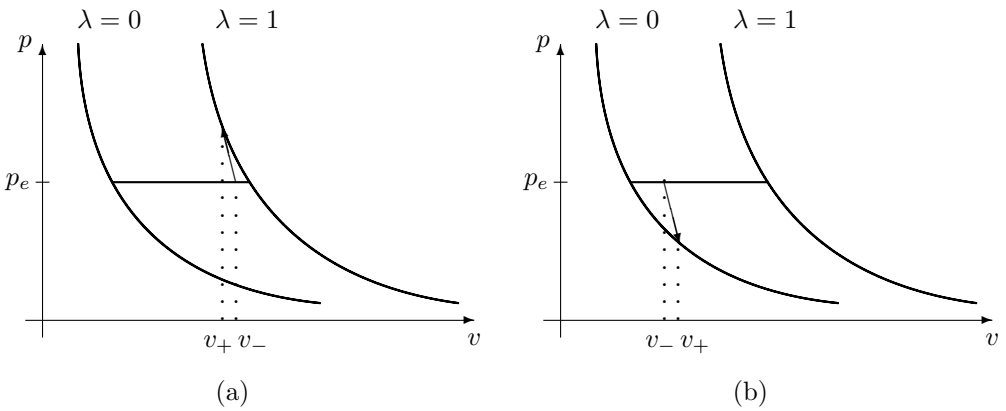


FIG. 3. *Other waves.* (a) *collapsing shock;* (b) *explosion shock.*

(vi) *Collapsing waves.* As a collapsing wave passes through the metastable vapor, $\lambda_+ = 1$, $p_+ > p_e$, the fluid changes to a liquid/vapor mixture at equilibrium pressure, $p_- = p_e$, $\lambda_- \in (0,1)$; see Figure 3(a). Although this is also a kind of liquefaction shock, we single it out to emphasize that it is supersonic relative to its front and that the pressure drops after the shock passes.

(vii) *Explosion waves.* In the front of an explosion wave there is metastable liquid, $\lambda_+ = 0$, $p_+ < p_e$. Behind the wave is a liquid/vapor mixture at equilibrium pressure, $p_- = p_e$, $\lambda_- \in (0,1)$. The Rankine–Hugoniot condition then requires $v_+ > v_-$. See Figure 3(b). Also these waves are supersonic with respect to the state in front.

3. Properties of admissible waves. In this section we study some properties of the admissible waves listed in section 2; these properties provide motivations for the kinetic conditions that are given in section 4.

Throughout this section we denote (v_-, λ_-) the state behind a shock, (v_+, λ_+) the state in front, while $c = c(v_-, v_+)$ is given by (2.5). For simplicity we deal only with the case $c > 0$.

LEMMA 3.1. *If* $c^2 + p_v(v_\pm, \lambda_\pm) < 0$, *with* $\lambda_- = 0$, $\lambda_+ = 1$, *then the equation*

$$R(v, \lambda; c) := c^2(v - v_-) + p(v, \lambda) - p(v_-, 0) = 0 \tag{3.1}$$

has a unique solution $v := v(\lambda)$. *This solution satisfies*

$$R_v\left(v(\lambda), \lambda; c\right) < 0. \tag{3.2}$$

Proof. Existence of a solution $v(\lambda)$ of $R(v, \lambda; c) = 0$ follows from $R_\lambda = p_\lambda > 0$ and then, for $\lambda \in (0,1)$,

$$R(v_-, \lambda; c) > R(v_-, 0; c) = 0 = R(v_+, 1; c) > R(v_+, \lambda; c).$$

To prove the uniqueness of the solution $v(\lambda)$, assume its contrary, i.e., that there are two solutions of the equation $R(v, \lambda_0; c) = 0$ for some $\lambda_0 \in (0,1)$, $v_1(\lambda_0)$, and $v_2(\lambda_0)$. Without loss of generality, let $v_1(\lambda_0) < v_2(\lambda_0)$ denote the smallest two such solutions.

Thus, there are no other zeros between $v_1(\lambda_0)$ and $v_2(\lambda_0)$. Note that $R_{vv} = p_{vv} > 0$ and $R_v(v_-, 0; c) = c^2 + p_v(v_-, 0) < 0$. Then

$$R_v(v_1, \lambda_0; c) < 0 < R_v(v_2, \lambda_0; c) \tag{3.3}$$

and hence

$$R_v(v, \lambda_0; c) > 0 \quad \text{for } v \in (v_2, v_+) \tag{3.4}$$

holds. It leads to a contradiction:

$$0 = R(v_+, 1; c) > R(v_+, \lambda_0; c) > R(v_2, \lambda_0; c) = 0, \tag{3.5}$$

which establishes the uniqueness of $v(\lambda)$.

If there was a point $\lambda_0 \in (0, 1)$ such that $R_v(v(\lambda_0), \lambda_0; c) \geq 0$ holds, then the argument from (3.3) to (3.5) would apply to yield a contradiction. This shows that $R_v(v(\lambda), \lambda; c) < 0$ for $\lambda \in [0, 1]$, as desired. □

LEMMA 3.2. *Under the assumptions of the previous lemma, every liquefaction (evaporation) traveling wave connecting (v_-, λ_-) with (v_+, λ_+) is monotone.*

Proof. We deal with the case of a liquefaction wave. In the (λ, v)-plane, consider the strip $S = \{(\lambda, v) : \ 0 \leq \lambda \leq 1, v > 0\}$. Denote $\mathcal{R}_0 = \{(\lambda, v);\ R(v, \lambda) = 0\}$ and

$$\mathcal{R}_- = \{(\lambda, v) \in S;\ R(v, \lambda) < 0\},$$
$$\mathcal{R}_+ = \{(\lambda, v) \in S;\ R(v, \lambda) > 0\}.$$

For contradiction, assume that there exists a profile $(v(\xi), \lambda(\xi))$ with $v'(\xi_0) = 0$ for some ξ_0. When $(v(\xi), \lambda(\xi))$ starts from $(v_-, 0)$ it must enter the region $\mathcal{R}_-$; otherwise $v'(\xi) < 0$ for every ξ and then it could not reach the point $(v_+, 1)$. From (2.4), when $\xi = \xi_0$ the trajectory $(v(\xi), \lambda(\xi))$ meets the curve $\mathcal{R}_0$. The curve $\mathcal{R}_0$, parameterized as $v(\lambda)$, is increasing as λ increases, as shown by

$$\frac{dv}{d\lambda} = -\frac{R_\lambda}{R_v} = -\frac{p_\lambda}{R_v} > 0$$

in view of Lemma 3.1. Thus, $v'(\xi) < 0$ for every $\xi > \xi_0$ and so the point $(v_+, 1)$ cannot be reached, a contradiction.

If λ is not monotone, then there exists the smallest critical point ξ_0, which must be a local maximum, and then a critical point ξ_1, adjacent to ξ_0, which must be a local minimum point. From the second equation of (2.4) we have $p(v(\xi_0), \lambda(\xi_0)) \leq p_e$. If $\xi \in (\xi_0, \xi_1)$, then $\lambda(\xi)$ is decreasing while $v(\xi)$ is increasing; then $p(v(\xi_1), \lambda(\xi_1)) < p_e$ from the first equation of (2.4). But then at ξ_1 we should have $0 = a(p - p_e)\lambda(\lambda - 1) + b\lambda'' > 0$, a contradiction. □

Remark that in the plane (λ, v) the curve $\{(\lambda, v);\ p(v, \lambda) = p_e\}$ lies all above the curve $\mathcal{R}_0$; see Figure 4. In fact, if $v(\lambda)$ is as in Lemma 3.1, then $v'(\lambda) > 0$. By differentiating (3.1) with respect to λ we obtain $c^2 v'(\lambda) + d/d\lambda[p\,(v(\lambda), \lambda)] = 0$, and thus p is decreasing along the curve $\mathcal{R}_0$. Since $v_- < v_m$ and $v_+ < v_M$, the claim is proved.

LEMMA 3.3. *Fix a state (v_-, λ_-), $\lambda_- = 0$ (or 1). If there are liquefaction (evaporation) waves having (v_-, λ_-) as the back state, then there is one with the least speed.*

Proof. We now consider the case for liquefaction waves. The proof for evaporation waves is the same.

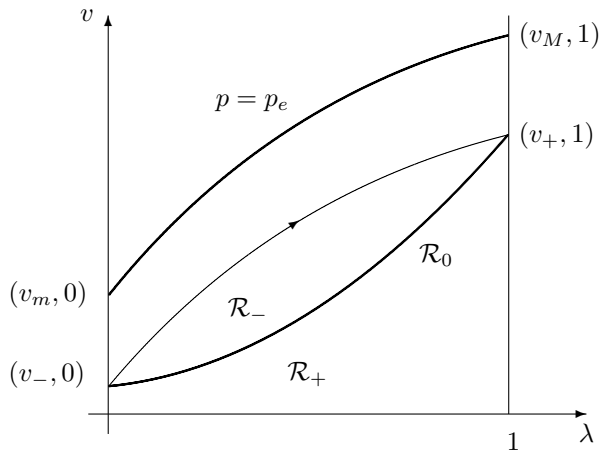


FIG. 4. *Profile in the plane* (λ, v). *The curves* $\mathcal{R}_0$ *and* $p = p_e$ *are represented by thickness, the profile with a thin line.*

Fix a state $(v_-, 0)$ and consider the set $\mathcal{P}(v_-)$ of all states $(v_+, 1)$ that can be connected to $(v_-, 0)$ with a liquefaction wave. These traveling waves satisfy (2.4), and from Lemma 3.2 they are monotone. Let

$$\omega = \inf_{v_+ \in \mathcal{P}(v_-)} c(v_-, v_+) \, .$$

Take a sequence $v_n \in \mathcal{P}(v_-)$ such that $c(v_-, v_n) \to \omega$ as $n \to +\infty$, and let

$$(v(\xi; v_-, v_n), \lambda(\xi; v_-, v_n))$$

be the related traveling waves. Note that a shift of a traveling wave solution of (2.3) is still a traveling wave with the same speed. By shifting we can assume that for every n

$$\lambda(0; v_-, v_n) = 1/2 \, .$$

Due to (1.3), $v_n \le v_1$ for large n to make possible $c(v_-, v_n) \to \omega$, as $n \to \infty$. We have that for large n, $\mathrm{TV}(v(\cdot; v_-, v_n)) \le |v_+ - v_1|$, $\mathrm{TV}(\lambda(\cdot; v_-, v_n)) = 1$ since they are monotone. Then there exists a subsequence converging a.e. to some functions $(v_*(\xi), \lambda_*(\xi))$.

The functions $v_*(\xi)$, $\lambda_*(\xi)$ still satisfy (2.4) and are monotone. They are not constant because $\lambda_*(0) = 1/2$ is not an equilibrium point of (2.4). Then they satisfy the conditions $v_*(\pm\infty) = v_\pm$, $\lambda_*(-\infty) = 0$, $\lambda_*(+\infty) = 1$. Thus, $(v_*(\xi), \lambda_*(\xi))$ is a traveling wave with the minimum speed. □

We denote the lowest speed of a liquefaction (evaporation) wave by $\omega(v_-, \lambda_-)$.

THEOREM 3.1. *The function* $v_- \to \omega(v_-, \lambda_-)$ *introduced above is decreasing in the case of a liquefaction wave, increasing in the case of an evaporation wave.*

Proof. For definiteness, we prove the statement for a liquefaction wave. For an evaporation wave the proof is analogous.

In the case of a liquefaction wave, the state λ behind the wave is $\lambda_- = 0$. Consider v_- in the domain of definition of the function $\omega(v_-, 0)$. By the definition of $\omega(v_-, 0)$, there is a solution of (2.4) starting at $v(-\infty) = v_-$, and the slowest nonnegative speed of such traveling waves is $\omega(v_-, 0)$, which connects $(v_\pm, \lambda_\pm)$ as depicted in Figure 2(a). We denote this slowest traveling wave by $(v_1, \lambda_1)(\xi; v_-)$.

We want to prove that

$$\omega := \omega(v_-, 0) \geq \omega(v_-^*, 0) \tag{3.6}$$

if $v_- < v_-^* < v_m$ with v_-^* sufficiently close to v_-. It suffices to prove that for such v_-^*, there is a solution of (2.4) with speed $c = \omega$ starting from $v(-\infty) = v_-^*$. The other side of the traveling wave $v(\infty) = v_+^*$ must satisfy the Rankine–Hugoniot condition

$$\omega^2(v_+^* - v_-^*) + p(v_+^*, 1) - p(v_-^*, 0) = 0. \tag{3.7}$$

Due to the subsonicity at $(v_\pm, p(v_\pm, \lambda_\pm))$, the smallness of $|v_- - v_-^*|$, and the shape of $p(v, \lambda)$ given by (1.2) and (1.3), there must be the smallest solution of (3.7) satisfying $v_+^* > v_+$. We rewrite and modify the system (2.4) of traveling waves connecting $v_\pm^*$ as

$$\begin{cases} -\omega v' = \omega^2(v - v_-^*) + p - p_-^* + \eta v'', \\ -\omega \lambda' = aw(v, \lambda) + b\lambda'', \\ (v, \lambda)(\pm\infty) = (v_\pm^*, \lambda_\pm), \quad (v', \lambda')(\pm\infty) = (0, 0). \end{cases} \tag{3.8}$$

We shall take the limit $\eta \to 0+$ later. By Theorems 3.1 and 3.1′ in [9] and the analysis therein, a necessary and sufficient condition for (3.8) with data given in (i) and (ii) of section 2 to have a solution is

$$\omega \geq \inf_{U \in K} \sup_{x,j} \frac{a_j U_j'' + F_j(U, \omega)}{-U_j'}, \tag{3.9}$$

where

$$\begin{aligned} U &= (v, \lambda), \\ F(U, \omega) &= \left(\omega^2(v - v_-^*) + p - p_-^* \;\; a(p - p_e)\lambda(\lambda - 1)\right), \end{aligned}$$

with $a_1 = \eta$, $a_2 = b$, and

$$K := \{U \in C^2 \;:\; U \text{ is monotone and } U(\pm\infty) = (v_\pm^*, \lambda_\pm)\}. \tag{3.10}$$

We now apply Lemma 3.1; then denote by $v_2 := v(\lambda)$ the unique solution of the equation

$$R(v, \lambda; \omega) := \omega^2(v - v_-^*) + p(v, \lambda) - p(v_-^*, 0) = 0. \tag{3.11}$$

We consider the function

$$v_2^*(\xi) := v_2(\lambda_1(\xi)).$$

It is clear that $v_2^*(\pm\infty) = v_\pm^*$. Let $\delta > 0$ be a sufficiently small constant. The intervals $(-\infty, \xi_-)$ over which $v_2^*(\xi) > v_1(\xi) + \delta$ are nonempty since $v_2^*(-\infty) = v_-^* > v_- = v_1(-\infty)$. Let $(-\infty, \xi_-)$ denote the largest such interval, and hence either $\xi_- = \infty$ or $v_1(\xi_-) + \delta = v_2^*(\xi_-)$ holds. If it is the later case, then there is an interval (ξ_+, ∞) over which $v_1(\xi) + \delta < v_2^*(\xi)$ and $v_1(\xi_+) + \delta = v_2^*(\xi_+)$ hold. Due to (3.2), the function $v_2^*(\xi)$ is strictly increasing since $\lambda_1' > 0$. We can construct a function $v_3^* \in C^2$ satisfying

$$v_3^*(\xi) = \begin{cases} v_2^*(\xi) & \text{if } \xi < \xi_- - 1, \\ \text{joining the left and right branches in } C^2 & \text{if } \xi_- - 1 < \xi < \xi_-, \\ v_1(\xi) + \delta & \text{if } \xi_- < \xi < \xi_+, \\ \text{joining the left and right branches in } C^2 & \text{if } \xi_+ < \xi < \xi_+ + 1, \\ v_2^*(\xi) & \text{if } \xi > \xi_+ + 1. \end{cases} \tag{3.12}$$

The joining pieces in (3.12) can be made to satisfy $v_3^{*'} > 0$ and $v_1 + \delta < v_3^* < v_2^*$ for $\xi \in (\xi_- - 1, \xi_-) \cup (\xi_+, \xi_+ + 1)$.

It is clear that $(v_3^*, \lambda_1) \in K$ as defined in (3.10). Also, the function v_3^* is independent of η. Since $(v_1, \lambda_1)(\xi, v_-)$ is a traveling wave of speed ω joining $(v_\pm, \lambda_\pm)$, it follows that

$$\begin{aligned} \omega &= \frac{-b\lambda_1'' - a(p(v_1, \lambda_1) - p_e)\lambda_1(\lambda_1 - 1)}{\lambda_1'} \\ &> \frac{-b\lambda_1'' - a(p(v_1 + \delta, \lambda_1) - p_e)\lambda_1(\lambda_1 - 1)}{\lambda_1'} \\ &\geq \frac{-b\lambda_1'' - a(p(v_3^*, \lambda_1) - p_e)\lambda_1(\lambda_1 - 1)}{\lambda_1'} \end{aligned} \tag{3.13}$$

since $v_3^* \geq v_1 + \delta$. To estimate the other component of (3.9), we consider

$$\frac{\eta v_3^{*''} + \omega^2(v_3^* - v_-^*) + p(v_3^*, \lambda_1) - p_-^*}{-v_3^{*'}}. \tag{3.14}$$

Over the range $(-\infty, \xi_-) \cup (\xi_+, \infty)$, we use $v_3^* \leq v_2^*$ and (3.11) to derive

$$\begin{aligned} \omega^2(v_3^* - v_-^*) + p(v_3^*, \lambda_1) - p_-^* &= \omega^2(v_3^* - v_2^*) + p(v_3^*, \lambda_1) - p(v_2^*, \lambda_1) \\ &= R_v(\theta, \lambda_1; \omega)(v_3^* - v_2^*) \geq 0. \end{aligned}$$

Here, we used the smallness of $|v_2^* - v_3^*|$ and Lemma 3.1. Then the inequality $v_3^{*\prime} > 0$ yields

$$\frac{\eta v_3^{*''} + \omega^2(v_3^* - v_-^*) + p(v_3^*, \lambda_1) - p_-^*}{-v_3^{*'}} \leq -\eta \frac{v_3^{*''}}{v_3^{*'}}. \tag{3.15}$$

On intervals $(-\infty, \xi_- - 1)$ and $(\xi_+ + 1, \infty)$, the definition of v_3^* implies that

$$\frac{v_3^{*''}}{v_3^{*'}} = O(1) + O(1)\frac{\lambda_1''}{\lambda_1'}.$$

For the traveling wave $(v_1, \lambda_1)(\xi)$, we can prove that

$$\frac{\lambda_1''}{\lambda_1'} = O(1)$$

by investigating the decay rate of $\lambda_1(\xi)$ at $\xi = \pm\infty$ and by $v_1' > 0$ proved in the proof of Lemma 3.2. Over the intervals $[\xi_- - 1, \xi_-)$ and $(\xi_+, \xi_+ + 1]$, $v_3^{*''}/v_3^{*'}$ changes smoothly by $O(1)$ from $v_2^{*''}(\xi_\pm \pm 1)/v_2^{*'}(\xi_\pm \pm 1) = O(1)$ to $v_1^{*''}(\xi_\pm)/v_1^{*'}(\xi_\pm) = O(1)$. Hence $v_3^{*''}/v_3^{*'}$ is also bounded on the intervals $[\xi_- - 1, \xi_-)$ and $(\xi_+, \xi_+ + 1]$. Therefore we have

$$\frac{\eta v_3^{*''} + \omega^2(v_3^* - v_-^*) + p(v_3^*, \lambda_1) - p_-^*}{-v_3^{*'}} = O(1)\eta \tag{3.16}$$

for $\xi \in (-\infty, \xi_-) \cup (\xi_+, \infty)$.

On the interval $[\xi_-, \xi_+]$, we have $v_3^* = v_1 + \delta$ and

$$\begin{aligned}
&\frac{\eta v_3^{*\prime\prime} + \omega^2(v_3^* - v_-^*) + p(v_3^*, \lambda_1) - p_-^*}{-v_3^{*\prime}} \\
&= \frac{\eta v_1'' + \omega^2(v_1 + \delta - v_-^*) + p(v_1 + \delta, \lambda_1) - p_-^*}{-v_1'} \\
&= \frac{\omega^2(v_1 - v_-) + p(v_1, \lambda_1) - p_-}{-v_1'} \\
&\quad + \frac{\eta v_1'' + \omega^2\delta + p(v_1 + \delta, \lambda_1) - p(v_1, \lambda_1) + \omega^2(v_- - v_-^*) + p_- - p_-^*}{-v_1'} \\
&= \omega + \frac{\eta v_1'' + \omega^2\delta + p(v_1 + \delta, \lambda_1) - p(v_1, \lambda_1) + \omega^2(v_- - v_-^*) + p_- - p_-^*}{-v_1'}.
\end{aligned} \tag{3.17}$$

The part

$$F(v_1, \lambda_1, \delta) := \eta v_1'' + \omega^2\delta + p(v_1 + \delta, \lambda_1) - p(v_1, \lambda_1) + \omega^2(v_- - v_-^*) + p_- - p_-^*$$

in (3.17) satisfies

$$F(v_1, \lambda_1, 0) = \eta v_1'' + \omega^2(v_- - v_-^*) + p_- - p_-^* > 0$$

when $\eta > 0$ is sufficiently small, due to $\omega^2 + p_v(v_-, \lambda_-) < 0$ and the smallness of $v_-^* - v_- > 0$. Then, by the continuity of F, $F(v_1, \lambda_1, \delta) > 0$ for any (v_1, λ_1) in the bounded closed set $[v_-, v_+] \times [0, 1]$ if $\delta > 0$ is sufficiently small. This, together with (3.17) and $v_1' > 0$, implies that, for $\xi \in [\xi_-, \xi_+]$,

$$\frac{\eta v_3^{*\prime\prime} + \omega^2(v_3^* - v_-^*) + p(v_3^*, \lambda_1) - p_-^*}{-v_3^{*\prime}} < \omega. \tag{3.18}$$

Combining estimates (3.13)–(3.18), we see that the condition (3.9) is satisfied when $\eta > 0$ is sufficiently small. Thus, the system (3.8) has a solution with speed ω when $\eta > 0$ is sufficiently small.

Let $(v, \lambda)(\xi, v_\pm^*; \eta)$ be solutions of (3.8) with the speed ω. They are monotone in ξ. Since (3.8) is invariant under shifting, $\xi \mapsto \xi + \xi_0$, we can assume $v(0, v_\pm^*; \eta) = (v_+^* + v_-^*)/2$. By the monotoneness of these solutions, there is a sequence $\{\eta_n\}_1^\infty$ such that

$$(v, \lambda)(\xi, v_\pm^*; 0) := \lim_{n\to\infty} (v, \lambda)(\xi, v_\pm^*; \eta_n)$$

exist for a.e. $\xi \in (-\infty, \infty)$. The limit $(v, \lambda)(\xi, v_\pm^*; 0)$ satisfies the differential equations in (2.4) with speed ω in the sense of distribution, and hence in the strong sense. It also satisfies the boundary conditions in (2.4) since there is no other equilibrium point between $(v, \lambda) = (v_-^*, 0)$ and $(v_+^*, 1)$. In other words, the system (2.4) has a traveling wave with speed $\omega = \omega(v_-, 0)$, starting at $v(-\infty) = v_-^*$. Therefore, the minimum speed $\omega(v_-^*, 0)$ of all traveling waves starting at v_-^* is no greater than $\omega(v_-, 0)$. The proof is complete. □

Remark 3.1. The above proof can be further improved to yield strict monotonicity of $\omega(v_-, \lambda_-)$.

4. Kinetic relations. At this stage, the existence results on traveling wave profiles listed in section 2 are not strong enough to guarantee the solvability of the Riemann problem for (1.8). On the other hand, for each given v_-, there are too many liquefaction and vaporization traveling waves listed in section 2, cases (i) and (ii). If we admit all these waves, then many solutions can exist for the same initial data. Admitting collapsing and explosion waves introduce more solution configurations for Riemann solvers; see section 6. When to use which waves remains to be investigated more thoroughly in future studies. For this reason, instead of strictly using traveling wave admissibility criterion, we shall impose the following selection criteria for admissible waves that mimics the properties of traveling wave profiles. The selection criteria are also called kinetic relations in most mathematical literature.

In the solution of the Riemann problem for (1.8) we admit every liquefaction or evaporation wave connecting a pure phase with a mixture phase, isobaric waves, and Lax-shock and rarefaction waves, as defined in section 2, disregarding the existence results of their traveling waves profiles. Of course each of these waves must satisfy the Rankine–Hugoniot conditions (2.1) when a discontinuity occurs.

Moreover we admit both liquefaction and evaporation waves connecting *two pure phases*. The selection criterion for these liquefaction and evaporation waves are as follows: Let (v_-, u_-, λ_-) be the state in the back of a shock, (v_+, u_+, λ_+) the state in the front. For simplicity we consider only the case of waves with positive speeds.

(i) For each liquid state $(v_-, u_-, 0)$ with $p(v_-, 0) > p_e$ there is only one vapor state $(v_+, u_+, 1)$, satisfying (2.6), that can be connected to $(v_-, u_-, 0)$ with a liquefaction wave. The speed $s = s(v_-)$ of the liquefaction wave is a decreasing C^1 function satisfying $s^2(v_-) + p_v(v_\pm, \lambda_\pm) \le 0$.

(ii) For each vapor state $(v_-, u_-, 1)$ with $p(v_-, 1) < p_e$ there is only one liquid state (v_+, u_+0), satisfying (2.11), that can be connected to $(v_-, u_-, 1)$ with an evaporation wave. The speed $s = s(v_-)$ of the evaporation wave is an increasing C^1 function; it satisfies $s^2(v_-) + p_v(v_\pm, \lambda_\pm) \le 0$.

There are physical reasons for imposing conditions (i) and (ii): for example, let us consider the case of liquefaction waves. The speeds of liquefaction waves depend on the density of liquid drops in front of the waves, which is determined by the nucleation effects [8]. The larger the density of liquid drops in front of the waves is, the faster the liquefaction wave moves. The nucleation term is typically very small and takes a long time to have some effect in a metastable vapor. Thus, when t is not large, the liquefaction shock travels at the lowest possible speed. The first part of condition (i) then can be justified furthermore because the nucleation effect is not included in (1.8). The approximation of the hyperbolic regime to the parabolic one is then to be understood for $O(1)$-order time intervals, in which the nucleation changes little the value of λ in the front of the liquefaction and evaporation traveling waves. Furthermore, if every liquefaction shock is allowed, then the Riemann solvers are not unique.

The second part of condition (i) is quite intuitive, since if v_- decreases then the difference $p(v_-, 0) - p_e > 0$ increases, speeding up the phase changes. Hence it is natural to assume that the liquefaction waves proceed faster. Moreover, in terms of traveling waves, we showed in Theorem 3.1 that the liquefaction shock moving with the slowest speed satisfies this assumption. At last, the condition $s^2(v_-) + p_v(v_\pm, \lambda_\pm) \le 0$ means simply that the wave can be either subsonic or sonic.

Condition (ii) is motivated on the same basis. But, there is a difference from the case of a liquefaction wave: The assumption $s^2(v_-) + p_v(v_-, 1) \le 0$ implies that the

connection exists only if $v_- \le v_-^T$, where v_-^T is the abscissa of the point where the line joining $(v_-, p(v_-, 1))$ with $(v_+, p(v_+, 0))$ is tangent to the curve $p(v, 1)$. Such a point v_-^T exists and is unique because of the assumption $s'(v_-) > 0$.

5. Solutions of the Riemann problem. In this section we show how to solve the Riemann problem (1.8) for any data (v_-, u_-, λ_-), (v_+, u_+, λ_+) with $v_\pm > 0$, $u_\pm \in (-\infty, +\infty)$, $\lambda_\pm \in [0, 1]$. This is done by considering several different sets of initial data; cases are classified according to the difference in speeds $u_- - u_+$. The waves considered in the Riemann problems below are always listed from the left to the right.

We define $v_+^T > v_M$, $v_-^T > v_M$ as the points that satisfy, respectively, the equations

$$p_v(v_+^T, 1) = \frac{p(v_+, \lambda_+) - p(v_+^T, 1)}{v_+ - v_+^T}\,, \tag{5.1}$$

$$p_v(v_-^T, 1) = \frac{p(v_-^T, 1) - p(v_-, \lambda_-)}{v_-^T - v_-}\,. \tag{5.2}$$

The point v_+^T (respectively, v_-^T) is the abscissa of the tangency point of the line passing through $(v_+, p(v_+, \lambda_+))$ (respectively, $(v_-, p(v_-, \lambda_-))$) with the curve $p(v, 1)$.

5.1. Data in two mixture phases. We assume about data $\lambda_- \in (0, 1)$, $\lambda_+ \in (0, 1)$, $p(v_-, \lambda_-) = p(v_+, \lambda_+) = p_e$.

5.1.1. $u_+ - u_- < 0$. In this case the solution has a transition

$$\text{mixture} \to \text{liquid} \to \text{mixture}.$$

See Figure 5(a) in the case $v_- < v_+$. We claim that there exists a unique $v_1 < v_m$ such that the Riemann problem has a solution made of a left-moving liquefaction shock (from v_- to v_1) followed by a right-moving liquefaction shock (from v_1 to v_+).

In fact the jump conditions require

$$\begin{cases} -s_1(v_1 - v_-) &= u_1 - u_-, \\ -s_2(v_+ - v_1) &= u_+ - u_1. \end{cases} \tag{5.3}$$

As both s_1 and s_2 are functions of v_1, summing up the previous equations we find $F(v_1) := -s_1(v_1 - v_-) - s_2(v_+ - v_1) = u_+ - u_-$. We remark first that

$$\inf_{v \le v_m} F(v_1) = -\infty, \quad \max_{v \le v_m} F(v_1) = F(v_m) = 0.$$

Moreover,

$$\frac{dF}{dv_1}(v_1) = -s_1 + s_2 - \frac{ds_1}{dv_1}(v_1 - v_-) - \frac{ds_2}{dv_1}(v_+ - v_1). \tag{5.4}$$

From (2.2) we see that $ds_1/dv_1 > 0$, $ds_2/dv_1 < 0$; then $dF/dv_1 > 0$. Therefore for every $u_+ - u_- < 0$ there exists a unique v_1 such that $F(v_1) = u_+ - u_-$.

5.1.2. $u_+ - u_- = 0$. In this case the solution is simply given by an isobaric wave from v_- to v_+. Velocity and pressure are constant through the wave.

Following the pattern of case 5.1.1, for $v_1 \to v_m$ one could argue that the solution consists of two jumps: v_- to v_m, v_m to v_+. However, both waves are stationary, so they coincide, and the jump is directly from v_- to v_+.

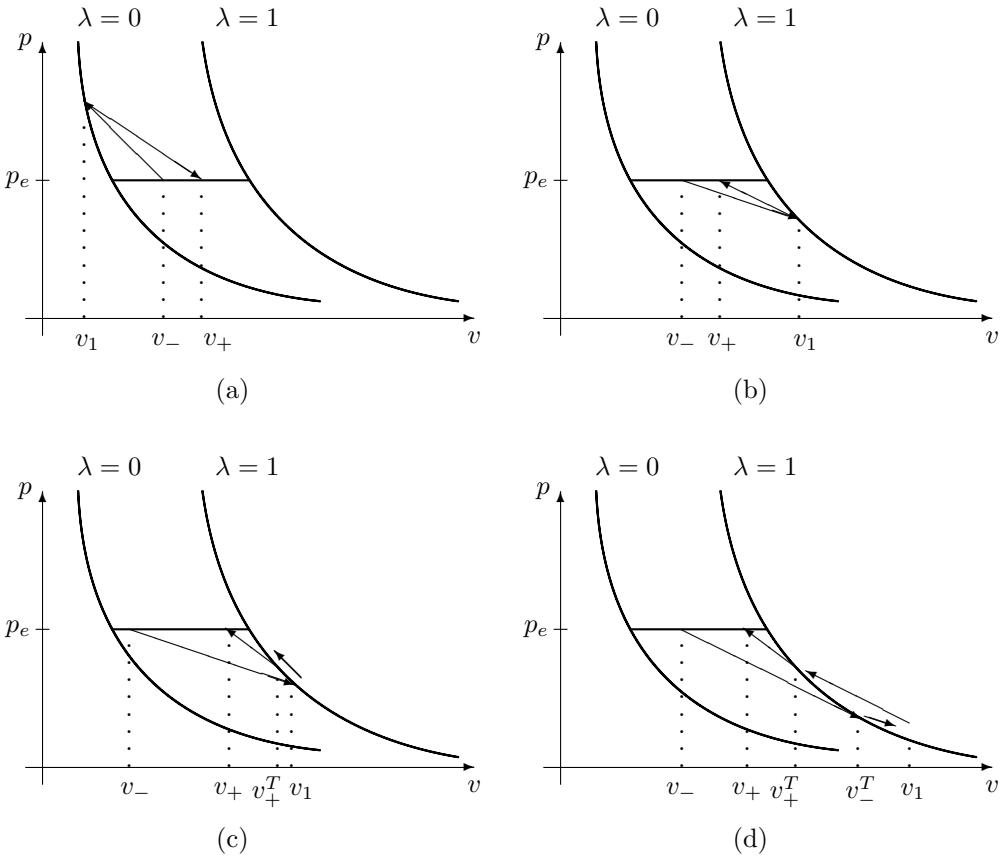


FIG. 5. *Two mixture phases data.* (a) $u_+ - u_- < 0$; (b) $0 < u_+ - u_- < u_*$; (c) $u_* < u_+ - u_- < u^*$; (d) $u_+ - u_- > u^*$. *In case* (c) *the small wave is a* 2*-rarefaction wave attached to the evaporation shock; in case* (d) *the small waves are first a* 1*-rarefaction then a* 2*-rarefaction wave.

5.1.3. $u_+ - u_- > 0$. In this case the solution has a transition (see Figures 5(b)–(d))

$$\text{mixture} \to \text{vapor} \to \text{mixture}.$$

Define

$$u_* = \sqrt{-\frac{p(v_+^T, 1) - p(v_-, \lambda_-)}{v_+^T - v_-}}(v_+^T - v_-) - \sqrt{-p_v(v_+^T, 1)}(v_+ - v_+^T),$$

$$u^* = \sqrt{-p_v(v_-^T, 1)}(v_-^T - v_-) - \int_{v_-^T}^{v_+^T} \sqrt{-p_v(v, 1)}\, dv - \sqrt{-p_v(v_+^T, 1)}(v_+ - v_+^T).$$

It is easy to check that $0 < u_* < u^*$.
Consider first the case $v_- < v_+$. Three patterns arise.

(i) $0 < u_+ - u_- \le u_*$: The solution consists of a left-moving evaporation shock (from v_- to v_1) followed by a right-moving evaporation shock (from v_1 to v_+); see Figure 5(b). As in case 5.1.1 the equations (5.3) must be satisfied. Now, however, $ds_1/dv_1 < 0$, $ds_2/dv_1 > 0$ and from (5.4) we again have $dF/dv_1 > 0$. As $F(v_M) = 0$, this means that we can solve uniquely the equation $F(v_1) = u_+ - u_-$ until the line joining $(v_1, p(v_1, 1))$ and $(v_+, p(v_+, \lambda_+))$ becomes tangent to the graph of $p(v, 1)$, i.e., for $v_1 \in (v_M, v_+^T]$. This concludes this case, because $F(v_+^T) = u_*$.

(ii) $u_* \le u_+ - u_- \le u^*$: The solution consists of a left-moving evaporation shock (from v_- to v_1), followed by a 2-rarefaction wave (from v_1 to v_+^T) ending on a right-moving evaporation shock (from v_+^T to v_+). We call s_2^T the speed of this last shock; see Figure 5(c). Then

$$\begin{cases} -s_1(v_1 - v_-) = u_1 - u_-, \\ -\displaystyle\int_{v_1}^{v_+^T} \sqrt{-p_v(v,1)}\, dv = u_+^T - u_-, \\ -s_2^T(v_+ - v_+^T) = u_+ - u_+^T. \end{cases}$$

Summing up we define the function

$$G(v_1) = -s_1(v_1 - v_-) - \int_{v_1}^{v_+^T} \sqrt{-p_v(v,1)}\, dv - s_2^T(v_+ - v_+^T).$$

We already proved that $ds_1/dv_1 < 0$, moreover $ds_2^T/dv_1 = 0$; then $dG/dv_1 > 0$. Therefore we can solve uniquely the equation $G(v_1) = u_+ - u_-$ until the line joining $(v_-, p(v_-, \lambda_-))$ and $(v_1, p(v_1, 1))$ becomes tangent to the graph of $p(v, 1)$, that is, for $v_1 \in [v_+^T, v_-^T]$. This concludes this case, since $G(v_-^T) = u^*$ and $G(v_+^T) = F(v_+^T)$.

(iii) $u^* \le u_+ - u_- < +\infty$: The solution consists of a left-moving evaporation shock (from v_- to v_-^T, whose speed we call s_1^T) followed by a 1-rarefaction wave (from v_-^T to v_1), coinciding at the left with the evaporation shock; then it follows a 2-rarefaction wave (from v_1 to v_+^T) ending on a right-moving evaporation shock (from v_+^T to v_+). See Figure 5(d). Then

$$\begin{cases} -s_1^T(v_-^T - v_-) = u_-^T - u_-, \\ \displaystyle\int_{v_-^T}^{v_1} \sqrt{-p_v(v,1)}\, dv = u_-^T - u_1, \\ -\displaystyle\int_{v_1}^{v_+^T} \sqrt{-p_v(v,1)}\, dv = u_+^T - u_-, \\ -s_2^T(v_+ - v_+^T) = u_+ - u_+^T. \end{cases}$$

Summing up we define the function

$$H(v_1) = -s_1^T(v_-^T - v_-) + \int_{v_-^T}^{v_1} \sqrt{-p_v(v,1)}\, dv - \int_{v_1}^{v_+^T} \sqrt{-p_v(v,1)}\, dv - s_2^T(v_+ - v_+^T).$$

We see that $dH(v_1)/dv_1 = 2\sqrt{-p_v(v_1, 1)} > 0$; moreover $H(v_-^T) = G(v_-^T)$ and $\sup_{v \ge v_-^T} H(v) = +\infty$ because of (1.5). Then we can solve uniquely the equation $H(v_1) = u_+ - u_-$ for $u_+ - u_- \in [u^*, +\infty)$.

This concludes the discussion of case 5.1.3 if $v_- < v_+$. The case $v_- > v_+$ is analogous and goes as follows. If $u_+ - u_-$ is positive and sufficiently small, we find a solution consisting of a left-moving evaporation shock followed by a right-moving evaporation shock, as in case (i); for larger values of $u_+ - u_-$ the solution consists of an evaporation shock moving toward the left with fixed speed, and it is sonic on the right. Then it follows a 1-rarefaction wave coinciding on the left with the shock and then a right-moving evaporation shock. For still larger values of $u_+ - u_-$ the pattern is as in case (iii). Details are left to the reader.

5.2. Data in one mixture, one pure phase. In this section we solve the Riemann problem in the case that one data is in a pure phase and the other in a mixture phase. More precisely we treat in detail the case of data (v_-, u_-, λ_-), (v_+, u_+, λ_+) with $\lambda_- = 0, 1$ and $\lambda_+ \in (0, 1)$. The mirror case $\lambda_- \in (0, 1)$, $\lambda_+ = 0, 1$ is easily deduced by the transformation $x \mapsto -x$.

5.2.1. $\boldsymbol{\lambda_- = 0}$, $\boldsymbol{\lambda_+ \in (0,1)}$, $\boldsymbol{-\infty < u_+ - u_- \le u_*}$. The threshold u_* is defined by

$$u_* = \begin{cases} \displaystyle\int_{v_-}^{v_m} \sqrt{-p_v(v,0)}\, dv & \text{if } v_- \le v_m, \\ \displaystyle\sqrt{-\frac{p(v_m,0) - p(v_-,0)}{v_m - v_-}}(v_m - v_-) & \text{if } v_- > v_m. \end{cases}$$

The solution has a transition

$$\text{liquid} \to \text{mixture}.$$

The solution consists of a 1-wave from v_- to v_1, then a right-moving liquefaction shock from v_1 to v_+ (an isobaric wave if $u_+ - u_- = u_*$); see Figure 6(a). This construction can be made if $v_1 \le v_m$, and for $v_1 = v_m$ we have an isobaric wave; if $v_1 > v_m$, the liquefaction shock is no more admissible. Then

$$\begin{cases} -\chi_{(-\infty,v_-)}(v_1) \cdot s_1(v_1 - v_-) + \chi_{[v_-,+\infty)}(v_1) \cdot \displaystyle\int_{v_-}^{v_1} \sqrt{-p_v(v,0)} dv = u_1 - u_-, \\ -s(v_+ - v_1) = u_+ - u_1. \end{cases}$$

The function

$$F(v_1) = -\chi_{(-\infty,v_-)}(v_1) \cdot s_1(v_1 - v_-) + \chi_{[v_-,+\infty)}(v_1) \cdot \int_{v_-}^{v_1} \sqrt{-p_v(v,0)} dv \\ -s(v_+ - v_1)$$

satisfies $dF/dv_1 > 0$ because $ds_1/dv_1 > 0$, $ds/dv_1 < 0$. Since $v_1 \le v_m$ we have a solution if $u_+ - u_- \in (-\infty, F(v_m)]$. It is then sufficient to remark that $F(v_m) = u_*$ to complete the proof.

5.2.2. $\boldsymbol{\lambda_- = 0}$, $\boldsymbol{\lambda_+ \in (0,1)}$, $\boldsymbol{u_* < u_+ - u_- \le u^*}$. The value u^* is defined in (5.11). In this case the solution has a transition

$$\text{liquid} \to \text{vapor} \to \text{mixture}.$$

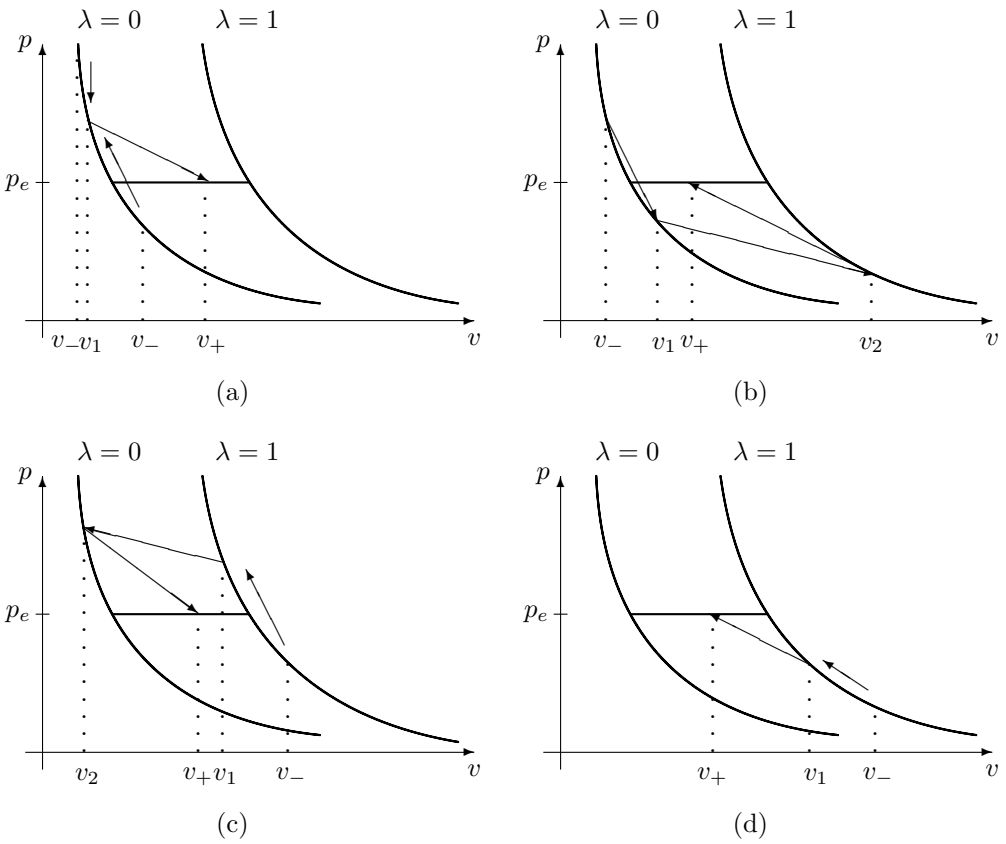


FIG. 6. *One pure phase, one mixture phase data. Liquid-mixture cases:* (a) $-\infty < u_+ - u_- < u_*$; (b) $u_* < u_+ - u_- < u^*$. *In case* (a) *both a* 1*-rarefaction and a* 1*-shock are drawn; if* $u_+ - u_- > u^*$ *the pattern is analogous to that in Figure* 5(d) *(two rarefactions in the vapor phase) and is not shown. Vapor-mixture cases:* (c) $-\infty < u_+ - u_- < u_\sharp$; (d) $u_\sharp < u_+ - u_- < u^\sharp$. *In case* (c) *a* 1*-Lax wave followed by a left-moving liquefaction shock, then a right-moving liquefaction shock; in case* (d) *a* 1*-Lax wave followed by an evaporation shock. Also in this case if* $u_+ - u_- > u^\sharp$ *the pattern is analogous to that in Figure* 5(d) *(two rarefactions in the vapor phase) and is not shown.*

It consists of a 1-Lax wave from v_- to v_1, then a left-moving evaporation shock from v_1 to v_2, at last a right-moving evaporation shock from v_2 to v_+; see Figure 6(b). Here $v_1 > v_m$, different from the previous case. We have

$$\begin{cases} -\chi_{(-\infty,v_-)}(v_1)\cdot s_1(v_1 - v_-) + \chi_{[v_-,+\infty)}(v_1)\cdot \displaystyle\int_{v_-}^{v_1} \sqrt{-p_v(v,0)}dv = u_1 - u_-, \\ -s(v_2 - v_1) = u_2 - u_1, \\ -s_2(v_+ - v_2) = u_+ - u_2. \end{cases} \tag{5.5}$$

Remark 5.1. If every evaporation and liquefaction wave connecting two pure phases were admissible then it is clear that uniqueness of solutions to the Riemann problem would fail: we should have two free parameters in system (5.5). This is why we imposed the kinetic condition (ii), section 4; then $s = s(v_2)$ and $v_1 = v_1(v_2)$. More

precisely, let us call $(v_{2-}^T, p(v_{2-}^T, 1))$ the point where the line joining $(v_1, p(v_1, 0))$ and $(v_2, p(v_2, 1))$ becomes tangent to the curve $p(v, 1)$. Then for any $v_2 \in [v_M, v_{2-}^T]$ there exist unique $v_1 > v_m$ and s such that v_1 and v_2 are connected by an evaporation shock of speed s.

Consider, however, for the moment v_2, s as independent variables in (5.5) and define the function

$$F(v_2, s) = -\chi_{(-\infty, v_-)}(v_1) \cdot s_1(v_1 - v_-) + \chi_{[v_-, +\infty)}(v_1) \cdot \int_{v_-}^{v_1} \sqrt{-p_v(v, 0)}dv$$

$$-s(v_2 - v_1) - s_2(v_+ - v_2); \tag{5.6}$$

see [6] for a similar procedure. Then $v_1 = v_1(v_2, s)$, $s_1 = s_1(v_1(v_2, s))$, $s_2 = s_2(v_2)$; for simplicity we omit the dependence on v_2, s in the following calculations.

LEMMA 5.1. *For the function $F(v_2, s)$ defined in (5.6), with $v_1 = v_1(v_2, s)$, $s_1 = s_1(v_1(v_2, s))$, $s_2 = s_2(v_2)$, we have*

$$\frac{\partial F}{\partial v_2}(v_2, s) > 0, \quad \frac{\partial F}{\partial s}(v_2, s) < 0. \tag{5.7}$$

Proof. We begin with some preliminary calculations. By differentiating with respect to v_2 and to s the formula

$$s^2 = -\frac{p(v_2, 1) - p(v_1, 0)}{v_2 - v_1},$$

we find

$$\frac{\partial v_1}{\partial v_2} = \frac{p_v(v_2, 1) + s^2}{p_v(v_1, 0) + s^2}, \qquad \frac{\partial v_1}{\partial s} = \frac{2s(v_2 - v_1)}{p_v(v_1, 0) + s^2}. \tag{5.8}$$

Then $\partial v_1/\partial v_2 > 0$, $\partial v_1/\partial s > 0$ because of the subsonic condition (ii) in section 4. On the other hand, by differentiating with respect to v_2 the formula

$$s_2^2 = -\frac{p(v_+, \lambda_+) - p(v_2, 1)}{v_+ - v_2},$$

we obtain

$$\frac{\partial s_2}{\partial v_2} = \frac{p_v(v_2, 1) + s_2^2}{2s_2(v_+ - v_2)} \tag{5.9}$$

that is positive again because of (ii). At last

$$\frac{ds_1}{dv_1} = -\frac{1}{2s_1(v_1 - v_-)} \left(p_v(v_1, 0) + s_1^2\right). \tag{5.10}$$

Next we compute

$$\frac{\partial F}{\partial v_2}(v_2, s) = \begin{cases} -\left(\dfrac{\partial s_1}{\partial v_1}\dfrac{\partial v_1}{\partial v_2} \cdot (v_1 - v_-) + s_1 \dfrac{\partial v_1}{\partial v_2}\right) + (s_2 - s) - \dfrac{\partial s_2}{\partial v_2}(v_+ - v_2) & \text{if } v_1 < v_-, \\ \left(\sqrt{-p_v(v_1, 0)} + s\right) \dfrac{\partial v_1}{\partial v_2} + (s_2 - s) - \dfrac{\partial s_2}{\partial v_2}(v_+ - v_2) & \text{if } v_1 \geq v_-. \end{cases}$$

We note that $ds_1/dv_1 > 0$, $ds_2/dv_2 > 0$, $s - s_1 > 0$, $s_2 - s > 0$, and $\sqrt{-p_v(v_1,0)} + s > 0$. The first inequality in (5.7) then follows from (5.8), (5.9).

We now compute

$$\frac{\partial F}{\partial s}(v_2,s) = -\chi_{(-\infty,v_-)}(v_1)\cdot\left(\frac{\partial s_1}{\partial s}(v_1-v_-)+s_1\frac{\partial v_1}{\partial s}\right) + \chi_{[v_-,+\infty)}(v_1)\cdot\sqrt{-p_v(v_1,0)}\frac{\partial v_1}{\partial s} - (v_2-v_1) + s\frac{\partial v_1}{\partial s}.$$

Since $\partial s_1/\partial s = (ds_1/dv_1)(\partial v_1/\partial s)$, from (5.8), (5.10) we have

$$\frac{\partial F}{\partial s}(v_2,s) = \begin{cases} \dfrac{v_2-v_1}{s_1(p_v(v_1,0)+s^2)}(s-s_1)\left(p_v(v_1,0)+ss_1\right) & \text{if } v_1 < v_-\,, \\ \dfrac{v_2-v_1}{p_v(v_1,0)+s^2}\left(\sqrt{-p_v(v_1,0)}+s\right)^2 & \text{if } v_1 \geq v_-. \end{cases}$$

Both expressions are negative. This proves the second inequality in (5.7). □

Recall now that we have only one free parameter in the resolution (5.5), i.e., v_2. Therefore

$$\frac{d}{dv_2}\left[F\left(v_2,s(v_2)\right)\right] = \frac{\partial F}{\partial v_2}\left(v_2,s(v_2)\right) + \frac{\partial F}{\partial s}\left(v_2,s(v_2)\right)\cdot\frac{ds}{dv_2}(v_2)\ .$$

Because of assumption (ii) in section 4, we have $\partial s/\partial v_2 < 0$, and from Lemma 5.1 we deduce

$$\frac{d}{dv_2}\left[F\left(v_2,s(v_2)\right)\right] > 0\ .$$

Then define $G(v_2) = F\left(v_2,s(v_2)\right)$. The function G is increasing and

$$\min_{v\geq v_M} G(v_2) = G(v_M) = u_*.$$

This case then matches with the previous one. This wave structure is valid until v_2 reaches the point v_{2+}^T, when the line joining $(v_2,p(v_2,1))$ with $(v_+,p(v_+,\lambda_+))$ becomes tangent to the curve $p(v,1)$ at $(v_2,p(v_2,1))$. The situation is entirely analogous to case 5.1.3. Therefore the upper limit of $u_+ - u_-$ for this case is

$$u^* = G(v_+^T)\ . \tag{5.11}$$

5.2.3. $\boldsymbol{\lambda_- = 0}$, $\boldsymbol{\lambda_+ \in (0,1)}$, $\boldsymbol{u^* \leq u_+ - u_- < +\infty}$. This case is the continuation of the previous one; the solution has again a transition

$$\text{liquid} \to \text{vapor} \to \text{mixture}.$$

The patterns of the solutions are analogous to case 5.1.3. Let v_{2-}^T be the abscissa of the tangency point of the line joining $(v_1,p(v_1,0))$ and $(v_2,p(v_2,1))$ with the curve $p(v,1)$. When $v_2 \in [v_+^T, v_{2-}^T]$ we need to introduce in the solution a 2-rarefaction wave ending with the right-moving evaporation shock; this happens for $u^* \leq u_+ - u_- \leq u^{**}$, for some u^{**}. When finally $v_2 \in [v_{2-}^T, +\infty)$ (that is, $u_+ - u_- \geq u^{**}$), we add also a 1-rarefaction wave beginning with the left-moving evaporation shock. Details are left to the reader.

5.2.4. $\lambda_- = 1$, $\lambda_+ \in (0,1)$, $-\infty < u_+ - u_- \le u_\sharp$. The threshold $u_\sharp$ is defined by

$$u_\sharp = \begin{cases} \displaystyle\int_{v_-}^{v_M} \sqrt{-p_v(v,1)}\,dv & \text{if } v_- \le v_M, \\ \displaystyle\sqrt{-\frac{p(v_M,1)-p(v_-,1)}{v_M - v_-}}(v_M - v_-) & \text{if } v_- \ge v_M. \end{cases} \tag{5.12}$$

In this case we have a transition

$$\text{vapor} \to \text{liquid} \to \text{mixture}.$$

The solution consists of a left-moving Lax wave from v_- to v_1, a left-moving liquefaction wave (uniquely determined by the kinetic condition (i)) from v_1 to v_2, and a right-moving liquefaction wave from v_2 to v_+; see Figure 6(c). This requires $v_1 \le v_M$; otherwise the left-moving liquefaction wave is not admissible. If $v_1 = v_M$ (i.e., $u_+ - u_- = u_\sharp$), the two liquefaction waves are replaced by a single isobaric wave from v_M to v_+. Therefore

$$\begin{cases} \chi_{(-\infty,v_-)}(v_1)\cdot s_1(v_1 - v_-) + \chi_{[v_-,+\infty)}(v_1)\cdot \displaystyle\int_{v_-}^{v_1} \sqrt{-p_v(v,1)}\,dv = u_1 - u_-, \\ -s(v_2 - v_1) = u_2 - u_1, \\ -s_2(v_+ - v_2) = u_+ - u_2. \end{cases}$$

We proceed as in case 5.2.2, the only differences being that 0 and 1 are interchanged for the Lax curves and we have liquefaction instead of evaporation waves. In order to prove estimates analogous to (5.7) we use v_2 and s as independent variables, set $v_1 = v_1(v_2,s)$, $s_1 = s(v_1(v_2,s))$, $s_2 = s_2(v_2)$ and define

$$\begin{aligned} F(v_2,s) = {} & \chi_{(-\infty,v_-)}(v_1)s_1(v_1 - v_-) + \chi_{[v_-,+\infty)}(v_1)\int_{v_-}^{v_1} \sqrt{-p_v(v,1)}\,dv \\ & -s(v_2 - v_1) - s_2(v_+ - v_2). \end{aligned}$$

Now, however, $v_2 - v_1 < 0$ and

$$\frac{\partial v_1}{\partial v_2}(v_2,s) = \frac{p_v(v_2,0) + s^2}{p_v(v_1,1) + s^2} > 0, \quad \frac{\partial v_1}{\partial s}(v_2,s) = \frac{2s(v_2 - v_1)}{p_v(v_1,1) + s^2} < 0.$$

Then

$$\frac{\partial F}{\partial v_2}(v_2,s) > 0, \quad \frac{\partial F}{\partial s}(v_2,s) < 0$$

hold and since $s = s(v_2)$ we have $d/dv_2[F(v_2,s(v_2))] > 0$. Notice that if $v_2 = v_m$, then $v_1 = v_M$. The function F satisfies $\inf_{v_2 \le v_m} F(v_2,s(v_2)) = -\infty$ and reaches its maximum at point v_m. If $v_- \le v_M$, the maximum is reached when the 1-wave is a shock; if $v_- \ge v_M$, it is reached when it is a 1-rarefaction wave. Therefore $\max_{v_2 \le v_M} F(v_2) = u_\sharp$, where $u_\sharp$ is defined in (5.12).

5.2.5. $\lambda_- = 1$, $\lambda_+ \in (0,1)$, $u_\sharp < u_+ - u_- \le u^\sharp$. Here $u^\sharp$ is defined by

$$u^\sharp = \int_{v_-}^{v_+^T} \sqrt{-p_v(v,1)}\,dv - \sqrt{-p_v(v_+^T,1)}(v_+ - v_+^T)$$

for v_+^T defined in (5.1). In this case the solution has a transition

$$\text{vapor} \to \text{mixture}.$$

The solution consists of a 1-Lax wave from v_- to v_1 followed by a right-moving evaporation shock from v_1 to v_+; see Figure 6(d). This construction holds under the following two conditions on v_1. First, the line joining $(v_1, p(v_1, 1))$ and $(v_+, p(v_+, \lambda_+))$ cannot go beyond the tangent to the curve $p(v, 1)$ at point $(v_1, p(v_1, 1))$, i.e., $v_1 < v_+^T$; this position can be reached only if the 1-wave is a rarefaction wave. Second, $v_1 > v_M$; this may happen either if the 1-wave is a shock or a 1-rarefaction wave.

The proof is analogous to case 5.2.1, using the function

$$F(v_1) = -\chi_{(-\infty, v_-)}(v_1) \cdot s_1(v_1 - v_-) + \chi_{[v_-, +\infty)}(v_1) \cdot \int_{v_-}^{v_1} \sqrt{-p_v(v, 1)} dv \\ -s(v_+ - v_1).$$

The function F is increasing and $F(v_M) = u_\sharp$, $F(v_+^T) = u^\sharp$. This case is therefore proved.

5.2.6. $\boldsymbol{\lambda_- = 1}$, $\boldsymbol{\lambda_+ \in (0, 1)}$, $\boldsymbol{u^\sharp \le u_+ - u_- < +\infty}$. Again the solution has a transition

$$\text{vapor} \to \text{mixture}.$$

With respect to the previous case, the solution refers to $v_1 > v_+^T$. It consists of a 1-rarefaction wave from v_- to v_1; a 2-rarefaction wave from v_1 to v_+^T; and a right-moving evaporation shock from v_+^T to v_+. The proof of this case is similar to the proof of case 5.1.3(iii), the only difference being that the 1-rarefaction starts from a generic point v_- and not just from v_-^T. The proof is omitted.

5.3. Data in two pure different phases. In this section we consider the case of initial data (v_-, u_-, λ_-), (v_+, u_+, λ_+) in two different pure phases, that is, either $\lambda_- = 0$ and $\lambda_+ = 1$ or $\lambda_- = 1$ and $\lambda_+ = 0$. For simplicity we treat only the first case; the second case is dealt analogously.

5.3.1. $\boldsymbol{\lambda_- = 0}$, $\boldsymbol{\lambda_+ = 1}$, $\boldsymbol{-\infty < u_+ - u_- \le u_\star}$. Here $u_\star$ is defined in (5.13). The solution consists of a 1-Lax wave from v_- to v_1, a right-moving liquefaction wave from v_1 to v_2, and a 2-Lax wave from v_2 to v_+; see Figure 7(a). The liquefaction wave is replaced by an isobaric wave in the case $u_+ - u_- = u_\star$. This means that

$$\begin{cases} -\chi_{(-\infty, v_-)}(v_1) \cdot s_1(v_1 - v_-) + \chi_{[v_-, +\infty)}(v_1) \cdot \displaystyle\int_{v_-}^{v_1} \sqrt{-p_v(v, 0)}\, dv = u_1 - u_-, \\ -s(v_2 - v_1) = u_2 - u_1, \\ -\chi_{(-\infty, v_+)}(v_2) \cdot s_2(v_+ - v_2) - \chi_{[v_+, +\infty)}(v_2) \cdot \displaystyle\int_{v_2}^{v_+} \sqrt{-p_v(v, 1)}\, dv = u_+ - u_2. \end{cases}$$

Arguing as we did in case 5.2.2 we take for the moment, however, v_1 and s as independent variables, so that $v_2 = v_2(v_1, s)$, $s_1 = s_1(v_1)$, $s_2 = s_2(v_2(v_1, s))$. Denote

$$F(v_1, s) \;=\; -\chi_{(-\infty, v_-)}(v_1) \cdot s_1(v_1 - v_-) + \chi_{[v_-, +\infty)}(v_1) \cdot \int_{v_-}^{v_1} \sqrt{-p_v(v, 0)}\, dv, \\ -s(v_2 - v_1) - \chi_{(-\infty, v_+)}(v_2) \cdot s_2(v_+ - v_2) - \chi_{[v_+, +\infty)}(v_2) \cdot \int_{v_2}^{v_+} \sqrt{-p_v(v, 1)}\, dv.$$

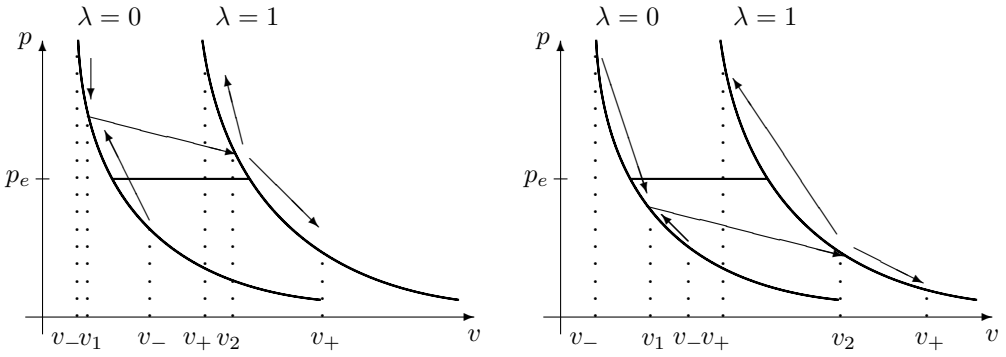


FIG. 7. *Two pure different phases data.* (a) $-\infty < u_+ - u_- < u_\star$; (b) $u_\star < u_+ - u_- \leq u^\star$. *In case* (a) *the phase boundary is a right-moving evaporation wave, in case* (b) *a left-moving liquefaction wave.*

Then

$$\frac{\partial F}{\partial v_1}(v_1, s) = \chi_{(-\infty, v_-)}(v_1)\left(-\frac{\partial s_1}{\partial v_1}(v_1 - v_-) - s_1\right) + \chi_{[v_-, +\infty)}(v_1)\sqrt{-p_v(v_1, 0)}$$
$$-s \cdot \left(\frac{\partial v_2}{\partial v_1} - 1\right)$$
$$+\chi_{(-\infty, v_+)}(v_2)\left(-\frac{\partial s_2}{\partial v_1}(v_+ - v_2) + s_2\frac{\partial v_2}{\partial v_1}\right) + \chi_{[v_+, +\infty)}(v_2)\sqrt{-p_v(v_2, 1)}\frac{\partial v_2}{\partial v_1}\ .$$

We remark now that because of the kinetic condition (i)

$$\frac{\partial v_2}{\partial v_1} = \frac{p_v(v_1, 0) + s^2}{p_v(v_2, 0) + s^2} > 0, \quad \frac{\partial s_1}{\partial v_1} > 0, \quad \frac{\partial s_2}{\partial v_1} = \frac{\partial s_2}{\partial v_2}\frac{\partial v_2}{\partial v_1} < 0.$$

We need only to control the term $-s \cdot \partial v_2/\partial v_1$. But if $v_2 < v_+$, then $s < s_2$, while if $v_2 \geq v_+$, then $s < \sqrt{-p_v(v_2, 1)}$. All that proves that

$$\frac{\partial F}{\partial v_1}(v_1, s) > 0\ .$$

Next compute

$$\frac{\partial F}{\partial s}(v_1, s) = -(v_2 - v_1) - s \cdot \frac{\partial v_2}{\partial s} - \chi_{(-\infty, v_+)}(v_2) \cdot \left(\frac{\partial s_2}{\partial v_2}\frac{\partial v_2}{\partial s}(v_+ - v_2) - s_2 \cdot \frac{\partial v_2}{\partial s}\right)$$
$$+\chi_{[v_+, +\infty)}(v_2) \cdot \sqrt{-p_v(v_2, 1)}\frac{\partial v_2}{\partial s}\ .$$

Moreover

$$\frac{\partial v_2}{\partial s} = -\frac{2s(v_2 - v_1)}{p_v(v_2, 1) + s^2} > 0.$$

If $v_2 < v_+$, then

$$\frac{\partial F}{\partial s}(v_1, s) = \frac{(v_2 - v_1)}{s_2\left(p_v(v_2, 1) + s^2\right)}\left(s_2 - s\right)\left(s_2 s - p_v(v_2, 1)\right) < 0,$$

while if $v_2 \geq v_+$, then

$$\frac{\partial F}{\partial s}(v_1, s) = (v_2 - v_1)\frac{s - \sqrt{-p_v(v_2, 1)}}{s + \sqrt{-p_v(v_2, 1)}} < 0.$$

Recalling the kinetic condition (i) for $s = s(v_1)$, we have $d/dv_1\left[F(v_1, s(v_1))\right] > 0$. So $F(v_1, s(v_1))$ is an increasing function and in the current case has

$$u_\star = F(v_m, 0) \tag{5.13}$$

as maximum. This value can be explicitly computed from above, as in the previous cases, and depends on the four possible configurations of (v_-, v_+): $v_- < v_m$, $v_- > v_m$ or $v_+ < v_M$, $v_+ > v_M$. This concludes the case.

5.3.2. $\boldsymbol{\lambda_- = 0}$, $\boldsymbol{\lambda_+ = 1}$, $\boldsymbol{u_\star < u_+ - u_- \leq u^\star}$. This case is analogous to the previous one, but now a left-moving evaporation wave replaces the liquefaction wave. More precisely the solution consists of a 1-Lax wave, the evaporation shock, and a 2-Lax wave. Now $v_1 > v_m$ (and then $v_2 > v_M$); see Figure 7(b). This pattern is, however, possible for $v_2 \in [v_M, v_+^T]$; the state v_+^T was defined in (5.1).

We define the function

$$\begin{aligned} F(v_2, s) \;=\; & -\chi_{(-\infty, v_-)}(v_1)\cdot s_1(v_1 - v_-) + \chi_{[v_-, +\infty)}(v_1)\cdot \int_{v_-}^{v_1} \sqrt{-p_v(v, 0)}\, dv \\ & -s(v_2 - v_1) - \chi_{(-\infty, v_+)}(v_2)\cdot s_2(v_+ - v_2) - \chi_{[v_+, +\infty)}(v_2)\cdot \int_{v_2}^{v_+} \sqrt{-p_v(v, 1)}\, dv, \end{aligned}$$

where the independent variables are now v_2 and s. Then $v_1 = v_1(v_2, s)$, $s_1 = s_1(v_1(v_2, s))$, $s_2 = s_2(v_2)$. The function $F(v_2, s(v_2))$ is easily proved to be increasing. The critical speed $u^\star$ is then defined as

$$u^\star = F\left(v_+^T, s(v_+^T)\right).$$

Call $(v_*, p(v_*, 0))$ the intersection of the tangent line to the curve $p(v, 1)$ at point $(v_+^T, p(v_+^T, 1))$ with the curve $p(v, 0)$.

5.3.3. $\boldsymbol{\lambda_- = 0}$, $\boldsymbol{\lambda_+ = 1}$, $\boldsymbol{u^\star \leq u_+ - u_- < +\infty}$. Now $v_2 \in [v_+^T, +\infty)$ and the solution consists of a 1-Lax wave from v_- to v_*, an evaporation wave from v_* to v_+^T, a 1-rarefaction wave from v_+^T to v_2 (attached to the evaporation wave on the left), and a 2-rarefaction wave from v_2 to v_+.

The proof of this case is lengthy and is a combination of the cases 5.3.2 and 5.1.3(iii). The details are left to the reader.

5.4. Data in two same pure phases. In this case data (v_-, u_-, λ_-) and (v_+, u_+, λ_+) satisfy either $\lambda_- = \lambda_+ = 0$ or $\lambda_- = \lambda_+ = 1$. We solve the Riemann problem simply using two Lax waves, without phase changes; see [14].

5.5. Comments on the Riemann solver. The Riemann solver defined in the previous sections has many interesting features that we emphasize now.

First, many nonclassical waves arise [2, 3, 12]—for instance, rarefaction waves attached to phase boundaries as in case 5.1.3 (Figures 5(c)–(d)). Also, the case of two phase boundaries with two attached rarefaction waves is present.

Second, the structure of the solution may change abruptly under small changes of the Riemann data. Consider, for instance, case 5.1.2 (Figure 5(b)) and the two near cases 5.1.1 and 5.1.3 (Figures 5(a), (c)). If $u_+ - u_- = 0$, then the solution is an isobaric wave from v_- to v_+. However, if $u_+ - u_-$ is small but nonvanishing, then the solution suffers two phase transitions. In this case the solution has a total variation of order $|v_- - v_m| + |v_+ - v_m|$ (or $|v_- - v_M| + |v_+ - v_M|$) that may be much larger than $|v_+ - v_-|$. The reason for this abrupt change of structure is that when $u_+ - u_-$ go across 0, the phase transition in the solution changes from condensation to evaporization.

A Riemann solver is said to be consistent [4] if the following holds. Take any three state U_-, U_0, U_+, and solve the Riemann problems of initial data (U_-, U_0) and (U_0, U_+); if it is possible to paste horizontally the two solutions (there are no interactions of waves), then the result of the pasting is the solution for the Riemann problem of data (U_-, U_+). Our solutions are easily proved to be consistent if, under the previous notations, $\lambda_- = \lambda_0 = \lambda_+$. In general, however, they are not consistent: For instance, take $\lambda_- = 1$, $\lambda_0 = 0$, $\lambda_+ = 1$. Fix U_-, connect it to U_0 with a left-moving liquefaction wave, and then connect U_0 to some U_+ with a right-moving liquefaction wave. The pasting gives a solution to the Riemann problem of data (U_-, U_+) with two phase boundaries, while our Riemann solver prescribes the Lax solution. This gives two solutions for the same Riemann problem. Intuitively, both solutions are good, but they are good at different times. At first, there are no liquid drops in the initial data, and it takes time for such liquid drops to form through nucleation process. Without seeds for condensation, i.e., liquid drops, the vapor will stay as vapor, and hence the good solution at the early stage is the one without phase changes. After some time, a liquid drop appears in the vapor due to nucleation. Then condensation will occur around the drop, and we will see the second solution with two phase boundaries. Since we are interested in Riemann solvers in the early stages, we picked the one without phase changes as our Riemann solver.

6. Collapsing and explosion waves in the Riemann problem. In section 5 we uniquely solved the Riemann problem by using the waves introduced in section 4. We emphasized, however, in section 2 that other kinds of waves can be considered, namely collapsing and explosion waves. Here we introduce first a selection criterion in order to consider collapsing and explosion waves for system (1.8). Second, we show how collapsing and explosion waves can be used to construct a Riemann solver different from the one defined above. How to deal with the loss of uniqueness introduced by these new waves is left for future investigations. Here we limit ourselves to consider some examples.

The following selection criteria for collapsing and explosion waves are analogous to (i), (ii) in section 4. For simplicity we state the criteria only for waves with positive speed.

(iii) There exists a critical threshold p_{co} such that for each metastable state (v_+, λ_+) with $\lambda_+ = 0$ and $|p(v_+, \lambda_+) - p_e| \geq p_{co}$ there is only one mixture state (v_-, λ_-), $0 < \lambda < 1$, that can be connected to (v_+, λ_+) with a collapsing wave.

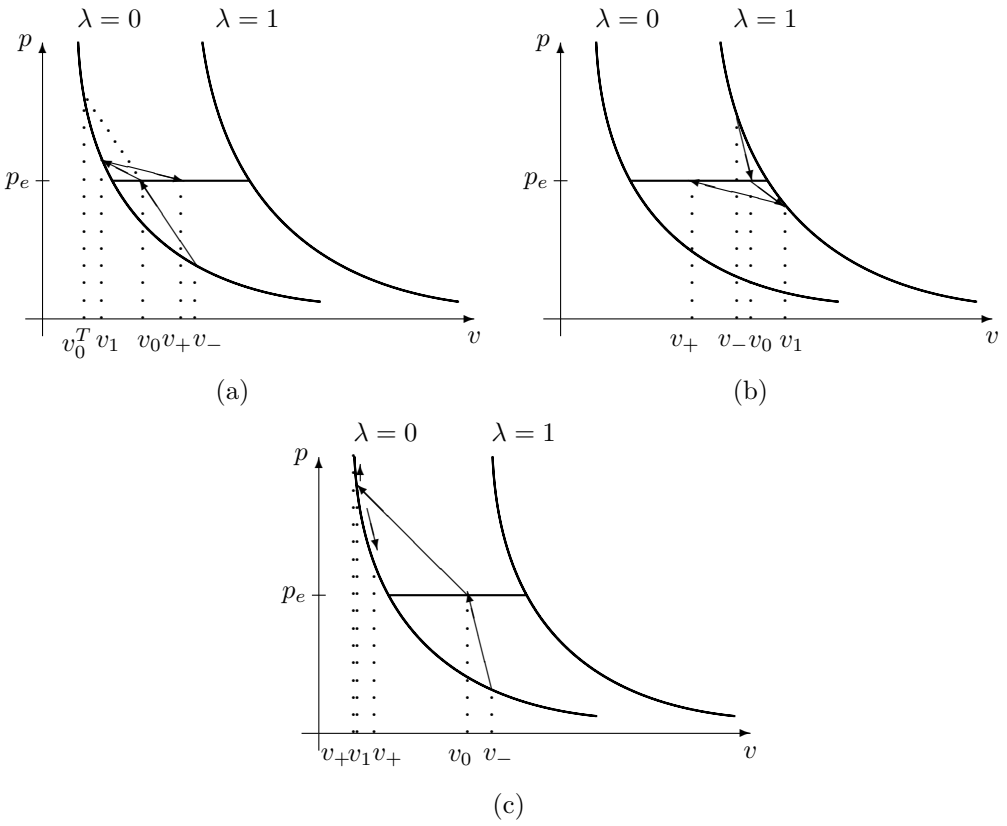


FIG. 8. *Collapsing and explosion waves in the solution of the Riemann problem.*

(iv) There is a critical threshold p_{ex} such that for each metastable state (v_+, λ_+) with $\lambda_+ = 1$ and $|p(v_+, \lambda_+) - p_e| \geq p_{ex}$ there is only one mixture state (v_-, λ_-), $0 < \lambda < 1$, that can be connected to (v_+, λ_+) with an explosion wave.[1]

We consider now some cases that can be solved collapsing or explosion waves.

Example 6.1. This example concerns the case of data (v_-, u_-, λ_-), (v_+, u_+, λ_+) one in a pure phase and one in a mixture. More precisely we assume

$$\lambda_- = 0, \quad \lambda_+ \in (0,1), \quad p_e - p(v_-, 0) \geq p_{ex}, \quad u_\circ \leq u_+ - u_- \leq u^\circ$$

for $u_\circ, u^\circ$ defined below. This case overlaps with case 5.2.1. The solution has a transition

$$\text{liquid} \to \text{mixture} \to \text{liquid} \to \text{mixture}. \tag{6.1}$$

It consists of a left-moving explosion wave from v_- to v_0, then a left-moving liquefaction wave from v_0 to v_1, and finally a right-moving liquefaction wave from v_1 to v_+; see Figure 8(a). When $u_+ - u_- = u^\circ$ the liquefaction waves are replaced by a single

[1] Our numerical tests on (1.7) indicate that at $|p(v_-, 1) - p_e| = p_{co}$, the speed of the collapsing wave is sonic.

isobaric wave. Because of condition (iv) we have $v_0 = v_0(v_-)$; let $\lambda_0 = \lambda_0(v_-)$ be the related mass density fraction and $s_0 = s_0(v_-)$ be the speed of the collapsing wave. Therefore

$$\begin{cases} -s_0(v_0 - v_-) = u_0 - u_-, \\ -s_1(v_1 - v_0) = u_1 - u_0, \\ -s_2(v_+ - v_1) = u_+ - u_1. \end{cases}$$

If we define $F(v_1) = -s_0(v_0 - v_-) - s_1(v_1 - v_0) - s_2(v_+ - v_1)$, then $dF/dv_1 > 0$ (the first summand is constant; for the sum of the other two see case 5.1.1). Now notice that this construction holds until the lines joining $(v_-, p(v_-, 0))$ with $(v_0, p(v_0, \lambda_0))$ and $(v_0, p(v_0, \lambda_0))$ with $(v_1, p(v_1, 0))$ are parallel, that is, until the liquefaction overtakes the explosion wave, so $s_0 = s_1$. Let us define v_0^T as the intersection point of the curve $p = p(v, 0)$ and the line joining $(v_-, p(v_-, 0))$ with $(v_0, p(v_0, \lambda_0))$:

$$\sqrt{-\frac{p(v_0, \lambda_0) - p(v_-, 0)}{v_0 - v_-}} = \sqrt{-\frac{p(v_-^T, 0) - p(v_0, \lambda_0)}{v_0^T - v_0}}\ .$$

From that point on the order of the liquefaction and explosion waves is no more respected. It suffices then to define

$$u_\circ = F(v_0^T), \quad u^\circ = F(v_m).$$

Remark that when $u_+ - u_- = u_\diamond$ the solution with the explosion and the liquefaction shock traveling with equal speed matches with the solution having a single 1-Lax shock, traveling with the same speed. We stress, however, the different behavior of this solution and the one provided by the previous Riemann solver when v_1 is close to v_0^T.

The solution given in Example 6.1 has three phase boundaries. On the other hand, for the same initial data, Figure 6(a) gives a solution of the form

$$\text{liquid} \to \text{liquid with higher pressure} \to \text{mixture} \tag{6.2}$$

with only one phase boundary. One point of Example 6.1 is that a physically relevant solution need not be the one with least number of phase boundaries. For example, when the liquid side of the initial data of Example 6.1 is at or beyond the spinodal limit, the liquid will very quickly evaporate into vapor or liquid/vapor mixture when in contact with the vapor drops of the other side of the initial data. This process is much faster than the sound speed as indicated by the explosion wave. In contrast, the liquid-to-liquid Lax shock in (6.2) is about the sound speed. When solution (6.1) and (6.2) compete, the process that proceeds faster will occur, eliminating the base for the slower process to occur, leading to (6.1). Thus, we choose solution (6.1), which has three phase boundaries, rather than the solution (6.2) with only one phase boundary.

Example 6.2. This example concerns again the case of data one in a pure phase and one in a mixture. We assume, however,

$$\lambda_- = 1, \quad \lambda_+ \in (0, 1), \quad p(v_+, 1) - p_e \geq p_{co}, \quad \tilde{u} \leq u_+ - u_- < +\infty$$

for $\tilde{u}$ defined below. This case overlaps with cases 5.2.4 and 5.2.5.

By using a collapsing wave we define now a solution which has transitions

$$\text{vapor} \to \text{mixture} \to \text{vapor} \to \text{mixture}.$$

The solution consists of a left-moving collapsing wave from v_- to v_0, a left-moving evaporation shock from v_0 to v_1, and a right-moving evaporation shock from v_1 to v_+; see Figure 8(b). If $u_+ - u_- = \tilde{u}$, then the two evaporation waves are replaced by a single isobaric wave from v_1 to v_+. Then

$$\begin{cases} -s_0(v_0 - v_-) = u_0 - u_-, \\ -s_1(v_1 - v_0) = u_1 - u_0, \\ -s_2(v_+ - v_1) = u_+ - u_1. \end{cases}$$

The function $F(v_1) = -s_0(v_0 - v_-) - s_1(v_1 - v_0) - s_2(v_+ - v_1)$ is easily proved to be increasing and reaches its minimum for $v_1 = v_M$; in that case $s_1 = s_2 = 0$ and

$$F(v_M) = \tilde{u} = -s_0(v_-) \cdot (v_0(v_-) - v_-)\,.$$

This pattern changes when the line joining $(v_0, p(v_0, \lambda_0))$ and $(v_1, p(v_1, 1))$ becomes tangent to the curve $p(v, 1)$ at point $(v_1, p(v_1, 1))$, that is, for $v_1 = v_0^T$. From that point on the pattern is as follows: a left-moving collapsing wave from v_- to v_0, a left-moving evaporation shock from v_0 to v_0^T, a 1-rarefaction wave from v_0^T to v_1, and a right-moving evaporation shock from v_1 to v_+. The pattern changes once more when $v_1 = v_+^T$ (defined in (5.1)). From that point on the pattern is as follows: a left-moving collapsing wave from v_- to v_0, a left-moving evaporation shock from v_0 to v_0^T, a 1-rarefaction wave from v_0^T to v_1, a 2-rarefaction wave from v_1 to v_+^T, and a right-moving evaporation shock from v_+^T to v_+.

The proof of these three last cases is analogous to that of case 5.1.3. The details are left to the reader.

Example 6.3. The following example refers to the case of two states both in a same pure phase; see subsection 5.4. We assume

$$\lambda_- = \lambda_+ = 0, \quad p_e - p(v_-, 0) \geq p_{ex}, \quad u_\diamond \leq u_+ - u_- \leq u^\diamond$$

for $u_\diamond$ and $u^\diamond$ defined below. The solution involves a transition

$$\text{liquid} \to \text{mixture} \to \text{liquid}.$$

It consists of a left-moving explosion wave connecting v_- to v_0, a left-moving liquefaction wave from v_0 to v_1, and a 2-Lax wave from v_1 to v_+; see Figure 8(c). The liquefaction wave becomes an isobaric wave if $u_+ - u_- = u^\diamond$. Therefore

$$\begin{cases} -s_0(v_0 - v_-) = u_0 - u_-, \\ -s(v_1 - v_0) = u_1 - u_0, \\ -\chi_{(-\infty, v_+)}(v_1) \cdot s_2(v_+ - v_1) - \chi_{[v_+, +\infty)}(v_1) \displaystyle\int_{v_1}^{v_+} \sqrt{-p_v(v, 0)}\, dv = u_+ - u_1. \end{cases}$$

Define then

$$\begin{aligned} &F(v_1) = -s_0(v_0 - v_-) - s(v_1 - v_0) \\ &-\chi_{(-\infty, v_+)}(v_1) \cdot s_2(v_+ - v_1) - \chi_{[v_+, +\infty)}(v_1) \int_{v_1}^{v_+} \sqrt{-p_v(v, 0)}\, dv. \end{aligned}$$

We have that $ds/dv_1 > 0$, $ds_2/dv_1 < 0$ and so $dF/dv_1 > 0$. As in Example 6.1, this construction fails when the liquefaction overtakes the explosion shock, i.e., when

the lines joining, respectively, $(v_-, p(v_-, 0))$ with $(v_0, p(v_0, 0))$ and $(v_0, p(v_0, 0))$ with $(v_1, p(v_1, 0))$ become parallel, that is, $s_0 = s$. Define $v_0^T < v_m$ by

$$\sqrt{-\frac{p(v_0(v_-), \lambda_0(v_-)) - p(v_-, 0)}{v_0(v_-) - v_-}} = \sqrt{-\frac{p(v_0^T, 0) - p(v_0(v_-), \lambda_0(v_-))}{v_0^T - v_0(v_-)}}.$$

We then define

$$u_\diamond = F(v_0^T), \quad u^\diamond = F(v_m).$$

For $u_+ - u_- \le u_\diamond$ the construction of solutions to the Riemann problem can be continued proceeding as in subsection 5.4.

Example 6.4. Consider again the case of two pure states, same phase. Assume $\lambda_- = \lambda_+ = 1$, $p(v_-, 1) - p_e \ge p_{co}$; for simplicity we omit the bounds for $u_+ - u_-$. The solution has a transition

$$\text{vapor} \to \text{mixture} \to \text{vapor}.$$

It connects v_- to v_0 with a left-moving collapsing wave, v_0 to v_1 with a left-moving evaporation wave, and, finally, v_1 to v_+ with a 2-Lax wave. This construction, however, can be done until the line joining $(v_0, p(v_0, 1))$ and $(v_1, p(v_1, 1))$ becomes tangent to the graph of the function $p(v, 1)$. From that point on a construction similar to the one of Example 6.2 can be done. We omit the details.

Acknowledgment. This work was partly accomplished while A. Corli and H. Fan visited the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK, during the programme Nonlinear Hyperbolic Waves in Phase Dynamics and Astrophysics.

REFERENCES

[1] R. Abeyaratne and J. K. Knowles, *Kinetic relations and the propagation of phase boundaries in solids*, Arch. Rational Mech. Anal., 114 (1991), pp. 119–154.

[2] R. M. Colombo and A. Corli, *Sonic hyperbolic phase transitions and Chapman-Jouguet detonations*, J. Differential Equations, 184 (2002), pp. 321–347.

[3] R. M. Colombo and A. Corli, *Sonic and kinetic phase transitions with applications to Chapman-Jouguet deflagrations*, Math. Methods Appl. Sci., 27 (2004), pp. 843–864.

[4] R. M. Colombo and F. Priuli, *Characterization of Riemann solvers for the two phase p-system*, Comm. Partial Differential Equations, 28 (2003), pp. 1371–1389.

[5] G. Dettleff, P. A. Thompson, G. E. A. Meier, and H.-D. Speckmann, *An experimental study of liquefaction shock waves*, J. Fluid Mech., 95 (1979), pp. 279–304.

[6] H. Fan, *The uniqueness and stability of the solution of the Riemann problem of a system of conservation laws of mixed type*, Trans. Amer. Math. Soc., 333 (1992), pp. 913–938.

[7] H. Fan, *Traveling waves, Riemann problems and computations of a model of the dynamics of liquid/vapor phase transitions*, J. Differential Equations, 150 (1998), pp. 385–437.

[8] H. Fan, *Convergence to traveling waves in two model systems related to the dynamics of liquid-vapor phase changes*, J. Differential Equations, 168 (2000), pp. 102–128.

[9] H. Fan, *On a model of the dynamics of liquid/vapor phase transitions*, SIAM J. Appl. Math., 60 (2000), pp. 1270–1301.

[10] H. Fan, *Symmetry breaking, ring formation and other phase boundary structures in shock tube experiments on retrograde fluids*, J. Fluid Mech., 513 (2004), pp. 47–75.

[11] E. Godlewski and P.-A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Appl. Math. Sci. 118, Springer-Verlag, New York, 1996.

[12] P. G. LeFloch, *Hyperbolic Systems of Conservation Laws. The Theory of Classical and Nonclassical Shock Waves*, Lectures Math. ETH Zürich, Birkhäuser, Basel, 2002.

[13] M. Shearer, *Nonuniqueness of admissible solutions of Riemann initial value problems for a system of conservation laws of mixed type*, Arch. Rational Mech. Anal., 93 (1986), pp. 45–59.

[14] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Grundlehren Math. Wiss. 258, Springer-Verlag, New York, 1983.

[15] P. A. Thompson, H. Chaves, G. E. A. Meier, Y.-G. Kim, and H.-D. Speckmann, *Wave splitting in a fluid of large heat capacity*, J. Fluid Mech., 185 (1987), pp. 385–414.

SIAM J. APPL. MATH.
Vol. 65, No. 2, pp. 458–474

SPATIAL DECAY BOUNDS IN TIME DEPENDENT PIPE FLOW OF AN INCOMPRESSIBLE VISCOUS FLUID*

CHANGHAO LIN† AND LAWRENCE E. PAYNE‡

Abstract. In this paper, the authors investigate the flow of an incompressible viscous fluid in a semi-infinite cylindrical pipe. If the net entrance flow is nonzero, then the fluid velocity will not tend to zero as the distance from the entrance end tends to infinity when the fluid adheres at the cylinder wall and the fluid is initially at rest. Assuming that the entrance velocity data are small enough and that the fluid flow converges to a laminar flow as the distance down the pipe tends to infinity, it is shown that the convergence in energy measure is at least exponential.

Key words. viscous pipe flow, decay bounds, Saint-Venant's principle, Navier–Stokes equation

AMS subject classifications. 35B40, 35Q30, 76D05, 76D07

DOI. 10.1137/040606326

1. Introduction.

A number of papers in the literature have dealt with the transient flow of an incompressible viscous fluid in a semi-infinite pipe or channel. If the net flow into the finite end of the pipe or channel at any time is not zero, then the flow velocity cannot go to zero as the distance from the finite end tends to infinity. If the flow is governed by the linear Stokes equations, then the case in which the net entry flow is zero has been considered in $\mathbb{R}^2$ by Lin [14] and in $\mathbb{R}^3$ by Ames, Payne, and Schaefer [2]. The case of nonzero channel entry flow has been investigated by Song [22]. When the flow is governed by the Navier–Stokes equation, the general entry flow problem in $\mathbb{R}^2$ has been treated by Lin and Payne [15]. In this paper, we study the analogous problem in $\mathbb{R}^3$.

It is well known that unless there is some restriction on data, coefficients, and geometry, a global solution may not exist. Also if the net entry flow in the pipe is nonzero, we expect the velocity profile to converge to that of transient laminar flow as the distance from the finite end of the cylinder tends to infinity. The object of this paper is to derive explicit conditions on the data, coefficients, and geometry that will imply exponential decay of the transient flow to transient laminar flow in some appropriate weighted measure. We should point out, however, that there are various ways of handling boundary terms that appear in the computations as well as various ways of combining the inequalities. We have not attempted to derive optimal results, since that would excessively lengthen the already involved computations.

For related work on steady solutions of the Navier–Stokes equations for flows in channels and pipes, see, for instance, [1, 2, 3, 4, 7, 12]. These pipe and channel flow results may be regarded as Saint-Venant type decay results. In fact, the first paper to point out this connection was that of Horgan and Wheeler [12]. For other results of Saint-Venant type, see [8, 9, 10]. Of interest also are the papers [16, 17].

*Received by the editors April 5, 2004; accepted for publication (in revised form) May 27, 2004; published electronically December 16, 2004.

http://www.siam.org/journals/siap/65-2/60632.html

†Department of Mathematics, South China Normal University, Guangzhou, 510631, People's Republic of China (linchh@scnu.edu.cn). The research of this author was supported by the National Science Foundation of China (grant 19970130) and Guangdong Provincial Natural Science Foundation of China (grant 03149).

‡Department of Mathematics, Cornell University, Ithaca, NY 14853 (lep8@cornell.edu).

In section 2 we give a formulation of the problem. In section 3 we present some auxiliary inequalities used throughout the paper. In section 4 we derive the energy decay bounds in terms of total weighted energies and properties of the limiting solution. The laminar flow bounds are obtained in section 5 and the required total energies are obtained in section 6.

2. Formulation of the problem. In this section we introduce the boundary value problem that provides the basis for our investigation of the pipe flow of an incompressible viscous fluid.

Let R denote the interior of a semi-infinite three-dimensional cylindrical pipe and let ∂R denote its boundary. The uniform cross section is denoted by D. We assume the generators of the cylinder are parallel to the x_3 axis and that the entry section of the cylinder lies in the plane $x_3 = 0$. The symbol R_z designates the subdomain of R for which $x_3 > z \geq 0$, i.e.,

$$R_z = \{(x_1, x_2, x_3) \mid (x_1, x_2) \in D, x_3 \geq z \geq 0\}.$$

Clearly $R \equiv R_0$. Furthermore, we denote the cross section D with $x_3 = z$ by the symbol D_z, i.e.,

$$D_z = \{(x_1, x_2, x_2) \mid (x_1, x_2) \in D, x_3 = z\}.$$

The velocity field $u_i(x_1, x_2, x_3, t)$, $(i = 1, 2, 3)$ and the pressure $p(x_1, x_2, x_3, t)$ for the transient Navier–Stokes flow of an incompressible viscous fluid in the pipe are assumed to be classical solutions of the following initial-boundary value problem:

$$u_{i,t} - \nu \Delta u_i + u_j u_{i,j} = p_{,i} \quad \text{in } R \times \{t > 0\}, \tag{2.1}$$

$$u_{i,i} = 0 \quad \text{in } R \times \{t > 0\} \tag{2.2}$$

with

$$u_i = 0 \quad \text{on } \partial D \times \{t \geq 0\}, \tag{2.3}$$

$$u_i = f_i(x_1, x_2, t) \quad \text{on } D_0 \times \{t > 0\}, \tag{2.4}$$

$$u_i = 0 \quad \text{in } R \times \{t = 0\}, \tag{2.5}$$

where Δ denotes the Laplace operator and ν is the constant kinematic viscosity. We have used the comma to denote partial differentiation and have adopted the summation convention of summing over a repeated Latin subscript from 1 to 3 and over a repeated Greek index (unless otherwise specified) from 1 to 2. By rescaling the space and time variables, we may take the constant ν to be 1. In general we shall assume that time lies in some finite interval $[0, T]$.

It is well known that a global bounded solution for $t \in [0, T]$ may not exist. However, if the data f_i are sufficiently small in L_2, a bounded solution will exist. In fact, if the mean value of f_3 over D is zero, we expect the solution in some appropriate measure to vanish exponentially. However, if f_3 does not have mean value zero, we expect that for sufficiently small data, the velocity field (u_1, u_2, u_3) will tend exponentially to $(0, 0, V)$ as $x_3 \to \infty$, where $V(x_1, x_2, t)$ satisfies

$$V_{,t} - V_{,\alpha\alpha} = P(t) \quad \text{in } D \times \{t > 0\}, \tag{2.6}$$

$$V = 0 \quad \text{on } \partial D \times \{t > 0\}, \tag{2.7}$$

$$V = 0 \quad \text{in } D \times \{t = 0\}. \tag{2.8}$$

The function $P(t)$ is not prescribed but is determined by the condition

$$\int_D V(x_1, x_2, t)dA = \int_D f_3(x_1, x_2, t)dA = Q(t). \tag{2.9}$$

The problem (2.6)–(2.8) may be viewed as an inverse problem for determining $P(t)$ and V.

We now set

$$w_i = u_i - v_i, \quad q_{,i} = p_{,i} - P(t)\delta_{i3}, \tag{2.10}$$

where

$$(v_1, v_2, v_3) = (0, 0, V).$$

Then (w_i, q) will satisfy the following initial-boundary value problem:

$$w_{i,t} - \Delta w_i + (w_j + v_j)(w_{i,j} + v_{i,j}) = q_{,i} \quad \text{in } R \times \{t > 0\}, \tag{2.11}$$

$$w_{i,i} = 0 \quad \text{in } R \times \{t > 0\}, \tag{2.12}$$

$$w_i = 0 \quad \text{on } \partial D \times \{t \geq 0\}, \tag{2.13}$$

$$w_i = f_i - V\delta_{i3} \quad \text{in } D_0 \times \{t \geq 0\}, \tag{2.14}$$

$$w_i = 0 \quad \text{in } R \times \{t = 0\}. \tag{2.15}$$

We suppose further that for any finite positive constant K, the energy expression

$$E(0, t) = \int_0^t \int_{R_0} x_3^2 w_{i,j} w_{i,j} dx d\eta + K \int_0^t \int_{R_0} x_3^2 w_{i,\eta} w_{i,\eta} dx d\eta \tag{2.16}$$

is bounded. Here η is a running time variable, and there is no summation over η.

We note that

$$\int_{D_z} w_3 dA = \int_{D_0} w_3 dA + \int_0^z \int_{D_\xi} w_{i,i} dA d\xi = 0. \tag{2.17}$$

3. Auxiliary results. We list in this section a number of standard inequalities used throughout this paper.

Let w be a Dirichlet integrable function defined on a bounded plane domain D and vanishing on the boundary ∂D; then

$$\int_D w_{,\alpha} w_{,\alpha} dA \geq \lambda_1 \int_D w^2 dA, \tag{3.1}$$

where λ_1 is the smallest eigenvalue of the problem

$$\Delta\varphi + \lambda\varphi = 0 \quad \text{in } D,$$
$$\varphi = 0 \quad \text{on } \partial D.$$

Lower bounds for λ_1 are well known; see, e.g., [6, 18, 19].

We also make use of the following representation theorem attributed to Babuška and Aziz [5].

THEOREM A. *Let D be a plane Lipschitz domain and let w be a differentiable function in D which satisfies $\int_D w dA = 0$. Then there exists a vector function φ_α such that*

$$\begin{aligned} \varphi_{\alpha,\alpha} &= w \quad \text{in } D, \\ \varphi_\alpha &= 0 \quad \text{on } \partial D, \end{aligned}$$

and a positive constant C depending only on the geometry of D such that

$$\int_D \varphi_{\alpha,\beta}\varphi_{\alpha,\beta} dA \le C \int_D \varphi_{\alpha,\alpha}^2 dA. \tag{3.2}$$

This theorem was first applied to viscous flow problems by Horgan and Wheeler [12]. In fact, an explicit upper bound for the optimal value of C was found by Horgan and Payne [11] if D is star-shaped. The analogue of Theorem A also holds in $\mathbb{R}^3$ (see [23]). This inequality will allow us to eliminate the pressure function difference term q, since w_3 satisfies the hypothesis of this theorem.

In addition to inequalities (3.1), (3.2), we also make use of the following Sobolev inequalities which hold for $w \in C_0^1(D)$ and $w \in C_0^1(R)$, respectively:

$$\int_D w^4 dA \le \frac{1}{2}\left[\int_D w^2 dA\right]\left[\int_D w_{,\alpha} w_{,\alpha} dA\right], \tag{3.3}$$

$$\int_{R_z} w^6 dx \le \Omega \left[\int_{R_z} w_{,i} w_{,i} dx\right]^3. \tag{3.4}$$

For (3.4), we assume that w vanishes appropriately as $x_3 \to \infty$. A derivation of (3.3) is given by Serrin [20] and Payne [17], while (3.4) follows as a special case of results of [13, 21], where the optimal value of Ω was determined to be

$$\Omega = \frac{1}{27}\left(\frac{3}{4}\right)^4.$$

In what follows, we also make frequent use of the fact that if $w \in C^1(R_z)$ and w_i vanishes on ∂D for $x_3 \ge z$, then, if w_i vanishes as $x_3 \to \infty$,

$$\begin{aligned} \int_{D_z} (w_i w_i)^2 dA &= -4 \int_{R_z} w_i w_{i,3} w_j w_j dx \\ &\le 4 \left[\int_{R_z} w_{i,3} w_{i,3} dx\right]^{1/2} \left[\int_{R_z} (w_j w_j)^3 dx\right]^{1/2} \\ &\le 4\sqrt{\Omega}\left[\int_{R_z} w_{i,j} w_{i,j} dx\right]^2. \end{aligned} \tag{3.5}$$

4. Energy decay bounds. In this section we derive the main exponential decay result for problems (2.1)–(2.15).

First, for arbitrary $z > 0$ and $t > 0$, we define a weighted energy integral

$$\begin{aligned} E(z,t) &= \int_0^t \int_{R_z} (\xi - z)^2 w_{i,j} w_{i,j} dx d\eta + K \int_0^t \int_{R_z} (\xi - z)^2 w_{i,\eta} w_{i,\eta} dx d\eta \\ &= E_1(z,t) + K E_2(z,t). \end{aligned} \tag{4.1}$$

We also define

$$E_3(z,t) = \int_0^t \int_{R_z} (\xi - z)^2 w_{i,j\eta} w_{i,j\eta} dx d\eta, \tag{4.2}$$

where there is no summation on the η subscript and K is a positive parameter. Note that (2.16) has been assumed, but now we further require that $E_3(0,t)$ be bounded. Clearly then

$$\frac{\partial E}{\partial z} = -2\int_0^t \int_{R_z} (\xi - z) w_{i,j} w_{i,j} dx d\eta - 2K \int_0^t \int_{R_z} (\xi - z) w_{i,\eta} w_{i,\eta} dx d\eta,$$
$$\frac{\partial^2 E}{\partial z^2} = 2\int_0^t \int_{R_z} w_{i,j} w_{i,j} dx d\eta + 2K \int_0^t \int_{R_z} w_{i,\eta} w_{i,\eta} dx d\eta.$$

Upon integrating by parts in (4.1) and using equations (2.11) and (2.12), we obtain

$$\begin{aligned} &E(z,t) + \frac{1}{2}\int_{R_z} (\xi - z)^2 w_i w_i dx\big|_{\eta=t} + \frac{K}{2}\int_{R_z} (\xi - z)^2 w_{i,j} w_{i,j} dx\big|_{\eta=t} \\ &\quad = -2\int_0^t \int_{R_z} (\xi - z) w_i w_{i,3} dx d\eta - \int_0^t \int_{R_z} (\xi - z)^2 w_i [w_j w_{i,j} + v_j w_{i,j} + w_j v_{i,j}] dx d\eta \\ &\qquad - 2\int_0^t \int_{R_z} (\xi - z) w_3 q dx d\eta - 2K \int_0^t \int_{R_z} (\xi - z) w_{i,\eta} w_{i,3} dx d\eta \\ &\qquad - 2K \int_0^t \int_{R_z} (\xi - z) w_{3,\eta} q dx d\eta \\ &\qquad - K \int_0^t \int_{R_z} (\xi - z)^2 w_{i,\eta} [w_j w_{i,j} + v_j w_{i,j} + w_j v_{i,j}] dx d\eta \\ &\quad = \sum_{i=1}^{6} I_i. \end{aligned} \tag{4.3}$$

We now proceed to bound each integral I_i. Using Schwarz's inequality and (3.1), we obtain

$$\begin{aligned} I_1 &\le 2\left(\int_0^t \int_{R_z} (\xi - z) w_i w_i dx d\eta\right)^{1/2} \left(\int_0^t \int_{R_z} (\xi - z) w_{i,3} w_{i,3} dx d\eta\right)^{1/2} \\ &\le \lambda_1^{-1/2} \left(-\frac{\partial E}{\partial z}\right). \end{aligned} \tag{4.4}$$

We next look at

$$\begin{aligned} I_2 &= \int_0^t \int_{R_z} (\xi - z) w_i w_i w_3 dx d\eta - \int_0^t \int_{R_z} (\xi - z)^2 w_i w_{i,3} V dx d\eta \\ &\qquad - \int_0^t \int_{R_z} (\xi - z)^2 w_3 w_\alpha V_{,\alpha} dx d\eta \\ &= I_{21} + I_{22} + I_{23}. \end{aligned} \tag{4.5}$$

By using Schwarz's inequality and (3.3), (3.5), we obtain

$$
\begin{aligned}
I_{21} &\le \int_0^t \int_z^\infty (\xi - z) \left(\int_{D_\xi} (w_i w_i)^2 dA \right)^{1/2} \left(\int_{D_\xi} w_3^2 dA \right)^{1/2} d\xi d\eta \\
&\le \frac{2^{1/4} \Omega^{1/8}}{\lambda_1^{3/4}} \int_0^t \left(\int_{R_z} w_{i,j} w_{i,j} dx \right)^{1/2} \left(\int_{R_z} (\xi - z)^2 w_{i,j} w_{i,j} dx \right)^{1/2} \\
&\qquad \cdot \left(\int_{R_z} w_{i,j} w_{i,j} dx \right)^{1/2} d\eta \\
&\le \frac{2^{1/4} \Omega^{1/8}}{\lambda_1^{3/4}} \max_t \left(\int_{R_z} (\xi - z)^2 w_{i,j} w_{i,j} dx \right)^{1/2} \left[\frac{\partial^2 E_1}{\partial z^2}(0,t) \right]^{1/2} \left[\frac{\partial^2 E}{\partial z^2} \right]^{1/2} \\
&\le \frac{\varepsilon_1}{2} \max_t \left(\int_{R_z} (\xi - z)^2 w_{i,j} w_{i,j} dx \right)^{1/2} + \frac{\Omega^{1/4}}{2^{1/2} \lambda_1^{3/2} \varepsilon_1} \left[\frac{\partial^2 E_1}{\partial z^2}(0,t) \right] \left[\frac{\partial^2 E}{\partial z^2} \right] \qquad (4.6)
\end{aligned}
$$

for arbitrary positive constant ε_1.

For I_{22}, we have

$$
\begin{aligned}
I_{22} &= -\int_0^t \int_{R_z} (\xi - z)^2 w_i w_{i,3} V dx d\eta \\
&= \int_0^t \int_{R_z} (\xi - z) w_i w_i V dx d\eta \\
&\le \frac{1}{\lambda_1} |V|_{\max} \left[-\frac{\partial E}{\partial z} \right]. \qquad (4.7)
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
I_{23} &= -\int_0^t \int_{R_z} (\xi - z)^2 w_3 w_\alpha V_{,\alpha} dx d\eta \\
&= \int_0^t \int_{R_z} (\xi - z)^2 w_{3,\alpha} w_\alpha V dx d\eta - \int_0^t \int_{R_z} (\xi - z)^2 w_3 w_{3,3} V dx d\eta \\
&\le \frac{1}{2} \lambda_1^{-1/2} |V|_{\max} E(z,t) + \frac{1}{\lambda_1} |V|_{\max} \left[-\frac{\partial E}{\partial z} \right]. \qquad (4.8)
\end{aligned}
$$

To seek a bound for I_3, we note that, for any $z > 0$,

$$
\int_{D_z} w_3 dA = 0. \qquad (4.9)
$$

Accordingly by Theorem A, there exists a vector function (φ_1, φ_2) such that

$$
\begin{cases} \varphi_{\alpha,\alpha} = w_3 & \text{in } D_z, \\ \varphi_\alpha = 0 & \text{on } \partial D_z, \end{cases}
$$

and for φ_α inequality (3.2) holds. Introducing φ_α into I_3, we obtain

$$
\begin{aligned}
I_3 &= -2 \int_0^t \int_{R_z} (\xi - z) \varphi_{\alpha,\alpha} q dx d\eta \\
&= 2 \int_0^t \int_{R_z} (\xi - z) \varphi_\alpha [w_{\alpha,\eta} + w_j w_{\alpha,j} + v_j w_{\alpha,j} - \Delta w_\alpha] dx d\eta \\
&= I_{31} + I_{32} + I_{33} + I_{34}. \qquad (4.10)
\end{aligned}
$$

Obviously, by using Schwarz's inequality and (3.2), we have

$$I_{31} \le \frac{2}{\sqrt{K}} \left(\int_0^t \int_{R_z} (\xi - z) \varphi_\alpha \varphi_\alpha dx d\eta \right)^{1/2} \left(K \int_0^t \int_{R_z} (\xi - z) w_{\alpha,\eta} w_{\alpha,\eta} dx d\eta \right)^{1/2}$$

$$\le \frac{1}{\lambda_1} (C/K)^{1/2} \left[-\frac{\partial E}{\partial z} \right]. \tag{4.11}$$

For I_{32}, with a derivation similar to (4.6), we obtain

$$\begin{aligned} I_{32} &\le 2 \int_0^t \int_z^\infty (\xi - z) \left(\int_{D_\xi} (\varphi_\alpha \varphi_\alpha)^2 dA \right)^{1/4} \left(\int_{D_\xi} (w_i w_i)^2 dA \right)^{1/4} \\ &\qquad \cdot \left(\int_{D_\xi} w_{\alpha,j} w_{\alpha,j} dA \right)^{1/2} d\xi d\eta \\ &\le \frac{2(2C)^{1/2} \Omega^{1/8}}{2^{1/4} \lambda_1^{1/2}} \int_0^t \left(\int_{R_z} w_{i,j} w_{i,j} dx \right)^{1/2} \left(\int_{R_z} (\xi - z)^2 w_{i,j} w_{i,j} dx \right)^{1/2} \\ &\qquad \cdot \left(\int_{R_z} w_{i,j} w_{i,j} dx \right)^{1/2} d\eta \\ &\le \frac{2(2C)^{1/2} \Omega^{1/8}}{2^{1/4} \sqrt{\lambda_1}} \max_t \left(\int_{R_z} (\xi - z)^2 w_{i,j} w_{i,j} dx \right)^{1/2} \left[\frac{\partial^2 E_1}{\partial z^2}(0,t) \right]^{1/2} \left[\frac{\partial^2 E}{\partial z^2} \right]^{1/2} \\ &\le \frac{\varepsilon_2}{2} \max_t \left(\int_{R_z} (\xi - z)^2 w_{i,j} w_{i,j} dx \right) + \frac{2^{3/2} \Omega^{1/4}}{\lambda_1 \varepsilon_2} \left[\frac{\partial^2 E}{\partial z^2}(0,t) \right] \left[\frac{\partial^2 E}{\partial z^2} \right]. \end{aligned} \tag{4.12}$$

Using (3.2), we find

$$\begin{aligned} I_{33} &\le 2 \int_0^t \int_{R_z} (\xi - z) \varphi_\alpha w_{\alpha,3} V dx d\eta \\ &\le 2 |V|_{\max} \left(\int_0^t \int_{R_z} (\xi - z) \varphi_\alpha \varphi_\alpha dx d\eta \right)^{1/2} \left(\int_0^t \int_{R_z} (\xi - z) w_{\alpha,3} w_{\alpha,3} dx d\eta \right) \\ &\le \frac{C^{1/2}}{\lambda_1} |V|_{\max} \left[-\frac{\partial E}{\partial z} \right], \end{aligned} \tag{4.13}$$

and

$$\begin{aligned} I_{34} &= -2 \int_0^t \int_{R_z} (\xi - z) \varphi_\alpha w_{\alpha,jj} dx d\eta \\ &= 2 \int_0^t \int_{R_z} \varphi_\alpha w_{\alpha,3} dx d\eta + 2 \int_0^t \int_{R_z} (\xi - z) \varphi_{\alpha,j} w_{\alpha,j} dx d\eta \\ &\le \frac{C^{1/2}}{\lambda_1} \left[\frac{\partial^2 E}{\partial z^2} \right] + (C/\lambda_1)^{1/2} \left[-\frac{\partial E}{\partial z} \right]. \end{aligned} \tag{4.14}$$

By using Schwarz's inequality, we obtain

$$I_4 = -2K \int_0^t \int_{R_z} (\xi - z) w_{i,\eta} w_{i,3} dx d\eta \le K^{1/2} \left[-\frac{\partial E}{\partial z} \right]. \tag{4.15}$$

For I_5, since

$$\int_{D_z} w_{3,t} dA = 0,$$

by Theorem A, there exists a vector function (ψ_1, ψ_2) that satisfies the boundary value problem

$$\begin{cases} \psi_{\alpha,\alpha} = w_{3,t} & \text{in } D_z, \\ \psi_\alpha = 0 & \text{on } \partial D_z. \end{cases}$$

Introducing ψ_α, we can write

$$\begin{aligned} I_5 &= -2K \int_0^t \int_{R_z} (\xi - z) \psi_{\alpha,\alpha} q dx d\eta \\ &= 2K \int_0^t \int_{R_z} (\xi - z) \psi_\alpha [w_{\alpha,\eta} + w_j w_{\alpha,j} + v_j w_{\alpha,j} - \Delta w_\alpha] dx d\eta \\ &= I_{51} + I_{52} + I_{53} + I_{54}. \end{aligned} \tag{4.16}$$

It is easy to see that

$$\begin{aligned} I_{51} &\le 2K \left(\int_0^t \int_{R_z} (\xi - z) \psi_\alpha \psi_\alpha dx d\eta \right)^{1/2} \left(\int_0^t \int_{R_z} (\xi - z) w_{\alpha,\eta} w_{\alpha,\eta} dx d\eta \right)^{1/2} \\ &\le (C/\lambda_1)^{1/2} \left[-\frac{\partial E}{\partial z} \right]. \end{aligned} \tag{4.17}$$

In a manner similar to (4.6), (4.11), we find

$$\begin{aligned} I_{52} &= 2K \int_0^t \int_{R_z} (\xi - z) \psi_\alpha w_\alpha w_{\alpha,j} dx d\eta \\ &\le 2K \int_0^t \int_z^{+\infty} (\xi - z) \left(\int_{D_\xi} (\psi_\alpha \psi_\alpha)^2 dA \right)^{1/4} \left(\int_{D_\xi} (w_\alpha w_\alpha)^2 dA \right)^{1/4} \\ &\qquad \cdot \left(\int_{D_\xi} w_{\alpha,j} w_{\alpha,j} dA \right)^{1/2} d\xi d\eta \\ &\le \frac{2(2CK)^{1/2} \Omega^{1/8}}{(2\lambda_1)^{1/4}} \max_t \left(\int_{R_z} (\xi - z)^2 w_{i,j} w_{i,j} dx \right)^{1/2} \left[\frac{\partial^2 E_1}{\partial z^2}(0,t) \right]^{1/2} \left[\frac{\partial^2 E}{\partial z^2} \right]^{1/2} \\ &\le \frac{\varepsilon_3}{2} \max_t \left(\int_{R_z} (\xi - z)^2 w_{i,j} w_{i,j} dx \right)^{1/2} + \frac{2^{3/2} CK \Omega^{1/4}}{\lambda_1^{1/2} \varepsilon_3} \left[\frac{\partial^2 E_1}{\partial z^2}(0,t) \right] \left[\frac{\partial^2 E}{\partial z^2} \right]. \end{aligned} \tag{4.18}$$

Similarly, we have

$$\begin{aligned} I_{53} &= 2K\int_0^t\int_{R_z}(\xi-z)\psi_\alpha w_{\alpha,3}V dxd\eta \\ &= (KC/\lambda_1)^{1/2}|V|_{\max}\left[-\frac{\partial E}{\partial z}\right], \end{aligned} \tag{4.19}$$

$$\begin{aligned} I_{54} &= 2K\int_0^t\int_{R_z}\psi_\alpha w_{\alpha,3}dxd\eta + 2K\int_0^t\int_{R_z}(\xi-z)\psi_{\alpha,j}w_{\alpha,j}dxd\eta \\ &\le (KC/\lambda_1)^{1/2}\left[\frac{\partial^2 E}{\partial z^2}\right] + (CK)^{1/2}\left[-\frac{\partial E}{\partial z}\right]. \end{aligned} \tag{4.20}$$

Finally, we derive a bound for I_6:

$$\begin{aligned} I_6 &= -K\int_0^t\int_{R_z}(\xi-z)^2 w_{i,\eta}[w_j w_{i,j} + v_j w_{i,j} + w_j v_{i,j}]dxd\eta \\ &= -K\int_0^t\int_{R_z}(\xi-z)^2 w_{i,\eta}w_j w_{i,j}dxd\eta - K\int_0^t\int_{R_z}(\xi-z)^2 w_{i,\eta}w_{i,3}V dxd\eta \\ &\quad - K\int_0^t\int_{R_z}(\xi-z)^2 w_{3,\eta}w_\alpha V_{,\alpha}dxd\eta \\ &= I_{61}+I_{62}+I_{63}. \end{aligned} \tag{4.21}$$

With a derivation similar to (4.11), we have

$$\begin{aligned} I_{61} &\le K\int_0^t\int_z^\infty(\xi-z)^2\left(\int_{D_\xi}(w_i w_i)^2 dA\right)^{1/4}\left(\int_{D_\xi}(w_{,i\eta}w_{i,\eta})^2 dA\right)^{1/4} \\ &\qquad\cdot\left(\int_{D_\xi}w_{i,j}w_{i,j}dA\right)^{1/2}d\xi d\eta \\ &\le (2/\lambda_1)^{1/4}\Omega^{1/8}K\int_0^t\left(\int_{R_z}w_{,ij\eta}w_{,ij\eta}dx\right)^{1/2}\left(\int_{R_z}(\xi-z)^2 w_{i,j}w_{i,j}dx\right)d\eta \\ &\le (2/\lambda_1)^{1/4}\Omega^{1/8}K\max_t\left[\int_{R_z}(\xi-z)^2 w_{i,j}w_{i,j}dx\right]^{1/2}\left[\frac{\partial^2 E_3(0,t)}{\partial z^2}\right]^{1/2}[E(z,t)]^{1/2} \\ &\le \frac{\varepsilon_4}{2}\max_t\int_{R_z}(\xi-z)^2 w_{i,j}w_{i,j}dx + \frac{\Omega^{1/4}}{(2\lambda_1)^{1/2}}\,\frac{K}{\varepsilon_4}\left[\frac{\partial^2 E_3(0,t)}{\partial z^2}\right]\cdot E(z,t). \end{aligned} \tag{4.22}$$

It is easy to bound I_{62}, i.e.,

$$I_{62} \le \frac{K^{1/2}}{2}|V|_{\max}E(z,t). \tag{4.23}$$

For I_{63}, we observe that

$$\begin{aligned} I_{63} &= K\int_0^t\int_{R_z}(\xi-z)^2 w_{3,\alpha\eta}w_\alpha V dxd\eta - K\int_0^t\int_{R_z}(\xi-z)^2 w_{3,\eta}w_{3,3}V dxd\eta \\ &= K\int_{R_z}(\xi-z)^2 w_{3,\alpha}w_\alpha V dx - K\int_0^t\int_{R_z}(\xi-z)^2 w_{3,\alpha}w_{\alpha,\eta}V dxd\eta \\ &\quad - K\int_0^t\int_{R_z}(\xi-z)^2 w_{3,\alpha}w_\alpha V_{,\eta}dxd\eta - K\int_0^t\int_{R_z}(\xi-z)^2 w_{3,\eta}w_{3,3}V dxd\eta \\ &\le \frac{K|V|_{\max}}{2\lambda_1^{1/2}}\left(\int_{R_z}(\xi-z)^2 w_{i,j}w_{i,j}dx\right) + \frac{1}{2}(K/\lambda_1)^{1/2}|V_{,t}|_{\max}E(z,t) \\ &\quad + K^{1/2}|V|_{\max}E(z,t). \end{aligned} \tag{4.24}$$

We now sum up the results established in this section to obtain

$$\begin{aligned} &E(z,t) + \frac{1}{2}\int_{R_z}(\xi-z)^2 w_i w_i dx + \frac{K}{2}\left(1-\frac{|V|_{\max}}{\lambda_1^{1/2}}\right)\int_{R_z}(\xi-z)^2 w_{i,j}w_{i,j}dx \\ &\quad \le M_1 E(z,t) + M_2\left(-\frac{\partial E}{\partial z}\right) + M_3\frac{\partial^2 E}{\partial z^2} \\ &\quad\quad + \frac{1}{2}(\varepsilon_1+\varepsilon_2+\varepsilon_3+\varepsilon_4)\max_t\int_{R_z}(\xi-z)^2 w_{i,j}w_{i,j}dx, \end{aligned} \tag{4.25}$$

where

$$\begin{aligned} M_1 &= \left(\frac{1}{\lambda_1}+K^{1/2}\right)|V|_{\max} + \frac{1}{2}(K/\lambda_1)^{1/2}|V_{,t}|_{\max} \\ &\quad + \left(\frac{\Omega^{1/2}}{2\lambda_1}\right)^{1/2}\frac{K}{\varepsilon_4}\left[\frac{\partial^2 E_3(0,t)}{\partial z^2}\right], \end{aligned} \tag{4.26}$$

$$\begin{aligned} M_2 &= \left[\left(\frac{1}{\lambda_1}\right)^{1/2} + \frac{2}{\lambda_1}|V|_{\max} + \frac{C^{1/2}}{\lambda_1}|V|_{\max} + \frac{1}{\lambda_1}(C/K)^{1/2} + 2(C/\lambda_1)^{1/2}\right. \\ &\quad \left. + K^{1/2} + (KC)^{1/2}\right], \end{aligned} \tag{4.27}$$

and

$$\begin{aligned} M_3 &= \Omega^{1/4}\left[\frac{1}{(2\lambda_1^3)^{1/2}\varepsilon_1} + \frac{2^{3/2}C}{\lambda_1\varepsilon_2} + \frac{2^{3/2}CK}{\lambda_1^{1/2}\varepsilon_3}\right]\left[\frac{\partial^2 E_1}{\partial z^2}(0,t)\right] \\ &\quad + \left(\frac{CK}{\lambda_1}\right)^{1/2} + \frac{C^{1/2}}{\lambda_1}. \end{aligned} \tag{4.28}$$

Explicit bounds for $|V|_{\max}$ and $|V_{,t}|_{\max}$ as well as bounds for the total energies $E(0,t)$, $\frac{\partial^2 E_1}{\partial z^2}(0,z)$, and $\frac{\partial^2 E_3}{\partial z^2}(0,z)$ in terms of data are derived in sections 5 and 6, respectively.

Suppose now that the quantity $\int(\xi - z)^2 w_{i,j} w_{i,j} dx$ takes its maximum value at $t^* \in [0, T]$. Then, evaluating (4.25) at $t = t^*$, we have

$$\gamma \int_{R_z} (\xi - z)^2 w_{i,j} w_{i,j} dx\big|_{t=t^*}$$
$$\leq (M_1 - 1)E(z, t^*) + M_2 \left(-\frac{\partial}{\partial z} E(z, t^*)\right) + M_3 \left(\frac{\partial^2}{\partial z^2} E(z, t^*)\right), \tag{4.29}$$

where

$$\gamma = \frac{1}{2}\left[K\left(1 - \frac{|V|_{\max}}{\lambda_1^{1/2}}\right) - (\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4)\right]. \tag{4.30}$$

At this point, we require that the data be small enough to satisfy

$$|V|_{\max} < \lambda_1^{1/2}, \tag{4.31}$$

in which case the ε_i can be chosen small enough that $\gamma > 0$. We further restrict the data by requiring that

$$M_1 < 1, \tag{4.32}$$

and since $E(z, t)$ and its first and second derivatives are monotone functions of t, we have

$$\gamma \int_{R_z} (\xi - z)^2 w_{i,j} w_{i,j} dx\big|_{t=t^*} \leq M_2 \left(-\frac{\partial}{\partial z} E(z, t)\right) + M_3 \left(\frac{\partial^2}{\partial z^2} E(z, t)\right). \tag{4.33}$$

Inserting (4.33) back into (4.25), we conclude that

$$\frac{K}{2\delta}\left(1 - \frac{|V|_{\max}}{\lambda_1^{1/2}}\right)\left[M_3 \frac{\partial^2 E}{\partial z^2} + M_2 \left(-\frac{\partial E}{\partial z}\right)\right] - (1 - M_1)E \geq 0. \tag{4.34}$$

We rewrite this expression as

$$\frac{\partial^2 E}{\partial z^2} - a\frac{\partial E}{\partial z} - bE \geq 0, \tag{4.35}$$

where

$$a = M_2/M_3, \quad b = \frac{(1 - M_1)}{M_3}\left[\frac{K}{2\delta}\left(1 - \frac{|V|_{\max}}{\lambda_1^{1/2}}\right)\right]^{-1}. \tag{4.36}$$

We further rewrite (4.35) as

$$\left(\frac{\partial}{\partial z} - k_1\right)\left(\frac{\partial E}{\partial z} + k_2 E\right) \geq 0, \tag{4.37}$$

where

$$k_1 = \frac{a}{2} + \frac{1}{2}\sqrt{a^2 + 4b}, \quad k_2 = -\frac{a}{2} + \frac{1}{2}\sqrt{a^2 + 4b}. \tag{4.38}$$

It is well known that (4.37) leads to the conclusion that

$$E(z, t) \leq E(0, t)e^{-k_2 z}. \tag{4.39}$$

Inequalities (4.31) and (4.32) may be regarded as Reynold's number type restrictions for our problem.

5. Bounds for laminar flow V. In this section we derive explicit bounds for the laminar flow V and its time derivative in terms of the geometry of domain and prescribed data.

We recall that the fully developed laminar flow V satisfies the initial-boundary value problem (2.6)–(2.8), i.e.,

$$\begin{aligned} V_{,t} - \Delta V &= P(t) \quad \text{in } D_z \times \{t > 0\}, \\ V(x,0) &= 0 \quad \text{in } D_z \times \{t = 0\}, \\ V(x,t) &= 0 \quad \text{on } \partial D_z \times \{t > 0\}, \end{aligned}$$

where $P(t)$ is an unknown function but is determined from the condition

$$\int_D V(x,t)dA = \int_D f_3(x,t)dA = Q(t), \tag{5.1}$$

and $f_3(x,t)$ is the f_3 of (2.4).

From (5.1), we readily obtain

$$\frac{1}{2}\frac{\partial}{\partial t}\int_D V^2 dA + \int_D V_{,\alpha}V_{,\alpha}dA = P(t)Q(t), \tag{5.2}$$

$$\frac{1}{2}\int_D V_{,\alpha}V_{,\alpha}dA + \int_0^t \int_D (V_{,\eta})^2 dA d\eta = \int_0^t P(\eta)Q_{,\eta}(\eta)d\eta. \tag{5.3}$$

Now let ψ be a solution of the problem

$$\begin{cases} \Delta\psi = -1 & \text{in } D, \\ \psi = 0 & \text{on } \partial D. \end{cases} \tag{5.4}$$

Then we have

$$Q(t) = \int_D V dA = -\int_D V\Delta\psi dA = -\int_D V_{,t}\psi dA + SP(t), \tag{5.5}$$

where

$$S = \int_D \psi dA = \int_D \psi_{,\alpha}\psi_{,\alpha}dA > 0. \tag{5.6}$$

From (5.5) we obtain

$$P(t) = \frac{1}{S}\left[Q(t) + \int_D \psi V_{,t}dA\right]. \tag{5.7}$$

Inserting (5.7) into (5.3) leads to

$$\begin{aligned} &\frac{1}{2}\int_D V_{,\alpha}V_{,\alpha}dA + \int_0^t \int_D (V_{,\eta})^2 dA d\eta = \frac{1}{S}\int_0^t Q_{,\eta}(\eta)\left[Q(\eta) + \int_D \psi V_{,\eta}dA\right]d\eta \\ &= \frac{1}{2S}[Q^2(t) - Q^2(0)] + \frac{1}{S}\int_0^t Q_{,\eta}\left(\int_D \psi V_{,\eta}dA\right)d\eta. \end{aligned} \tag{5.8}$$

But $Q(0)=0$, since $V(x,0)=0$, so (5.8) further reduces to

$$\begin{aligned}&\frac{1}{2}\int_D V_{,\alpha}V_{,\alpha}dA+\int_0^t\int_D (V_{,\eta})^2 dAd\eta\\ &\quad\le \frac{1}{2S}Q^2(t)+\frac{1}{S}\left[\int_0^t\int_D (V_{,\eta})^2 dAd\eta\int_0^t (Q_{,\eta})^2 d\eta\int_D \psi^2 dA\right]^{1/2}.\end{aligned}\tag{5.9}$$

Using the arithmetic-geometric mean inequality in (5.9) yields, finally,

$$\begin{aligned}&\int_D V_{,\alpha}V_{,\alpha}dA+\int_0^t\int_D (V_{,\eta})^2 dAd\eta\\ &\quad\le \frac{1}{S}Q^2(t)+\frac{1}{S^2}\int_0^t (Q_{,\eta})^2 d\eta\int_D \psi^2 dA=:\overline{M}_1.\end{aligned}\tag{5.10}$$

We remark that S is a monotone function of domain as is ψ, so that bounds for S and for $\int_D \psi^2 dA$ are easily obtained (see, e.g., [6]).

By a similar derivation, using (5.7) and (5.10) in (5.2) we obtain

$$\begin{aligned}&\frac{1}{2}\int_D V^2 dA+\int_0^t\int_D V_{,\alpha}V_{,\alpha}dAd\eta=\frac{1}{S}\int_0^t Q(\eta)\left[Q(\eta)+\int_D \psi V_{,\eta}dA\right]d\eta\\ &\quad\le \frac{1}{S}\int_0^t Q^2(\eta)d\eta+\frac{1}{S}\left[\int_0^t\int_D (V_{,\eta})^2 dAd\eta\int_D \psi^2 dA\int_0^t Q^2(\eta)d\eta\right]^{1/2}\\ &\quad\le \frac{1}{S}\int_0^t Q^2(\eta)d\eta+\frac{1}{S}\left\{\left[\frac{1}{S}Q^2(t)+\frac{1}{S^2}\int_0^t Q_{,\eta}^2 d\eta\int_D \psi^2 dA\right]\right.\\ &\qquad\qquad\left.\cdot\int_D \psi^2 ds\int_0^t Q^2(\eta)d\eta\right\}^{1/2}\\ &\quad=:\overline{M}_2.\end{aligned}\tag{5.11}$$

From (5.11) it follows that

$$\int_D V^2(x,t)dA\le 2\overline{M}_2.\tag{5.12}$$

To derive a bound for $|V|_{\max}$, we observe that

$$\begin{aligned}V(x)&=-\int_D G(V_{,t}-P(t))dA\\ &=-\int_D GV_{,t}dA+P(t)\int_D GdA\\ &=-\int_D GV_{,t}dA+\psi(x)P(t),\end{aligned}\tag{5.13}$$

where G is the harmonic Green's function for D. Using Schwarz's inequality in (5.13), we obtain

$$|V|\le\left(\int_D G^2 dA\right)^{1/2}\left(\int_D V_{,t}^2 dA\right)^{1/2}+|\psi|_{\max}|P(t)|.\tag{5.14}$$

As a limiting expression of results of Weinberger [24], we have

$$\int_D G^2 dA \le \frac{|D|}{8\pi^2}, \tag{5.15}$$

where $|D|$ is the area of domain D. Using (5.9), (5.7), and the monotonicity of ψ, we obtain a bound for $|V|_{\max}$. The bound for $|V_{,t}|_{\max}$ may be found in an analogous way by substituting $V_{,t}$ for V and $P'(t)$ for $P(t)$ in the preceding arguments.

6. Bounds for the total weighted energies. In this section we sketch how one can derive the total weighted energy bounds needed to complete our decay results. Since we need it in subsequent arguments, we first show how to bound $\partial^2 E(0,t)/\partial z^2$.

Let $\widehat{u}_i$ be the solution of the associated Stokes flow problem

$$\widehat{u}_{i,t} = \Delta \widehat{u}_i + \widehat{p}_{,i} \quad \text{in } R \times \{t > 0\}, \tag{6.1}$$

$$\widehat{u}_{i,i} = 0 \quad \text{in } R \times \{t > 0\}, \tag{6.2}$$

$$\widehat{u}_i = 0 \quad \text{on } \partial D \times \{t \ge 0\}, \tag{6.3}$$

$$\widehat{u}_i = f_i(x_1, x_2, t) \quad \text{in } D_0 \times \{t > 0\}, \tag{6.4}$$

$$\widehat{u}_i = 0 \quad \text{in } R \times \{t = 0\}. \tag{6.5}$$

Suppose now that we set

$$w_i = (u_i - \widehat{u}_i) + (\widehat{u}_i - v_i) = \chi_i + \theta_i,$$
$$q_i = (p_{,i} - \widehat{p}_i) + (\widehat{p}_{,i} - P(t)\delta_{i3}) = \sigma_{,i} + \gamma_{,i}.$$

Clearly (χ_i, σ) is the solution of the initial-boundary value problem

$$\chi_{i,t} - \Delta \chi_i + u_j u_{i,j} = \sigma_{,i} \quad \text{in } R \times \{t > 0\}, \tag{6.6}$$

$$\chi_{i,i} = 0 \quad \text{in } R \times \{t > 0\}, \tag{6.7}$$

$$\chi_i = 0 \quad \text{on } \partial D \times \{t \ge 0\}, \tag{6.8}$$

$$\chi_i = 0 \quad \text{in } D_0 \times \{t > 0\}, \tag{6.9}$$

$$\chi_i = 0 \quad \text{in } R \times \{t = 0\}. \tag{6.10}$$

Furthermore, by the triangle inequality we have

$$[E(0,t)]^{1/2} \le [\widehat{E}(0,t)]^{1/2} + [\widetilde{E}(0,t)]^{1/2}, \tag{6.11}$$

where

$$\widehat{E}(0,t) = \int_0^t \int_R \xi^2 [\chi_{i,j}\chi_{i,j} + K\chi_{i,\eta}\chi_{i,\eta}] dx d\eta,$$
$$\widetilde{E}(0,t) = \int_0^t \int_R \xi^2 [\theta_{i,j}\theta_{i,j} + K\theta_{i,\eta}\theta_{i,\eta}] dx d\eta.$$

The same triangle inequality holds for E_1 and E_2 separately as well as for their z derivatives. Bounds for $-\frac{\partial \widetilde{E}(0,t)}{\partial z}$ and $\frac{\partial^2 \widetilde{E}(0,t)}{\partial z^2}$ were derived in [2, section 6]. Note, however, that the $E(0,t)$ of [2] is our $-\frac{\partial \widetilde{E}(0,t)}{\partial z}$.

We start by deriving a bound for $\frac{\partial^2 \widehat{E}_1(0,t)}{\partial z^2}$. Upon integration by parts, we obtain

$$\frac{\partial^2 \widehat{E}_1(0,t)}{\partial z^2} = -2 \int_0^t \int_R \chi_i [\chi_{i,\eta} + (\chi_j + \widehat{u}_{,j})(\chi_i + \widehat{u}_i)_{,j} - \sigma_{,i}] dx d\eta. \tag{6.12}$$

In a manner analogous to that of section 4, it is readily seen that

$$\begin{aligned}
&\frac{\partial^2 \widehat{E}_1(0,t)}{\partial z^2} + \int_R \chi_i \chi_i dx\big|_{\eta=t} = -2\int_0^t \int_R \chi_i(\chi_j + \widehat{u}_j)\widehat{u}_{i,j} dx d\eta \\
&= 2\int_0^t \int_R \chi_{i,j}[\chi_j(\theta_i + v_i) + (\theta_j + v_j)(\theta_i + v_i)] dx d\eta \\
&= 2\int_0^t \int_R \chi_{i,j}\chi_j \theta_i dx d\eta + 2\int_0^t \int_R \chi_{3,j}\chi_j V dx d\eta \\
&\quad + 2\int_0^t \int_R \chi_{i,j}\theta_i\theta_j dx d\eta + 2\int_0^t \int_R \chi_{i,3}\theta_i V dx d\eta + 2\int_0^t \int_R \chi_{3,j}\theta_j V dx d\eta \\
&\quad + 2\int_0^t \int_R \chi_{3,3} V^2 dx d\eta \\
&\le C_1 \max_t \left[\int_r \theta_{i,j}\theta_{i,j} dx\right]^{1/2} \frac{\partial^2 \widehat{E}_1(0,t)}{\partial z^2} + C_2 |V|_{\max} \frac{\partial^2 \widehat{E}_1(0,t)}{\partial z^2} \\
&\quad + C_3 \max_t \left[\int_R \theta_{i,j}\theta_{i,j} dx\right]^{1/2} \left[\frac{\partial^2 \widehat{E}_1(0,t)}{\partial z^2} \cdot \frac{\partial^2 \widetilde{E}_1(0,t)}{\partial z^2}\right]^{1/2} \\
&\quad + C_4 |V|_{\max} \left[\frac{\partial^2 \widehat{E}_1(0,t)}{\partial z^2} \cdot \frac{\partial^2 \widetilde{E}(0,t)}{\partial z^2}\right]^{1/2} \\
&\quad + C_5 |V|_{\max}^2 \left[\frac{\partial^2 \widehat{E}_1(0,t)}{\partial z^2}\right]^{1/2}
\end{aligned} \tag{6.13}$$

for computable C_i $(i = 1, 2, \ldots, 5)$. Applying the arithmetic-geometric mean inequality, (6.13) yields the bound

$$\left\{1 - C_1 \max_t \left[\int_R \theta_{i,j}\theta_{i,j} dx\right]^{1/2} - C_2 |V|_{\max} - \varepsilon\right\} \frac{\partial^2 \widehat{E}_1(0,t)}{\partial z^2} \le \text{data}, \tag{6.14}$$

provided the term in brackets is positive for some positive ε.

A bound for $\int_R \theta_{i,j}\theta_{i,j} dx$ in terms of data was derived in [2] for general t. Again, if the maximum value of $\int_R \theta_{i,j}\theta_{i,j} dx$ occurs at $t^* \in (0,t)$, we may employ the bounds of [2] for $t = t^*$. By use of Schwarz's inequality, if necessary, these data bounds may be made monotone in t and thus yield a bound for $\max_t \int_R \theta_{i,j}\theta_{i,j} dx$ that is independent of the unknown t^*. If the data terms are small enough, we can then satisfy

$$C_1 \max_t \left[\int_R \theta_{i,j}\theta_{i,j} dx\right]^{1/2} + C_2 |V|_{\max} \le 1 - \varepsilon. \tag{6.15}$$

It can be shown that if $u_{3,t}$ is continuous on ∂D at $t = 0$, the same arguments yield a bound for $\frac{\partial^2 \widehat{E}_3(0,t)}{\partial z^2}$, since we already have a bound for $|V_{,t}|_{\max}$. The key is to show that $\int_R \chi_{i,t}\chi_{i,t} dx\big|_{t=0}$ is zero. This follows since

$$\int_R \chi_{i,t}\chi_{i,t} dx\big|_{t=0} = \int_R \chi_{i,t}\sigma_{,i} dx\big|_{t=0} = -\int_{D_0} \chi_{3,t}\sigma dA = 0, \tag{6.16}$$

recalling that $\chi_i = 0$ in D_0.

We next indicate how to find a bound for $\frac{\partial^2 \widehat{E}_2(0,t)}{\partial z^2}$ and then complete our bound for $\frac{\partial^2 \widehat{E}(0,t)}{\partial z^2}$. Clearly, we have

$$\frac{\partial^2 \widehat{E}_2(0,t)}{\partial z^2} + \int_R \chi_{i,j}\chi_{i,j}dx\big|_{\eta=t} = -2\int_0^t \int_R \chi_{i,\eta}(\chi_j+\theta_j+v_j)(\chi_i+\theta_i+v_i)_{,j}dxd\eta. \tag{6.17}$$

The arguments of section 4 allow us to bound $\frac{\partial^2 \widehat{E}_2(0,t)}{\partial z^2}$ in terms of data and $\frac{\partial^2 \widehat{E}_3(0,t)}{\partial z^2}$ provided data terms are small enough. This then provides a bound for $\frac{\partial^2 \widehat{E}(0,t)}{\partial z^2}$. But the bound for $\frac{\partial^2 \widetilde{E}(0,t)}{\partial z^2}$ is known from [2]. It follows then that we have a bound for $\frac{\partial^2 E(0,t)}{\partial z^2}$. This yields a bound for $E(0,t)$ as follows. Evaluating (4.34) at $z=0$, we find

$$E(0,t) = \frac{1}{b}\,\frac{\partial^2 E(0,t)}{\partial z^2} - \frac{a}{b}\,\frac{\partial E(0,t)}{\partial z}. \tag{6.18}$$

By Schwarz's inequality, we have

$$-\frac{\partial E(0,t)}{\partial z} \le \sqrt{2}\left[E(0,t)\cdot\frac{\partial^2 E(0,t)}{\partial z^2}\right]^{1/2}. \tag{6.19}$$

Combining (6.18) and (6.19) and using the arithmetic-geometric mean inequality yields

$$E(0,t) \le 2\left[\frac{1}{b}+\frac{a^2}{b^2}\right]\frac{\partial^2 E(0,t)}{\partial z^2}. \tag{6.20}$$

Finally, with the bounds for $E(0,t)$ and $\frac{\partial^2 E_3(0,t)}{\partial z^2}$ we can make (4.39) explicit provided the data terms are sufficiently small. These data terms depend on f_i and its derivatives, the time interval, and the size of D. So our results will hold not only if the f_i and its derivatives are small enough but also for general data if the time interval or the size of the domain is sufficiently small.

REFERENCES

[1] K. A. Ames and L. E. Payne, *Decay estimates in steady pipe flow*, SIAM J. Math. Anal., 20 (1989), pp. 789–815.

[2] K. A. Ames, L. E. Payne, and P. W. Schaefer, *Spatial decay estimates in time-dependent Stokes flow*, SIAM J. Math. Anal., 24 (1993), pp. 1395–1413.

[3] C. J. Amick, *Steady solution of the Navier–Stokes equations in unbounded channels and pipes*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 4 (1977), pp. 473–513.

[4] C. J. Amick, *Properties of steady Navier–Stokes solutions for certain unbounded channels and pipes*, Nonlinear Anal., 2 (1978), pp. 689–720.

[5] I. Babuška and A. K. Aziz, *Survey lectures on the mathematical foundations of the finite element method*, in The Mathematical Foundation of the Finite Element Method with Applications to Partial Differential Equations, Academic Press, New York, 1972, pp. 5–359.

[6] C. Bandle, *Isoperimetric Inequalities and Their Applications*, Pitman Press, London, 1980.

[7] C. O. Horgan, *Plane steady flows and energy estimates for the Navier–Stokes equation*, Arch. Rational Mech. Anal., 68 (1978), pp. 359–381.

[8] C. O. Horgan, *Recent developments concerning Saint-Venant's principle: An update*, AMR, 42 (1989), pp. 295–303.

[9] C. O. Horgan, *Recent developments concerning Saint-Venant's principle: A second update*, AMR, 49 (1996), pp. 101–111.
[10] C. O. Horgan and J. K. Knowles, *Recent developments concerning Saint-Venant's principle*, Adv. in Appl. Mech., 23 (1983), pp. 179–269.
[11] C. O. Horgan and L. E. Payne, *Inequalities of Korn, Friedrichs and Babuška–Aziz*, Arch. Rational Mech. Anal., 82 (1983), pp. 165–179.
[12] C. O. Horgan and L. T. Wheeler, *Spatial decay estimates for the Navier–Stokes equations with application to the problem of entry flow*, SIAM J. Appl. Math., 35 (1978), pp. 97–116.
[13] H. A. Levine, *An estimate for the best constant in a Sobolev inequality involving three integral norms*, Ann. Mat. Pura. Appl. (4), 124 (1980), pp. 181–197.
[14] C. Lin, *Spatial decay estimates and energy bounds for the Stokes flow equation*, Stability Appl. Anal. Contin. Media, 2 (1992), pp. 249–264.
[15] C. Lin and L. E. Payne, *Spatial decay bounds in the channel flow of an incompressible viscous fluid*, Math. Models Methods Appl. Sci., 14 (2004), pp. 795–818.
[16] O. A. Oleinik, *Applications of the energy estimates analogous to Saint-Venant's principle to problems of elasticity and hydrodynamics*, Lecture Notes in Phys. 90, Springer-Verlag, Berlin, New York, 1979, pp. 422–432.
[17] L. E. Payne, *Uniqueness criteria for steady state solutions of the Navier–Stokes equations*, Simpos. Internaz. Appl. Anal. Fis. Mat. (Cagliari–Sarrari, 1964), Edizioni Cremonse, Rome, 1965, pp. 130–153.
[18] L. E. Payne, *Isoperimetric inequalities and their applications*, SIAM Rev., 9 (1967), pp. 453–488.
[19] G. Pólya and G. Szegö, *Isoperimetric inequalities in mathematical physics*, Ann. of Math. Stud. 27, Princeton University Press, Princeton, NJ, 1951.
[20] J. Serrin, *The initial-value problem for the Navier–Stokes equations*, in Nonlinear Problems (Proc. Sympos., Madison, WI), University of Wisconsin Press, Madison, WI, 1963, pp. 69–98.
[21] G. Talenti, *Best constant in Sobolev inequality*, Ann. Mat. Pura. Appl. (4), 110 (1976), pp. 353–372.
[22] J. C. Song, *Improved decay estimates in time-dependent Stokes flow*, J. Math. Anal. Appl., 288 (2003), pp. 505–517.
[23] W. Velte, *On inequalities of Friedrichs and Babuška–Aziz in dimension three*, Z. Anal. Anwendungen, 17 (1998), pp. 843–857.
[24] H. F. Weinberger, *Symmetrization in uniformly elliptic problems*, in Studies in Mathematical Analysis and Related Topics, Stanford University Press, Stanford, CA, 1962, pp. 424–428.

SIAM J. APPL. MATH.
Vol. 65, No. 2, pp. 475–493

HOMOGENIZATION THEORY AND THE ASSESSMENT OF EXTREME FIELD VALUES IN COMPOSITES WITH RANDOM MICROSTRUCTURE*

ROBERT LIPTON†

Abstract. Suitable macroscopic quantities are identified and used to assess the field distribution within a composite specimen of finite size with random microstructure. Composites made of N anisotropic dielectric materials are considered. The characteristic length scale of the microstructure relative to the length scale of the specimen is denoted by ε, and realizations of the random composite microstructure are labeled by ω. Consider any cube C_0 located inside the composite. The function $P^\varepsilon(t, C_0, \omega)$ gives the proportion of C_0 where the square of the electric field intensity exceeds t. The analysis focuses on the case when $0 < \varepsilon \ll 1$. Rigorous upper bounds on $\lim_{\varepsilon\to 0} P^\varepsilon(t, C_0, \omega)$ are found. They are given in terms of the macrofield modulation functions. The macrofield modulation functions capture the excursions of the local electric field fluctuations about the homogenized or macroscopic electric field. Information on the regularity of the macrofield modulations translates into bounds on $\lim_{\varepsilon\to 0} P^\varepsilon(t, C_0, \omega)$. Sufficient conditions are given in terms of the macrofield modulation functions that guarantee polynomial and exponential decay of $\lim_{\varepsilon\to 0} P^\varepsilon(t, C_0, \omega)$ with respect to "t." For random microstructure with oscillation on a sufficiently small scale we demonstrate that a pointwise bound on the macrofield modulation function provides a pointwise bound on the actual electric field intensity. These results are applied to assess the distribution of extreme electric field intensity for an L-shaped domain filled with a random laminar microstructure.

Key words. random composite materials, field fluctuations, material breakdown

AMS subject classifications. 35J15, 60G10, 78M40

DOI. 10.1137/S0036139903426976

1. Introduction.

Failure of composite materials can often be attributed to the presence of large local fields. This includes extreme temperature gradients and large electric and current fields as well as mechanical stresses [9]. These fields are strongly influenced by the local microgeometry inside the composite. It is often the case that the microgeometry of heterogeneous specimens is known only in a statistical sense. Motivated by these considerations, we examine the distribution of extreme field values in random heterogeneous media. The focus here is to assess the likelihood that the magnitude of the electric field inside the composite exceeds a prescribed nominal value for almost every realization of the random microstructure.

Here we consider a random composite made up of N anisotropic dielectric materials with dielectric tensors $A_1, A_2, \ldots, A_N$. To describe the dielectric tensor for a finite size sample of random composite, we begin with the description of a random medium of infinite extent. The dielectric tensor field $A(\mathbf{y}, \omega)$ associated with the composite is a function of both position $\mathbf{y}$ and geometric realization ω taken from the sample space Ω. For each realization ω, the tensor field $A(\mathbf{y}, \omega)$ is piecewise constant taking

*Received by the editors April 25, 2003; accepted for publication (in revised form) February 27, 2004; published electronically December 16, 2004. This research effort is sponsored by NSF grant DMS-0296064 and the Air Force Office of Scientific Research, Air Force Materiel Command USAF, grant F49620-02-1-0041. The U.S. government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research or the U.S. government.

http://www.siam.org/journals/siap/65-2/42697.html

†Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803 (lipton@math.lsu.edu).

only the values $A_1, A_2, \ldots, A_N$ for different points $\mathbf{y}$ in R^3. The random medium is assumed to be stationary, i.e., for any finite choice of points $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k$ and any vector $\mathbf{h}$, the distribution of the random tensor

$$A(\mathbf{y}_1 + \mathbf{h}, \omega), A(\mathbf{y}_2 + \mathbf{h}, \omega), \ldots, A(\mathbf{y}_k + \mathbf{h}, \omega) \tag{1.1}$$

does not depend on $\mathbf{h}$. The finite size composite specimen occupies the bounded domain $\mathcal{D}$, and points inside it are denoted by $\mathbf{x}$. The dielectric tensor for a composite with a random microstructure of characteristic length scale ε relative to the size of $\mathcal{D}$ is given by

$$A^\varepsilon(\mathbf{x}, \omega) = A\left(\frac{\mathbf{x}}{\varepsilon}, \omega\right). \tag{1.2}$$

The potential inside the composite is denoted by $\phi^\varepsilon(\mathbf{x}, \omega)$. For a prescribed charge distribution $f = f(\mathbf{x})$ and prescribed values of the electric potential on the boundary of the domain $\mathcal{D}$ given by $\phi^\varepsilon(\mathbf{x}, \omega) = \phi_0(\mathbf{x})$, the potential is the solution of

$$-\mathrm{div}(A^\varepsilon(\mathbf{x}, \omega)\nabla\phi^\varepsilon(\mathbf{x}, \omega)) = f \tag{1.3}$$

in $\mathcal{D}$. Here (1.3) holds in the sense of distributions. The associated electric field $\mathbf{E}^\varepsilon(\mathbf{x}, \omega) = -\nabla\phi^\varepsilon(\mathbf{x}, \omega)$ is not necessarily a stationary random field; this is due to the finite size of the domain $\mathcal{D}$ and the prescribed charge distribution.

Failure initiation criteria are often given in terms of a critical field strength such that if a significant portion of the sample has field strength above this value, then the failure process is initiated [7]. Motivated by this observation, we focus on the subset of the composite where $|\mathbf{E}^\varepsilon|^2$ exceeds the value $t > 0$, and we denote it by $S_t^\varepsilon(\omega)$. Consider any cube C_0 inside the composite. It is assumed here that the boundary of the cube does not intersect the boundary of the specimen. The field distribution function $\lambda^\varepsilon(t, C_0, \omega)$ gives the volume of the intersection of $S_t^\varepsilon(\omega)$ with C_0, i.e., $\lambda^\varepsilon(t, C_0, \omega) = |S_t^\varepsilon(\omega) \cap C_0|$. Here $|S|$ denotes the volume of the set S. Division of $\lambda^\varepsilon(t, C_0, \omega)$ by the volume of the cube gives the function $P^\varepsilon(t, C_0, \omega)$. Here $P^\varepsilon(t, C_0, \omega)$ gives the proportion of the cube experiencing field strength greater than t. One also defines the electric field distribution inside the part of the ith phase contained in the cube C_0. The volume of the set in the ith phase contained in C_0 where $|\mathbf{E}^\varepsilon|^2$ exceeds the value $t > 0$ is denoted by $\lambda_i^\varepsilon(t, C_0, \omega)$. The set occupied by the ith phase is denoted by $S_i^\varepsilon(\omega)$. Analogously $P_i^\varepsilon(t, C_0, \omega) \equiv \lambda_i^\varepsilon(t, C_0, \omega)/|S_i^\varepsilon(\omega) \cap C_0|$ gives the proportion if the ith phase contained in C_0 with field strength greater than t.

In this paper we obtain bounds on $P^\varepsilon(t, C_0, \omega)$ and $P_i^\varepsilon(t, C_0, \omega)$ in the limit of vanishing ε. These bounds are expressed in terms of suitable macroscopic quantities dubbed macrofield modulation functions. To illustrate the ideas, one applies the Chebyshev inequality to obtain the bound on $P^\varepsilon(t, C_0, \omega)$ given by

$$P^\varepsilon(t, C_0, \omega) \leq t^{-p} \frac{1}{|C_0|} \int_{C_0} |\mathbf{E}^\varepsilon(\mathbf{x}, \omega)|^{2p}\, d\mathbf{x}. \tag{1.4}$$

In section 2 we state the homogenized version of (1.4) given by

$$\lim_{\varepsilon \to 0} P^\varepsilon(t, C_0, \omega) \leq t^{-p} A_p(C_0). \tag{1.5}$$

Here $A_p(C_0)$ is independent of ω and is described in terms of the macrofield modulation functions. The macrofield modulation of order p is the L^p norm of the square

of the electric field intensity for the associated corrector problem (2.2) posed on the infinite random medium when the random medium is subjected to an imposed macroscopic electric field; see (2.9). Proposition 2.1 explicitly shows how integrability of order p at the level of the corrector problem contributes to the t^{-p} order decay of $\lim_{\varepsilon\to 0} P^{\varepsilon}(t, C_0, \omega)$. Similarly, Proposition 2.3 shows how L^{∞} regularity of the square of the electric field intensity for the associated corrector problem allows $\lim_{\varepsilon\to 0} P^{\varepsilon}(t, C_0, \omega)$ to vanish above a critical value of t. For this case we can pass to a subsequence, if necessary, to derive a pointwise bound on the local electric field intensity for almost every realization of the random microstructure when the scale of the microstructure is sufficiently small; see Proposition 2.4. When the macrofield modulation function has bounded mean oscillation, an explicit upper bound is obtained that is exponential in $-t$ and is given in terms of the BMO norm of the macrofield modulation function; see Proposition 2.5. The corrector problem that is used to define the macrofield modulation functions is well known and naturally arises in the definition of the effective dielectric tensor [1, 10, 17, 18].

It is pointed out that the main results given by Propositions 2.1 through 2.6 are strong limit theorems in that they hold for almost all realizations of the random medium. Propositions 2.1 through 2.6 are a direct consequence of the homogenization constraints given in Proposition 3.1. These constraints relate the macrofield modulation functions to the distribution of states for the square of the electric field intensity. This type of constraint is introduced in [11, 14] for the case of graded locally periodic microstructures and in the context of G convergence for multiphase linearly elastic composites. The results reported here apply to the mathematically identical situations appearing in the contexts of thermal conductivity and DC electric conductivity.

The paper is organized as follows: In section 2 the macrofield modulation functions are introduced and the main results are presented. The homogenization constraint is introduced and derived in section 3. The homogenized version of Chebyshev's inequality is established in section 4. The bounds on the support of $\lim_{\varepsilon\to 0} P_i^{\varepsilon}(t, C_0, \omega)$ and $\lim_{\varepsilon\to 0} P^{\varepsilon}(t, C_0, \omega)$ are obtained in section 5. These are given in terms of the L^{∞} norm of the macrofield modulation functions. The pointwise upper bounds are derived in section 6. The exponentially decaying bound on $\lim_{\varepsilon\to 0} P^{\varepsilon}(t, C_0, \omega)$ is derived in section 7. In section 8 we consider a highly oscillatory, randomly layered dielectric occupying an L-shaped domain. The dielectric is subjected to a prescribed charge density and the electric potential satisfies homogeneous Dirichlet boundary conditions. The macrofield modulation functions together with the results of section 2 are applied to assess the distribution of the electric field intensity inside the domain.

2. The macrofield modulation functions and main results. To introduce the macrofield modulation functions, we consider a random composite of infinite extent. For stationary random media it is shown in [17] that one can regard the dielectric tensor $A(\mathbf{y}, \omega)$ as the realization of a random function $\tilde{A}$ with respect to a three-dimensional dynamical system T acting on a suitable sample space; see also [2] for a more recent discussion. In view of this let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space. For a given partition of Ω into N measurable subsets $\Omega_1, \Omega_2, \ldots, \Omega_N$ we introduce the indicator functions $\tilde{\chi}_i$ taking the values 1 in Ω_i and zero outside and set $\tilde{A}(\omega) = \sum_{i=1}^{N} A_i \tilde{\chi}_i(\omega)$. Following [5, 10, 17] we regard the dielectric $A(\mathbf{y}, \omega)$ as a realization of $\tilde{A}$ with respect to a three-dimensional dynamical system T on Ω, i.e., $A(\mathbf{y}, \omega) = \tilde{A}(T(\mathbf{y})\omega)$ for $(\mathbf{y}, \omega)$ in $R^3 \times \Omega$. Here the family of mappings $T = T(\mathbf{y})$, $\mathbf{y}$ in R^3 from Ω into Ω, is one to one and preserves the measure $\mathcal{P}$ on Ω; i.e., for any A in $\mathcal{F}$ one has $\mathcal{P}(T(-\mathbf{y})A) = \mathcal{P}(A)$. The family of transforms is a group with $T(0)\omega = \omega$, $T(\mathbf{y} + \mathbf{h}) = T(\mathbf{y})T(\mathbf{h})$, and

for any $\mathcal{P}$ measurable function $\tilde{f}$ on Ω, the function $\tilde{f}(T(\mathbf{y})\omega)$ defined on $R^3 \times \Omega$ is also measurable with respect to $\mathcal{L} \times \mathcal{F}$, where $\mathcal{L}$ stands for the σ-algebra of Lebesgue-measurable subsets of R^3. Lastly, it is assumed that the dynamical system is ergodic.

Let $\mathbf{e}^1$, $\mathbf{e}^2$, $\mathbf{e}^3$ represent unit vectors along the coordinate directions in R^3. A constant electric field $\mathbf{e}^k$ is imposed on the infinite random medium. The dielectric response in the composite is given by an electric field that can be decomposed into the imposed electric field $\mathbf{e}^k$ and a stationary random fluctuation $-\nabla\varphi^k(\mathbf{y},\omega) = \mathbf{G}^k(T(\mathbf{y})\omega)$, where $\mathbf{G}^k$ is in $L^2(\Omega,\mathcal{P})$ with zero mean, i.e., $\langle \mathbf{G}^k \rangle = \int_\Omega \mathbf{G}^k d\mathcal{P} = 0$; see [5, 10, 17, 8]. From the Birkhoff ergodic theorem it follows that for any sequence of cubes $Q(r)$ of side length $2r$ and volume $|Q(r)|$,

$$\lim_{r\to\infty} \frac{1}{|Q(r)|} \int_{Q(r)} (-\nabla\varphi^k(\mathbf{y},\omega))\, d\mathbf{y} = \langle \mathbf{G}^k \rangle = 0. \tag{2.1}$$

The fluctuation solves

$$-\mathrm{div}(A(\mathbf{y},\omega)(\nabla\varphi^k(\mathbf{y},\omega) + \mathbf{e}^k)) = 0 \tag{2.2}$$

for $\mathbf{y}$ in R^3. For an imposed constant electric field of the general form $\overline{\mathbf{E}} = (E_1\mathbf{e}^1 + E_2\mathbf{e}^2 + E_3\mathbf{e}^3)$, the stationary random fluctuation is obtained by superposition and is given by $-\nabla\varphi(\mathbf{y},\omega) = \sum_{k=1}^3 E_k \mathbf{G}^k(T(\mathbf{y})\omega)$. For future reference we introduce the matrix with column vectors $\mathbf{G}^k$ given by $\underline{\tilde{\mathbf{G}}}(\omega) = (\mathbf{G}^1(\omega), \mathbf{G}^2(\omega), \mathbf{G}^3(\omega))$. Then $-\nabla\varphi(\mathbf{y},\omega) = \underline{\tilde{\mathbf{G}}}(T(\mathbf{y})\omega)\overline{\mathbf{E}}$ and $\mathbf{E}(\mathbf{y},\omega) = (I + \underline{\tilde{\mathbf{G}}}(T(\mathbf{y})\omega))\overline{\mathbf{E}}$. The dielectric displacement is a stationary random field, and its mean is given by

$$\langle \mathbf{D} \rangle = \int_\Omega \tilde{A}(\omega)(I + \underline{\tilde{\mathbf{G}}}(\omega))\overline{\mathbf{E}}\, d\mathcal{P}(\omega) = \lim_{r\to\infty} \frac{1}{|Q(r)|} \int_{Q(r)} A(\mathbf{y},\omega)\mathbf{E}(\mathbf{y},\omega)\, dy. \tag{2.3}$$

The effective dielectric tensor A^E provides the linear relation between the imposed electric field $\overline{\mathbf{E}}$ and the mean dielectric displacement $\langle \mathbf{D} \rangle$, i.e., $\langle \mathbf{D} \rangle = A^E\overline{\mathbf{E}}$; see [5, 10, 17, 8].

When considering failure initiation it is important to assess the magnitude of the local electric field inside the random medium arising from the imposed electric field $\overline{\mathbf{E}}$. Here one is interested in the probability that the square of the electric field intensity $|\mathbf{E}|^2$ in the ith phase exceeds a nominal value t. For the stationary random case this probability is the same for every point and is given by $\theta_{t,i} = \mathcal{P}(\tilde{\chi}_i(\omega)|(I + \underline{\tilde{\mathbf{G}}}(\omega))\overline{\mathbf{E}}|^2 > t)$. Other quantities that are useful for local field assessment are given by the L^p norms, $1 \le p \le \infty$. The $L^p(\Omega)$ norm of a $\mathcal{P}$ measurable function $\tilde{g}$ is denoted by $\|\tilde{g}\|_{L^p(\Omega)}$. Since $T(\mathbf{y})$ preserves the measure $\mathcal{P}$ on Ω, it follows that

$$\|\chi_i(\mathbf{y},\omega)|\mathbf{E}(\mathbf{y},\omega)|^2\|_{L^p(\Omega)} = \|\tilde{\chi}_i(\omega)|(I + \underline{\tilde{\mathbf{G}}}(\omega))\overline{\mathbf{E}}|^2\|_{L^p(\Omega)} \quad \text{for every } \mathbf{y} \text{ in } R^3. \tag{2.4}$$

Motivated by these considerations, we introduce moments of the local electric field of order p.

Definition: Moments of the local electric field.

$$f_p^i(\overline{\mathbf{E}}) = \|\tilde{\chi}_i(\omega)|(I + \underline{\tilde{\mathbf{G}}}(\omega))\overline{\mathbf{E}}|^2\|_{L^p(\Omega)} = \left(\int_\Omega \tilde{\chi}_i(\omega)|(I + \underline{\tilde{\mathbf{G}}}(\omega))\overline{\mathbf{E}}|^{2p}\, d\mathcal{P}(\omega) \right)^{1/p} \tag{2.5}$$

for $1 \le p \le \infty$.

Moments of the electric field have been calculated for two-dimensional random dispersions of disk-, needle-, and square-shaped inclusions in [4].

It is pointed out that the electric field generated by a constant imposed electric field is self-similar under a rescaling of the infinite random medium. Indeed, set $\varepsilon_k = 1/k$ and rescale the material properties by $A^{\varepsilon_k}(\mathbf{y},\omega) = A(\mathbf{y}/\varepsilon_k,\omega)$. It is easily checked that the electric field also rescales as $\mathbf{E}^{\varepsilon_k}(\mathbf{y},\omega) = \mathbf{E}(\mathbf{y}/\varepsilon_k,\omega)$. Thus the analysis of electric field distribution for the ε_k scale microstructure reduces to an analysis for the unrescaled random media. However, this symmetry is broken for generic situations when the specimen is finite in extent and the loading is not uniform throughout the sample. Because of this the electric field in the composite is not obtained directly through an analysis of the electric field in an infinite random medium. Instead, it is shown here that a suitable multiscale analysis using macrofield modulation functions provides rigorous bounds on the field distributions $P^{\varepsilon}(t,C_0,\omega)$ and $P_i^{\varepsilon}(t,C_0,\omega)$ for almost every realization in the limit of vanishing ε.

Consider a finite size specimen $\mathcal{D}$ filled with random composite with characteristic length scale $\varepsilon_k = 1/k$. Here the composite is described by $A^{\varepsilon_k}(\mathbf{x},\omega) = \tilde{A}(T(\mathbf{x}/\varepsilon_k)\omega)$, and the electric potential $\phi^{\varepsilon_k}(\mathbf{x},\omega)$ solves the boundary value problem described in the introduction with equilibrium condition given by (1.3). The electric field is given by $\mathbf{E}^{\varepsilon_k} = -\nabla(\phi^{\varepsilon_k})$. The multiscale analysis proceeds in two steps. The first step is the up scaling or homogenization step where the macroscopic electric field is determined. From the theory of random homogenization, the fields $\mathbf{E}^{\varepsilon_k}(\mathbf{x},\omega)$ and $\mathbf{D}^{\varepsilon_k}(\mathbf{x},\omega) = \mathbf{A}^{\varepsilon_k}(\mathbf{x},\omega)\mathbf{E}^{\varepsilon_k}(\mathbf{x},\omega)$ converge to the deterministic macroscopic fields $\mathbf{E}(\mathbf{x})^M$ and $\mathbf{D}^M(x)$ as ε_k goes to zero for almost every ω; see [10, 17]. Here the convergence of the sequences of electric and displacement fields is given by weak convergence in $L^2(\mathcal{D})^3$. The deterministic macroscopic potential $\phi^M(\mathbf{x})$ satisfies the boundary condition $\phi^M(\mathbf{x}) = \phi_0(\mathbf{x})$. The macroscopic dielectric displacement satisfies the equilibrium equation

$$\text{div}\mathbf{D}^M = f \tag{2.6}$$

and $\mathbf{E}^M = -\nabla\phi^M$. The displacement and electric field are related through the homogenized constitutive law

$$\mathbf{D}^M(\mathbf{x}) = A^E\mathbf{E}^M(\mathbf{x}). \tag{2.7}$$

The second step is a down scaling step and gives the interaction between the macroscopic electric field $\mathbf{E}^M(\mathbf{x})$ and the microstructure. For each $\mathbf{x}$, the microscopic dielectric response is given by

$$\mathbf{E}(\mathbf{x},\mathbf{y},\omega) = (I + \underline{\tilde{\mathbf{G}}}(T(\mathbf{y})\omega))\mathbf{E}^M(\mathbf{x}). \tag{2.8}$$

The relevant interaction is described by the macrofield modulation function $f_p^i(\mathbf{E}^M(\mathbf{x}))$ given by the following definition.

Definition: Macrofield modulation function.

$$f_p^i(\mathbf{E}^M(\mathbf{x})) = \|\tilde{\chi}_i(\omega)|(I + \underline{\tilde{\mathbf{G}}}(\omega))\mathbf{E}^M(\mathbf{x})|^2\|_{L^p(\Omega)} \tag{2.9}$$

for $1 \le p \le \infty$. The macrofield modulation function $f_p^i(\mathbf{E}^M(\mathbf{x}))$ provides a measure of the amplification or diminution of $\mathbf{E}^M(\mathbf{x})$ by the random medium. Explicit formulas for the macrofield modulation functions for randomly layered two-phase dielectrics are given in section 8.

Consider any cube C_0 inside the composite. The L^1 norm of a function $g(\mathbf{x})$ over the cube C_0 is denoted by $\|g\|_{L^1(C_0)}$. In what follows, it is always assumed that

$\theta_i = \int_\Omega \tilde{\chi}_i(\omega)\, d\mathcal{P} > 0$, and from ergodicity the volume occupied by the ith phase in the cube C_0 tends to the nonzero limit $\lim_{\varepsilon_k \to 0} \int_{C_0} \tilde{\chi}_i(T(\mathbf{x}/\varepsilon_k)\omega)\, d\mathbf{x} = \theta_i |C_0|$ as ε_k tends to zero. Passing to a subsequence, if necessary, we consider $\lim_{\varepsilon_k \to 0} P_i^{\varepsilon_k}(t, C_0, \omega)$.

If it is known that $\| |f_p^i(\mathbf{E}^M(\mathbf{x}))|^p \|_{L^1(C_0)} < \infty$ for some p, then the following proposition shows that $\lim_{\varepsilon_k \to 0} P_i^{\varepsilon_k}(t, C_0, \omega)$ decays on the order of t^{-p}.

PROPOSITION 2.1 (homogenization of Chebyshev's inequality). *Given that*

$$\| |f_p^i(\mathbf{E}^M(\mathbf{x}))|^p \|_{L^1(C_0)} < \infty$$

for some p with $1 \le p < \infty$, then for almost every realization ω one has

$$\begin{aligned} \lim_{\varepsilon_k \to 0} P_i^{\varepsilon_k}(t, C_0, \omega) &\le t^{-p} \frac{1}{\theta_i |C_0|} \| |f_p^i(\mathbf{E}^M(\mathbf{x}))|^p \|_{L^1(C_0)} \\ &= t^{-p} \frac{1}{\theta_i |C_0|} \int_{C_0} \int_\Omega \tilde{\chi}_i(\omega) |(I + \underline{\tilde{\mathbf{G}}}(\omega)) \mathbf{E}^M(\mathbf{x})|^{2p}\, d\mathcal{P}(\omega)\, d\mathbf{x}. \end{aligned} \tag{2.10}$$

If $\| |f_p^i(\mathbf{E}^M(\mathbf{x}))|^p \|_{L^1(C_0)} < \infty$ for all $i = 1, 2, \ldots, N$, then

$$\lim_{\varepsilon_k \to 0} P^{\varepsilon_k}(t, C_0, \omega) \le t^{-p} \frac{1}{|C_0|} \int_{C_0} \int_\Omega |(I + \underline{\tilde{\mathbf{G}}}(\omega)) \mathbf{E}^M(\mathbf{x})|^{2p}\, d\mathcal{P}(\omega)\, d\mathbf{x} \tag{2.11}$$

for almost every realization ω.

It is clear that the coefficients of t^{-p} in (2.10) and (2.11) depend upon the Dirichlet data ϕ_0, charge density f, and the domain $\mathcal{D}$ through the solution of the homogenized problem (2.6). The proof of Proposition 2.1 is given in section 4.

The L^∞ norm of a function $g(\mathbf{x})$ over the cube C_0 is denoted by $\|g\|_{L^\infty(C_0)}$. A characterization of the set of parameters t where $\lim_{\varepsilon_k \to 0} P_i^{\varepsilon_k}(t, C_0, \omega)$ vanishes for almost every realization is given in the following proposition.

PROPOSITION 2.2. *If $t > \|f_\infty^i(\mathbf{E}^M(\mathbf{x}))\|_{L^\infty(C_0)}$, then $\lim_{\varepsilon_k \to 0} P_i^{\varepsilon_k}(t, C_0, \omega) = 0$ for almost every ω in Ω*

From the proposition it is evident that if $t > \|f_\infty^i(\mathbf{E}^M(\mathbf{x}))\|_{L^\infty(C_0)}$, then the volume of the subsets in the ith phase for which $|\mathbf{E}^{\varepsilon_k}(\mathbf{x}, \omega)|^2 > t$ vanishes as ε_k tends to zero with probability one. The proof of Proposition 2.2 is given in section 5.

We introduce the macrostress modulation $M(\mathbf{E}^M(\mathbf{x}))$ given by

$$M(\mathbf{E}^M(\mathbf{x})) = \max_{i=1,\ldots,N} f_\infty^i(\mathbf{E}^M(\mathbf{x})) \tag{2.12}$$

and characterize $\lim_{\varepsilon_k \to 0} P^{\varepsilon_k}(t, C_0, \omega)$ in a way analogous to Proposition 2.2. This is stated in the following proposition.

PROPOSITION 2.3. *If $t > \|M(\mathbf{E}^M(\mathbf{x}))\|_{L^\infty(C_0)}$, then $\lim_{\varepsilon_k \to 0} P^{\varepsilon_k}(t, C_0, \omega) = 0$ for almost every realization.*

For random microstructure with oscillation on a sufficiently small scale, it is found that a pointwise bound on the macrofield modulation function delivers a pointwise bound on the actual electric field intensity for almost every realization of the microstructure.

PROPOSITION 2.4 (pointwise bounds on the electric field intensity). *Suppose that*

$$t > M(\mathbf{E}^M(\mathbf{x})) \tag{2.13}$$

on C_0. Then one can pass to a subsequence $\{\varepsilon_{k'}\}_{k'=1}^\infty$ if necessary to find that there is a critical ε_0 such that for every $\varepsilon_{k'} < \varepsilon_0$,

$$|\mathbf{E}^{\varepsilon_{k'}}(\mathbf{x}, \omega)|^2 \le t \tag{2.14}$$

for almost every $\mathbf{x}$ *in* C_0 *and for almost every realization* ω. *Here* ε_0 *can depend upon* $\mathbf{x}$ *and* ω.

The proof of Proposition 2.4 is given in section 6.

Last, we give conditions for which $\lim_{\varepsilon_k \to 0} P^{\varepsilon_k}(t, C_0, \omega)$ decreases exponentially with t. To do this we introduce the BMO norm of $M(\mathbf{E}^M(\mathbf{x}))$ over the cube C_0 given by

$$\|M\|_{BMO} = \sup_{C \subset C_0} \left(\frac{1}{|C|} \int_C |M(\mathbf{E}^M(\mathbf{x})) - M_C| \, d\mathbf{x} \right), \tag{2.15}$$

where M_C is the average of $M(\mathbf{E}^M(\mathbf{x}))$ over C and the supremum is taken over all subcubes C of C_0. The BMO norm and the space of functions of bounded mean oscillation were introduced by John and Nirenberg [6]. The space of functions with bounded L^∞ norm are a subspace of the functions with bounded BMO norm since $\|M(\mathbf{E}^M)\|_{BMO} \le c\|M(\mathbf{E}^M)\|_{L^\infty(C_0)}$, where c is a constant depending on C_0.

For any positive number α between zero and one, we define the constant $C(\alpha)$ by

$$C(\alpha) = \frac{\alpha |\ln \alpha|}{8\|M\|_{BMO}}. \tag{2.16}$$

With the average of $M(\mathbf{E}^M(\mathbf{x}))$ over the cube C_0 denoted by M_{C_0}, the bound on $\lim_{\varepsilon_k \to 0} P^{\varepsilon_k}(t, C_0, \omega)$ is given in the following proposition.

PROPOSITION 2.5. *If* $t > 8\|M\|_{BMO}\alpha^{-1} + M_{C_0}$, *then*

$$\lim_{\varepsilon_k \to 0} P^{\varepsilon_k}(t, C_0, \omega) \le \alpha^{-1} e^{-C(\alpha) \times (t - M_{C_0})} \tag{2.17}$$

for almost every realization.

For t fixed the proposition shows that $P^{\varepsilon_k}(t, C_0, \omega)$ approaches or drops below

$$\alpha^{-1} e^{-C(\alpha) \times (t - M_{C_0})}$$

for ε_k sufficiently small for almost every realization. It also shows that the upper bound is exponentially decreasing for large t. Optimization over α (see section 7) provides the tighter upper bound given by the following proposition.

PROPOSITION 2.6. *If* $t > 8\|M\|_{BMO} + M_{C_0}$, *then for almost every realization of the random medium*

$$\lim_{\varepsilon_k \to 0} P^{\varepsilon_k}(t, C_0, \omega) \le (\alpha(t))^{-1} e \times e^{[-\alpha(t)(t - M_{C_0})/(8\|M\|_{BMO})]}, \tag{2.18}$$

where the factor $\alpha(t)$ *lies in the interval* $e^{-1} < \alpha(t) < 1$ *and is the root of the equation*

$$\kappa^{-1} - \alpha(1 + \ln \alpha) = 0, \tag{2.19}$$

with $\kappa = (t - M_{C_0})/(8\|M\|_{BMO})$.

It is pointed out that if the macroscopic electric field $\mathbf{E}^M$ is constant inside C_0, then $\|M\|_{BMO} = 0$, $M_{C_0} = M(\mathbf{E}^M) = \|M(\mathbf{E}^M)\|_{L^\infty(C_0)}$, and Propositions 2.3, 2.5, and 2.6 reduce to the observation that if $t > M(\mathbf{E}^M)$, then $\lim_{\varepsilon_k \to 0} P^{\varepsilon_k}(t, C_0, \omega) = 0$ for almost all ω.

Propositions 2.1 through 2.6 provide the opportunity to recover information on the behavior of the electric field intensity $|\mathbf{E}^\varepsilon(\mathbf{x}, \omega)|$ inside the random microstructure from knowledge of the behavior of the macrofield modulation functions. An application is given in section 8 where the electric field distribution inside an L-shaped domain containing a highly oscillatory random laminate is analyzed.

3. Homogenization constraints. The homogenization constraints are motivated by considering the case of a random composite of infinite extent. For the $p = \infty$ case the homogenization constraint follows immediately from the definition of $f^i_\infty(\overline{\mathbf{E}})$. Indeed, it is clear from the definition of the L^∞ norm that $t \geq f^i_\infty(\overline{\mathbf{E}})$ implies that $\theta_{t,i} = 0$, and equivalently, if $\theta_{t,i} > 0$, it follows that $f^i_\infty(\overline{\mathbf{E}}) > t$. This delivers the homogenization constraints given by

$$\theta_{t,i}(f^i_\infty(\overline{\mathbf{E}}) - t) \geq 0. \tag{3.1}$$

For $1 \leq p < \infty$, Chebyshev's inequality implies

$$t^{-p}(f^i_p(\overline{\mathbf{E}}))^p \geq \theta_{t,i}. \tag{3.2}$$

Inequalities (3.1) and (3.2) are the specialization of the homogenization constraints to stationary random composites of infinite extent. In the general context the macroscopic electric field is not uniform and the composite specimen has finite size. For general specimen shapes and nonuniform loading, the constraints analogous to (3.1) and (3.2) are given in terms of $f^i_\infty(\mathbf{E}^M(\mathbf{x}))$ and $f^i_p(\mathbf{E}^M(\mathbf{x}))$. In order to complete the description of the homogenization constraint, a suitable generalization of $\theta_{t,i}$ is needed. For this case, one considers a realization of the random composite $A^{\varepsilon_k}(\mathbf{x},\omega)$ and the set in the ith phase where the square of the electric field intensity $|\mathbf{E}^{\varepsilon_k}(\mathbf{x},\omega)|^2$ exceeds t is denoted by $S^{\varepsilon_k}_{t,i}(\omega)$. Consider any subdomain Q of the specimen such that the boundary of Q does not intersect the boundary of the specimen. The distribution function $\lambda^{\varepsilon_k}_i(t,Q,\omega)$ is defined by $\lambda^{\varepsilon_k}_i(t,Q,\omega) = |S^{\varepsilon_k}_{t,i}(\omega) \cap Q|$. The indicator function for the set $S^{\varepsilon_k}_{t,i}(\omega)$ is written $\chi^{\varepsilon_k}_{t,i}(\mathbf{x},\omega)$ taking the value 1 in $S^{\varepsilon_k}_{t,i}(\omega)$ and 0 outside and we write $\lambda^{\varepsilon_k}_i(t,Q,\omega) = \int_Q \chi^{\varepsilon_k}_{t,i} d\mathbf{x}$. From the theory of weak convergence there exists a (Lebesgue measurable) density $\theta_{t,i}(\mathbf{x},\omega)$ taking values in the interval $[0,1]$ such that (on passage to a subsequence if necessary) $\lim_{k\to\infty} \lambda^{\varepsilon_k}_i(t,Q,\omega) = \int_Q \theta_{t,i}(\mathbf{x},\omega)dx$. The density $\theta_{t,i}(\mathbf{x},\omega)$ is the local distribution of states of the square of the electric field intensity $|\mathbf{E}^{\varepsilon_k}(\mathbf{x},\omega)|^2$ in the ith phase as ε_k goes to zero. Here, the random fields $\mathbf{E}^{\varepsilon_k}(\mathbf{x},\omega)$ and $\theta_{t,i}(\mathbf{x},\omega)$ can no longer be regarded as stationary; this is due to the finite size of the domain and nonuniform charge distribution within the dielectric. However, for almost every realization one has the homogenization constraints given in the following proposition.

PROPOSITION 3.1 (homogenization constraints). *For almost every point* $\mathbf{x}$ *in* Q *and almost every realization* ω *in* Ω*, one has*

$$\theta_{t,i}(\mathbf{x},\omega)(f^i_\infty(\mathbf{E}^M(\mathbf{x})) - t) \geq 0, \quad i = 1,\ldots,N, \tag{3.3}$$

and for $1/q + 1/p = 1$*,*

$$\theta^{1/q}_{t,i}(\mathbf{x},\omega) f^i_p(\mathbf{E}^M(\mathbf{x})) \geq t\theta_{t,i}(\mathbf{x},\omega), \quad i = 1,\ldots,N. \tag{3.4}$$

It is clear that (3.3) and (3.4) are the extensions of (3.1) and (3.2) to situations where the macroscopic electric field is no longer uniform.

Proof. For a given realization ω, it follows from the definition of the set $S^{\varepsilon_k}_{t,i}(\omega)$ that

$$\chi^{\varepsilon_k}_{t,i}(\mathbf{x},\omega)|\mathbf{E}^{\varepsilon_k}(\mathbf{x},\omega)|^2 - t\chi^{\varepsilon_k}_{t,i}(\mathbf{x},\omega) > 0. \tag{3.5}$$

Multiplying (3.5) by any nonnegative test function $p(\mathbf{x})$ and integrating over $\mathcal{D}$ gives

$$\int_{\mathcal{D}} p(\mathbf{x})(\chi^{\varepsilon_k}_{t,i}(\mathbf{x},\omega)|\mathbf{E}^{\varepsilon_k}(\mathbf{x},\omega)|^2 - t\chi^{\varepsilon_k}_{t,i}(\mathbf{x},\omega))\, d\mathbf{x} > 0. \tag{3.6}$$

Taking limits and passing to subsequences if necessary gives

$$\lim_{\varepsilon_k \to 0} \int_{\mathcal{D}} p(\mathbf{x}) \chi_{t,i}^{\varepsilon_k}(\mathbf{x}, \omega) |\mathbf{E}^{\varepsilon_k}(\mathbf{x}, \omega)|^2 \, d\mathbf{x} \geq t \int_{\mathcal{D}} p(\mathbf{x}) \theta_{t,i}(\mathbf{x}, \omega) \, d\mathbf{x}. \tag{3.7}$$

We will use the following lemma.

LEMMA 3.2.

$$\int_{\mathcal{D}} p(\mathbf{x}) f_\infty^i(\mathbf{E}^M(\mathbf{x})) \theta_{t,i}(\mathbf{x}, \omega) \, d\mathbf{x} \geq \lim_{\varepsilon_k \to 0} \int_{\mathcal{D}} p(\mathbf{x}) \chi_{t,i}^{\varepsilon_k}(\mathbf{x}, \omega) |\mathbf{E}^{\varepsilon_k}(\mathbf{x}, \omega)|^2 \, d\mathbf{x}, \tag{3.8}$$

and for $1/q + 1/p = 1$,

$$\int_{\mathcal{D}} p(\mathbf{x}) f_p^i(\mathbf{E}^M(\mathbf{x})) \theta_{t,i}^{1/q}(\mathbf{x}, \omega) \, d\mathbf{x} \geq \lim_{\varepsilon_k \to 0} \int_{\mathcal{D}} p(\mathbf{x}) \chi_{t,i}^{\varepsilon_k}(\mathbf{x}, \omega) |\mathbf{E}^{\varepsilon_k}(\mathbf{x}, \omega)|^2 \, d\mathbf{x} \tag{3.9}$$

for all nonnegative $p(\mathbf{x})$ *in* $C_0^\infty(\mathcal{D})$ *and for almost every* ω.

Applying the inequality (3.7) together with Lemma 3.2 delivers

$$\int_{\mathcal{D}} p(\mathbf{x}) f_\infty^i(\mathbf{E}^M(\mathbf{x})) \theta_{t,i}(\mathbf{x}, \omega) \, d\mathbf{x} \geq t \int_{\mathcal{D}} p(\mathbf{x}) \theta_{t,i}(\mathbf{x}, \omega) \, d\mathbf{x} \tag{3.10}$$

and

$$\int_{\mathcal{D}} p(\mathbf{x}) f_p^i(\mathbf{E}^M(\mathbf{x})) \theta_{t,i}^{1/q}(\mathbf{x}, \omega) \, d\mathbf{x} \geq t \int_{\mathcal{D}} p(\mathbf{x}) \theta_{t,i}(\mathbf{x}, \omega) \, d\mathbf{x} \tag{3.11}$$

for almost every ω. The proposition now follows since (3.10) and (3.11) hold for every nonnegative test function. □

Proof of Lemma 3.2. We write

$$A^{\varepsilon_k}(\mathbf{x}, \omega) = A^{\varepsilon_k}(A_1, A_2, \ldots, A_N, \mathbf{x}, \omega) = \sum_{\ell=1}^{N} \tilde{\chi}_\ell(T(\mathbf{x}/\varepsilon_k)\omega) A_\ell. \tag{3.12}$$

We introduce the $N+1$ phase composite identical to the previous except that in $S_{t,i}^{\varepsilon_k}(\omega)$ it has dielectric constant P_{N+1}. The piecewise constant dielectric tensor for this composite is given by

$$\begin{aligned} \hat{A}^{\varepsilon_k}(\mathbf{x}, \omega) &= \hat{A}^{\varepsilon_k}(A_1, A_2, \ldots, A_N, P_{N+1}, \mathbf{x}, \omega) \\ &= \sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} \tilde{\chi}_\ell(T(\mathbf{x}/\varepsilon_k)\omega) A_\ell \\ &\quad + \tilde{\chi}_i(T(\mathbf{x}/\varepsilon_k)\omega)(1 - \chi_{t,i}^{\varepsilon_k}(\mathbf{x}, \omega)) A_i + \tilde{\chi}_i(T(\mathbf{x}/\varepsilon_k)\omega) \chi_{t,i}^{\varepsilon_k}(\mathbf{x}, \omega) P_{N+1}. \end{aligned} \tag{3.13}$$

For P_{N+1} in a neighborhood of A_i, we invoke the compactness property of G-convergence with respect to the sequence $\{\hat{A}^{\varepsilon_k}(A_1, A_2, \ldots, A_N, P_{N+1}, \mathbf{x}, \omega)\}_{k=1}^\infty$ [19, 16] to assert the existence of a G-converging subsequence also denoted by

$$\{\hat{A}^{\varepsilon_k}(A_1, A_2, \ldots, A_N, P_{N+1}, \mathbf{x}, \omega)\}_{k=1}^\infty$$

and a G-limit denoted by $\hat{A}^E(A_1, A_2, \ldots, A_N, P_{N+1}, \mathbf{x}, \omega)$. The partial derivatives of $\hat{A}^E(A_1, A_2, \ldots, A_N, P_{N+1}, \mathbf{x}, \omega)$ with respect to each element of P_{N+1} evaluated at $P_{N+1} = A_i$ are given by [11, 12, 13]:

$$\begin{aligned} &\nabla_{mn}^{N+1} \hat{A}_{op}^E(A_1, A_2, \ldots, A_N, A_i, \mathbf{x}, \omega) \\ &\quad = \lim_{r \to 0} \lim_{\varepsilon_k \to 0} \left(\frac{1}{|Q(\mathbf{x}, r)|} \int_{Q(\mathbf{x}, r)} \chi_{t,i}^{\varepsilon_k}(\mathbf{y}, \omega)(\partial_m w_o^{k,r} + \mathbf{e}_m^o)(\partial_n w_p^{k,r} + \mathbf{e}_n^p) \, d\mathbf{y} \right). \end{aligned} \tag{3.14}$$

Here $Q(\mathbf{x},r)$ is a cube of side length $2r$ inside $\mathcal{D}$ centered at $\mathbf{x}$ with volume given by $|Q(\mathbf{x},r)|$, and the functions $w_p^{k,r}$ vanish on the boundary of the cube and are the solutions of

$$-\mathrm{div}(A^{\varepsilon_k}(\mathbf{y},\omega)(\nabla w_p^{k,r}(\mathbf{y})+\mathbf{e}^p))=0,\quad p=1,2,3, \tag{3.15}$$

for $\mathbf{y}$ in $Q(\mathbf{x},r)$. From [11, 12, 13] one has for every test function p vanishing on the boundary of $\mathcal{D}$ that

$$\begin{aligned}&\lim_{\varepsilon_k\to 0}\int_{\mathcal{D}} p(\mathbf{x})\chi_{t,i}^{\varepsilon_k}(\mathbf{x},\omega)|\mathbf{E}^{\varepsilon_k}(\mathbf{x},\omega)|^2\,d\mathbf{x}\\ &\quad=\int_{\mathcal{D}} p(\mathbf{x})\left(\sum_{m=1}^{3}\nabla_{mm}^{N+1}\hat{A}^E(A_1,A_2,\ldots,A_N,A_i,\mathbf{x},\omega)\right)\mathbf{E}^M(\mathbf{x})\cdot\mathbf{E}^M(\mathbf{x})\,d\mathbf{x}.\end{aligned} \tag{3.16}$$

Here

$$\begin{aligned}&\left(\sum_{m=1}^{3}\nabla_{mm}^{N+1}\hat{A}^E(A_1,A_2,\ldots,A_N,A_i,\mathbf{x},\omega)\right)\mathbf{E}^M(\mathbf{x})\cdot\mathbf{E}^M(\mathbf{x})\\ &\quad=\sum_{op}\left(\lim_{r\to 0}\lim_{\varepsilon_k\to 0}\left(\frac{1}{|Q(\mathbf{x},r)|}\int_{Q(\mathbf{x},r)}\chi_{t,i}^{\varepsilon_k}(\mathbf{y},\omega)(\nabla w_o^{k,r}+\mathbf{e}^o)\right.\right.\\ &\qquad\left.\left.\cdot(\nabla w_p^{k,r}+\mathbf{e}^p)\,d\mathbf{y}\right)\mathbf{E}_o^M(\mathbf{x})\mathbf{E}_p^M(\mathbf{x})\right).\end{aligned} \tag{3.17}$$

From the appendix of [5] it follows, on passing to a subsequence, if necessary, that for every $r>0$

$$\lim_{\varepsilon_k\to 0}\int_{Q(\mathbf{x},r)}|(-\nabla w_p^{k,r}(\mathbf{y}))-\mathbf{G}^p(T(\mathbf{y}/\varepsilon_k)\omega)|^2\,d\mathbf{y}=0 \tag{3.18}$$

for almost every ω. From this we deduce that for a denumerable sequence $\{r_j\}_{j=1}^{\infty}$, $r_j\to 0$

$$\begin{aligned}&\left(\sum_{m=1}^{3}\nabla_{mm}^{N+1}\hat{A}^E(A_1,A_2,\ldots,A_N,A_i,\mathbf{x},\omega)\right)\mathbf{E}^M(\mathbf{x})\cdot\mathbf{E}^M(\mathbf{x})\\ &\quad=\lim_{r_j\to 0}\lim_{\varepsilon_k\to 0}\left(\frac{1}{|Q(\mathbf{x},r_j)|}\int_{Q(\mathbf{x},r_j)}\chi_{t,i}^{\varepsilon_k}(\mathbf{y},\omega)|(I+\underline{\tilde{\mathbf{G}}}(T(\mathbf{y}/\varepsilon_k)\omega))\mathbf{E}^M(\mathbf{x})|^2\,d\mathbf{y}\right)\end{aligned} \tag{3.19}$$

for almost every ω. Applying the Hölder inequality gives

$$\begin{aligned}&\int_{Q(\mathbf{x},r_j)}\chi_{t,i}^{\varepsilon_k}(\mathbf{y},\omega)|(I+\underline{\tilde{\mathbf{G}}}(T(\mathbf{y}/\varepsilon_k)\omega))\mathbf{E}^M(\mathbf{x})|^2\,d\mathbf{y}\\ &\quad\le\int_{Q(\mathbf{x},r_j)}\chi_{t,i}^{\varepsilon_k}(\mathbf{y},\omega)\,d\mathbf{y}\|\tilde{\chi}_i(T(\mathbf{y})\omega)|(I+\underline{\tilde{\mathbf{G}}}(T(\mathbf{y})\omega))\mathbf{E}^M(\mathbf{x})|^2\|_{L^\infty(R^3)}\\ &\quad\le\int_{Q(\mathbf{x},r_j)}\chi_{t,i}^{\varepsilon_k}(\mathbf{y},\omega)\,d\mathbf{y}\|\tilde{\chi}_i(\omega)|(I+\underline{\tilde{\mathbf{G}}}(\omega))\mathbf{E}^M(\mathbf{x})|^2\|_{L^\infty(\Omega)}.\end{aligned} \tag{3.20}$$

The last inequality in (3.20) follows from a straightforward argument given in the appendix. Noting that

$$\lim_{r_j\to 0}\lim_{\varepsilon_k\to 0}\frac{1}{|Q(\mathbf{x},r_j)|}\int_{Q(\mathbf{x},r_j)}\chi_{t,i}^{\varepsilon_k}(\mathbf{y},\omega)\,d\mathbf{y}=\theta_{t,i}(\mathbf{x},\omega) \tag{3.21}$$

for almost all $\mathbf{x}$ and applying (3.20) to (3.19), we arrive at the estimate

$$\begin{aligned}&\left(\sum_{m=1}^{3}\nabla_{mm}^{N+1}\hat{A}^E(A_1,A_2,\ldots,A_N,A_i,\mathbf{x},\omega)\right)\mathbf{E}^M(\mathbf{x})\cdot\mathbf{E}^M(\mathbf{x})\\&\qquad\le\theta_{t,i}(\mathbf{x},\omega)f_\infty^i(\mathbf{E}^M(\mathbf{x}))\end{aligned} \tag{3.22}$$

for almost every ω in Ω, and the proof of (3.8) of Lemma 3.2 is complete. To prove (3.9) we return to (3.19) and apply the Hölder inequality with $1/p+1/q=1$ to obtain

$$\begin{aligned}&\lim_{\varepsilon_k\to 0}\frac{1}{|Q(\mathbf{x},r_j)|}\int_{Q(\mathbf{x},r_j)}\chi_{t,i}^{\varepsilon_k}(\mathbf{y},\omega)|(I+\underline{\tilde{\mathbf{G}}}(T(\mathbf{y}/\varepsilon_k)\omega))\mathbf{E}^M(\mathbf{x})|^2\,d\mathbf{y}\\&\quad\le\lim_{\varepsilon_k\to 0}\left(\frac{1}{|Q(\mathbf{x},r_j)|}\int_{Q(\mathbf{x},r_j)}\chi_{t,i}^{\varepsilon_k}(\mathbf{y},\omega)\,d\mathbf{y}\right)^{1/q}\\&\qquad\times\lim_{\varepsilon_k\to 0}\left(\frac{1}{|Q(\mathbf{x},r_j)|}\int_{Q(\mathbf{x},r_j)}\tilde{\chi}_i(T(\mathbf{y}/\varepsilon_k)\omega)|(I+\underline{\tilde{\mathbf{G}}}(T(\mathbf{y}/\varepsilon_k)\omega))\mathbf{E}^M(\mathbf{x})|^{2p}d\mathbf{x}\right)^{1/p}.\end{aligned} \tag{3.23}$$

From the Birkhoff ergodic theorem it follows that

$$\begin{aligned}&f_p^i(\mathbf{E}^M(\mathbf{x}))\\&\quad=\lim_{\varepsilon_k\to 0}\left(\frac{1}{|Q(\mathbf{x},r_j)|}\int_{Q(\mathbf{x},r_j)}\tilde{\chi}_i(T(\mathbf{y}/\varepsilon_k)\omega)|(I+\underline{\tilde{\mathbf{G}}}(T(\mathbf{y}/\varepsilon_k)\omega))\mathbf{E}^M(\mathbf{x})|^{2p}d\mathbf{x}\right)^{1/p},\end{aligned} \tag{3.24}$$

and the proof of (3.9) is complete. □

4. Homogenization of Chebyshev's inequality. In this section we establish Proposition 2.1. We start by providing the relationship between the limits $\lim_{\varepsilon_k\to 0}P_i^{\varepsilon_k}(t,C_0,\omega)$, $\lim_{\varepsilon_k\to 0}P^{\varepsilon_k}(t,C_0,\omega)$ and the distribution of states for the square of the electric field intensity in the ith phase. The volume of the subset of the ith phase contained in C_0 where the equivalent stress exceeds t is given by $\lambda_i^{\varepsilon_k}(t,C_0,\omega)=\int_{C_0}\chi_{t,i}^{\varepsilon_k}(\mathbf{x},\omega)d\mathbf{x}$. Passing to a subsequence if necessary, the theory of weak convergence delivers the distribution of states $\theta_{t,i}(\mathbf{x},\omega)$ for which $\lim_{\varepsilon_k\to 0}\lambda_i^{\varepsilon_k}(t,C_0,\omega)=\int_{C_0}\theta_{t,i}(\mathbf{x},\omega)d\mathbf{x}$. For fixed ε_k the volume of the ith phase in the cube C_0 is denoted by $V_i^{\varepsilon_k}$ and $P_i^{\varepsilon_k}(t,C_0,\omega)=\lambda_i^{\varepsilon_k}(t,C_0,\omega)/V_i^{\varepsilon_k}$. From ergodicity, $\lim_{\varepsilon_k\to 0}V_i^{\varepsilon_k}=\theta_i|C_0|$. It is clear that

$$\lim_{\varepsilon_k\to 0}P_i^{\varepsilon_k}(t,C_0,\omega)=\left(\frac{1}{\theta_i|C_0|}\right)\int_{C_0}\theta_{t,i}(\mathbf{x},\omega)\,d\mathbf{x}. \tag{4.1}$$

Set $\theta_t(\mathbf{x},\omega)=\sum_{i=1}^{N}\theta_{t,i}(\mathbf{x},\omega)$; then one has

$$\lim_{\varepsilon_k\to 0}P^{\varepsilon_k}(t,C_0,\omega)=(1/|C_0|)\int_{C_0}\theta_t(\mathbf{x},\omega)\,d\mathbf{x}. \tag{4.2}$$

It follows easily from the homogenization constraint (3.4) that

$$\theta_{t,i}(\mathbf{x},\omega) \leq t^{-p}|f_p^i(\mathbf{E}^M(\mathbf{x}))|^p, \qquad i=1,\ldots,N. \tag{4.3}$$

Taking averages of both sides gives

$$\lim_{\varepsilon_k\to 0} P_i^{\varepsilon_k}(t,C_0,\omega) \leq t^{-p}\left(\frac{1}{\theta_i|C_0|}\right)\int_{C_0}|f_p^i(\mathbf{E}^M(\mathbf{x}))|^p\,d\mathbf{x}, \tag{4.4}$$

and (2.10) of Proposition 2.1 is proved. The inequality (2.11) of Proposition 2.1 follows immediately upon summation of the left and right sides of (4.3) over $i=1,\ldots,N$ and averaging both sides.

5. Bounds on the support set of the electric field intensity distribution function. This section contains the proofs of Propositions 2.2 and 2.3. The homogenization constraint (3.3) is used to prove Proposition 2.2. Integration of (3.3) gives

$$\int_{C_0}\theta_{t,i}(\mathbf{x},\omega)f^i(\mathbf{E}^M(\mathbf{x}))\,d\mathbf{x} - t\int_{C_0}\theta_{t,i}(\mathbf{x},\omega)\,d\mathbf{x} \geq 0, \quad i=1,\ldots,N. \tag{5.1}$$

Application of Hölder's inequality to the first term and division by $\theta_i|C_0|$ gives

$$\lim_{\varepsilon_k\to 0} P_i^{\varepsilon_k}(t,C_0,\omega)(\|f^i(\mathbf{E}^M)\|_{L^\infty(C_0)} - t) \geq 0, \quad i=1,\ldots,N, \tag{5.2}$$

and Proposition 2.2 follows.

To prove Proposition 2.3 we add the constraints (5.1) to get

$$\sum_{i=1}^N\left(\int_{C_0}\theta_{t,i}(\mathbf{x},\omega)f^i(\mathbf{E}^M(\mathbf{x}))\,d\mathbf{x}\right) - t\int_{C_0}\theta_t(\mathbf{x},\omega)\,d\mathbf{x} \geq 0. \tag{5.3}$$

Noting that $M(\mathbf{E}^M(\mathbf{x})) \geq f^i(\mathbf{E}^M(\mathbf{x}))$ gives

$$\int_{C_0}\theta_t(\mathbf{x},\omega)M(\mathbf{E}^M(\mathbf{x}))\,d\mathbf{x} - t\int_{C_0}\theta_t(\mathbf{x},\omega)\,d\mathbf{x} \geq 0. \tag{5.4}$$

Application of Hölder's inequality to the first term and division by $|C_0|$ gives

$$\lim_{\varepsilon_k\to 0} P^{\varepsilon_k}(t,C_0,\omega)(\|M(\mathbf{E}^M(\mathbf{x}))\|_{L^\infty(C_0)} - t) \geq 0, \tag{5.5}$$

and Proposition 2.3 follows.

6. Pointwise bounds on the electric field intensity. In this section we give the proof of Proposition 2.4. From the hypothesis of Propositions 2.4 and 2.3 it follows that $\lim_{k\to\infty}|S_t^{\varepsilon_k}(\omega)\cap C_0| = 0$. We choose a subsequence $\{\varepsilon_{k'}\}_{k'=1}^\infty$ such that $|S_t^{\varepsilon_{k'}}(\omega)\cap C_0| < 2^{-k'}$. Then if $\mathbf{x}$ doesn't belong to $\cup_{k'\geq\tilde{K}}^\infty S_t^{\varepsilon_{k'}}(\omega)\cap C_0$, one has that $|\mathbf{E}^{\varepsilon_{k'}}|^2 \leq t$ for every $k' > \tilde{K}$. Hence for any $\mathbf{x}$ not in $A = \cap_{K=1}^\infty \cup_{k'\geq K}^\infty S_t^{\varepsilon_{k'}}(\omega)\cap C_0$ there is an index K for which $|\mathbf{E}^{\varepsilon_{k'}}|^2 \leq t$ for every $k' > K$. But

$$|A| \leq \left|\cup_{k'\geq\tilde{K}}^\infty S_t^{\varepsilon_{k'}}(\omega)\cap C_0\right| \leq \sum_{k'=\tilde{K}}^\infty |S_t^{\varepsilon_{k'}}(\omega)\cap C_0| \leq 2^{-\tilde{K}+1}.$$

Hence $|A| = 0$. Thus for almost every $\mathbf{x}$ in C_0 there is a finite index K (that may depend upon $\mathbf{x}$ and ω) for which $|\mathbf{E}^{\varepsilon_{k'}}|^2 \leq t$ for every $k' > K$, and the proposition follows.

7. Upper bounds on the stress distribution function. In this section Propositions 2.5 and 2.6 are derived. For a cube C_0 contained inside the composite, the set of points where $M(\mathbf{E}^M(\mathbf{x})) \geq t$ is denoted by $\{\mathbf{x} \text{ in } C_0;\ M(\mathbf{E}^M(\mathbf{x})) \geq t\}$. We start by establishing the inequality

$$\lim_{\varepsilon_k \to 0} P^{\varepsilon_k}(t, C_0, \omega) \leq \frac{|\{\mathbf{x} \text{ in } C_0;\ M(\mathbf{E}^M(\mathbf{x})) \geq t\}|}{|C_0|}. \tag{7.1}$$

Adding the homogenization constraints gives

$$\theta_t(\mathbf{x}, \omega)(M(\mathbf{E}^M(\mathbf{x})) - t) \geq 0. \tag{7.2}$$

Thus from (7.2) it is evident that at almost every point for which $\theta_t(\mathbf{x}, \omega) > 0$, one has that $M(\mathbf{E}^M(\mathbf{x})) \geq t$. The set of points in C_0 for which $\theta_t(\mathbf{x}, \omega) > 0$ is denoted by $\{\mathbf{x} \text{ in } C_0;\ \theta_t(\mathbf{x}, \omega) > 0\}$, and it is clear that

$$|\{\mathbf{x} \text{ in } C_0;\ \theta_t(\mathbf{x}, \omega) > 0\}| \leq |\{\mathbf{x} \text{ in } C_0;\ M(\mathbf{E}^M(\mathbf{x})) \geq t\}|. \tag{7.3}$$

Since $0 \leq \theta_t(\mathbf{x}, \omega) \leq 1$, one has the estimate

$$\int_{C_0} \theta_t(\mathbf{x}, \omega)\, d\mathbf{x} \leq |\{\mathbf{x} \text{ in } C_0;\ \theta_t(\mathbf{x}, \omega) > 0\}|, \tag{7.4}$$

and (7.1) follows from (7.3).

We will apply the John–Nirenberg theorem [6] to estimate the right-hand side of (7.1). To do this we show first that

$$|\{\mathbf{x} \text{ in } C_0;\ M(\mathbf{E}^M(\mathbf{x})) \geq t\}| \leq |\{\mathbf{x} \text{ in } C_0;\ |M(\mathbf{E}^M(\mathbf{x})) - M_{C_0}| \geq t - M_{C_0}\}|. \tag{7.5}$$

To see this, note that $M(\mathbf{E}^M(\mathbf{x})) \leq |M(\mathbf{E}^M(\mathbf{x})) - M_{C_0}| + M_{C_0}$, so

$$\{\mathbf{x} \text{ in } C_0;\ M(\mathbf{E}^M(\mathbf{x})) \geq t\} \subset \{\mathbf{x} \text{ in } C_0;\ |M(\mathbf{E}^M(\mathbf{x})) - M_{C_0}| \geq t - M_{C_0}\}, \tag{7.6}$$

and (7.5) follows. Application of the John–Nirenberg theorem gives

$$\frac{|\{\mathbf{x} \text{ in } C_0;\ |M(\mathbf{E}^M(\mathbf{x})) - M_{C_0}| \geq s\}|}{|C_0|} \leq \begin{cases} 1 & \text{for } 0 < s \leq 8\|M\|_{BMO}\alpha^{-1}, \\ \alpha^{-1} e^{[-(C(\alpha)\times(s))]} & \text{for } 8\|M\|_{BMO}\alpha^{-1} < s. \end{cases} \tag{7.7}$$

Proposition 2.5 follows immediately from the change of variables $s = t - M_{C_0}$ and the inequalities (7.1), (7.5), and (7.7). The function obtained by the change of variables $s = t - M_{C_0}$ in (7.7) is denoted by $\overline{P}_\alpha(t, C_0)$, and

$$\overline{P}_\alpha(t, C_0) = \begin{cases} 1 & \text{for } 0 < t - M_{C_0} \leq 8\|M\|_{BMO}\alpha^{-1}, \\ \alpha^{-1} e^{[-(C(\alpha)\times(t - M_{C_0}))]} & \text{for } 8\|M\|_{BMO}\alpha^{-1} < t - M_{C_0}. \end{cases} \tag{7.8}$$

It is evident from the estimates that $\lim_{\varepsilon_k \to 0} P^{\varepsilon_k}(t, C_0, \omega) \leq \overline{P}_\alpha(t, C_0)$ for $M_{C_0} < t$. Tighter upper bounds are given by optimizing over α, i.e.,

$$\lim_{\varepsilon_k \to 0} P^{\varepsilon_k}(t, C_0, \omega) \leq \overline{U}(t, C_0) = \inf_{0 < \alpha < 1} \overline{P}_\alpha(t, C_0). \tag{7.9}$$

Here $\overline{U}(t, C_0)$ is continuous and decreasing and is given by

(7.10)
$$\overline{U}(t, C_0) = \begin{cases} 1 & \text{for } 0 < t - M_{C_0} \leq 8\|M\|_{BMO}, \\ (\alpha(t))^{-1} e \times e^{[-\alpha(t)(t-M_{C_0})/(8\|M\|_{BMO})]} & \text{for } 8\|M\|_{BMO} + M_{C_0} < t. \end{cases}$$

The factor $\alpha(t)$ lies in the interval $e^{-1} < \alpha(t) < 1$ and is the root of the equation

$$\kappa^{-1} - \alpha(1 + \ln \alpha) = 0, \tag{7.11}$$

where $\kappa = (t - M_{C_0})/(8\|M\|_{BMO})$. Proposition 2.6 now follows immediately from (7.10).

8. Macrofield modulation functions for random two-phase layered composites. In this section we treat randomly layered media and give an example of how the macrofield modulation functions are used to assess the field distribution inside a finite size sample. We start by considering a two-dimensional electrostatic problem on the plane R^2 and derive explicit formulas for the moments of the electric field. The plane is partitioned into layers of unit thickness parallel to the y_2 axis. Each layer contains an isotropic dielectric material having either dielectric constant α or β with $\alpha < \beta$. The particular value of the dielectric constant in each layer is given by a Bernoulli process; i.e., a biased coin that takes heads with probability θ and tails with probability $1 - \theta$ is used to assign the dielectric constant in each layer. Over each layer the coin is flipped, and if the coin lands heads up, the layer is assigned the β dielectric; otherwise it assigned the α dielectric. In section 8.1 we calculate the moments of the electric field directly using the strong law of large numbers. In section 8.2 we apply these results and use Proposition 2.2 to assess the distribution of the electric field intensity inside an L-shaped domain filled with a highly oscillatory random laminate in the presence of a prescribed electric charge density.

8.1. Moments of the electric field for random two-phase layered composites. For a given infinite sequence of biased coin flips, we arrive at a realization of the random medium. The indicator function ω of the β phase is a function of the y_1 coordinate and takes the value one in the β phase and zero outside it. For convenience we choose the origin of the $y_1 - y_2$ coordinate system to lie on a two-phase interface, with the β phase on the left and the α phase on the right. The coordinates of the interfaces between α and β phases on the positive y_1 axis are given by the sequence $\{N_n\}_{n=1}^{\infty}$ and $N_0 = 0$. The coordinates of the interfaces between phases on the negative y_1 axis are given by $\{N_n\}_{n=-1}^{-\infty}$. Let $\mathbf{e}^1$ and $\mathbf{e}^2$ be unit vectors pointing in the directions of the y_1 axis and y_2 axis, respectively. For imposed electric field gradients $\mathbf{e}^k$, $k = 1, 2$, the fluctuating part of the electric potential φ^k is continuous and solves the two-dimensional version of the field problem (2.2) given by

$$\begin{aligned} &\Delta\varphi^k = 0 \quad \text{inside each layer,} \\ &\beta(\partial_{y_1}\varphi^k_{|_L} + \mathbf{e}^k_1) = \alpha(\partial_{y_1}\varphi^k_{|_R} + \mathbf{e}^k_1) \text{ on interfaces.} \end{aligned} \tag{8.1}$$

It is clear from the above that $\varphi^1 = \varphi^1(y_1)$ and $\varphi^2 = \text{const}$. In this context the analogue of (2.1) is given by

$$\lim_{r\to\infty} \frac{\int_{-r}^{r} \partial_{y_1}\varphi^1 dy_1}{2r} = \lim_{r\to\infty} \frac{\varphi^1(r) - \varphi^1(-r)}{2r} = 0 \tag{8.2}$$

and $\varphi^k(0) = 0$. Clearly $\varphi^2 = 0$, and the potential φ^1 is a continuous piecewise linear function of y_1, i.e., in each phase φ^1 is of the form $\varphi^1(y_1) = ay_1 + b$ where the constants a and b change between phases. Application of (8.1), the continuity conditions at two-phase interfaces, and (8.2) together with the strong law of large numbers shows that $\varphi^1(y_1)$ is given a.s. by the following formulas.

For $N_n \le y_1 < N_{n+1}$ and $n+1$ even, the potential is given by

$$\varphi^1(y_1) = (k_2 - 1)y_1 + k_1(N_1 - N_2 + N_3 - N_4 + \cdots + N_n), \tag{8.3}$$

and for $n+1$ odd

$$\varphi^1(y_1) = (k_3 - 1)y_1 + k_1(N_1 - N_2 + N_3 - N_4 + \cdots - N_n). \tag{8.4}$$

For $N_{-(n+1)} < y_1 \le N_{-n}$ and $n+1$ even, the potential is given by

$$\varphi^1(y_1) = (k_3 - 1)y_1 + k_1(-N_{-1} + N_{-2} - N_{-3} + N_{-4} + \cdots - N_{-n}), \tag{8.5}$$

and for $n+1$ odd

$$\varphi^1(y_1) = (k_2 - 1)y_1 + k_1(-N_{-1} + N_{-2} - N_{-3} + N_{-4} + \cdots + N_{-n}), \tag{8.6}$$

where the constants k_1, k_2, and k_3 are defined by

$$\begin{aligned} k_1 &= \frac{\beta - \alpha}{\alpha + (\beta - \alpha)(1 - \theta)}, \\ k_2 &= \frac{\alpha}{\alpha + (\beta - \alpha)(1 - \theta)}, \\ k_3 &= \frac{\beta}{\alpha + (\beta - \alpha)(1 - \theta)}. \end{aligned} \tag{8.7}$$

The derivative $\partial_{y_1}\varphi^1$ is given by the following formula:

$$\begin{aligned} \partial_{y_1}\varphi^1 = \gamma_\alpha &= \frac{\theta(\beta - \alpha)}{\alpha + (\beta - \alpha)(1 - \theta)} \text{ in the } \alpha \text{ phase}, \\ \partial_{y_1}\varphi^1 = \gamma_\beta &= \frac{-(1 - \theta)(\beta - \alpha)}{\alpha + (\beta - \alpha)(1 - \theta)} \text{ in the } \beta \text{ phase}. \end{aligned} \tag{8.8}$$

For an imposed constant applied field of the general form $\overline{\mathbf{E}} = E_1\mathbf{e}^1 + E_2\mathbf{e}^2$, the local electric field $\mathbf{E}(\mathbf{y}, \omega)$ is given by

$$\mathbf{E}(\mathbf{y}, \omega) = (1 - \omega(y_1))((1 + \gamma_\alpha)E_1\mathbf{e}^1 + E_2\mathbf{e}^2) + \omega(y_1)((1 + \gamma_\beta)E_1\mathbf{e}^1 + E_2\mathbf{e}^2). \tag{8.9}$$

We average over the plane and apply the strong law of large numbers to obtain the moments of the local electric field given by

$$\begin{aligned} f_p^1(\overline{\mathbf{E}}) &= \lim_{r\to\infty}\left(\frac{1}{2r}\int_{-r}^{r}(1 - \omega(y_1))|\mathbf{E}(\mathbf{y}, \omega)|^{2p}\,dy_1\right)^{1/p} \\ &= (1 - \theta)^{1/p}((1 + \gamma_\alpha)^2E_1^2 + E_2^2), \end{aligned} \tag{8.10}$$

$$\begin{aligned} f_p^2(\overline{\mathbf{E}}) &= \lim_{r\to\infty}\left(\frac{1}{2r}\int_{-r}^{r}\omega(y_1)|\mathbf{E}(\mathbf{y}, \omega)|^{2p}\,dy_1\right)^{1/p} \\ &= \theta^{1/p}((1 + \gamma_\beta)^2E_1^2 + E_2^2). \end{aligned} \tag{8.11}$$

FIG. 8.1. *A realization for* $\theta = 1/3$.

The local dielectric constant in the random laminate is given by

$$A(\mathbf{y},\omega) = \alpha(1-\omega(y_1)) + \beta\omega(y_1), \tag{8.12}$$

and the effective tensor A^E is given by

$$\begin{aligned} A^E\overline{\mathbf{E}} &= \lim_{r\to\infty}\frac{1}{2r}\int_{-r}^{r} A(\mathbf{y},\omega)\mathbf{E}(\mathbf{y},\omega)\,dy_1 \\ &= (\alpha^{-1}(1-\theta)+\beta^{-1}\theta)^{-1}E_1 + (\alpha(1-\theta)+\beta\theta)E_2. \end{aligned} \tag{8.13}$$

The random laminate described above is an example of a symmetric cell material [15]. A standard construction delivers the probability space $(\Omega,\mathcal{F},\mathcal{P})$ and dynamical system associated with the symmetric cell material; see [8, 17]. Using this, one rewrites the averages given in (8.10), (8.11), and (8.13) in terms of the ensemble averages used to define the moments of the local electric field and effective dielectric constant in section 2.

8.2. Electric field assessment for a randomly layered dielectric in an *L*-shaped domain. In this subsection we apply the theory presented in section 2 to assess the electric field distribution inside an L-shaped domain containing a highly oscillatory random laminate with length scale $\varepsilon_k = 1/k$, $k = 1, 2, \ldots$. Here the L-shaped domain is taken to have side length one. The dielectric constant for the highly oscillatory random laminate inside the L-shaped domain is given by

$$A^{\varepsilon_k}(\mathbf{x},\omega) = A(x_1/\varepsilon_k,\omega), \tag{8.14}$$

where $A(\mathbf{y},\omega)$ is given by the Bernoulli process (8.12) with $\theta = 1/3$. A realization of the random laminate with characteristic length scale ε_{40} is given in Figure 8.1. Here the subdomain in white is the α dielectric and the subdomain in black is the β dielectric.

The electric potential $\phi^{\varepsilon_k}(\mathbf{x},\omega)$ is the solution of

$$-\mathrm{div}(A^{\varepsilon_k}(\mathbf{x},\omega)\nabla\phi^{\varepsilon_k}(\mathbf{x},\omega)) = 10 \tag{8.15}$$

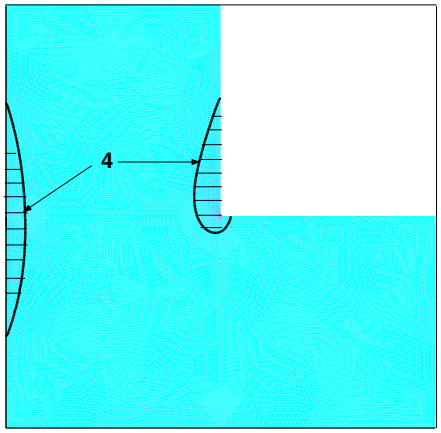


FIG. 8.2. *Distribution of the electric field intensity in the α dielectric.*

inside the L-shaped domain and $\phi^{\varepsilon_k}(\mathbf{x},\omega) = 0$ on the boundary. The associated electric field is given by $\mathbf{E}^{\varepsilon_k}(\mathbf{x},\omega) = -\nabla\phi^{\varepsilon_k}(\mathbf{x},\omega)$. The goal of this application is to characterize the distributions $\lim_{\varepsilon_k\to 0} P_i^{\varepsilon_k}(t, C_0, \omega)$, $i = 1, 2$. Here C_0 can be any square contained inside the L-shaped domain. To do this we solve numerically for the macroscopic potential and electric field and construct the macrofield modulation functions. The macroscopic electric potential $\phi^M(\mathbf{x})$ satisfies the boundary condition $\phi^M(\mathbf{x}) = 0$ and

$$-\mathrm{div}(A^E \nabla\phi^M(\mathbf{x})) = 10. \tag{8.16}$$

The macroscopic electric field is given by $\mathbf{E}^M(\mathbf{x}) = -\nabla\phi^M(\mathbf{x})$. The macrofield modulation functions are given by

$$f_p^1(\mathbf{E}^M(\mathbf{x})) = (1-\theta)^{1/p}((1+\gamma_\alpha)^2|\partial_{x_1}\phi^M(\mathbf{x})|^2 + |\partial_{x_2}\phi^M(\mathbf{x})|^2), \tag{8.17}$$

$$f_p^2(\mathbf{E}^M(\mathbf{x})) = \theta^{1/p}((1+\gamma_\beta)^2|\partial_{x_1}\phi^M(\mathbf{x})|^2 + |\partial_{x_2}\phi^M(\mathbf{x})|^2). \tag{8.18}$$

For the computation we choose $\alpha = 2$ and $\beta = 10$ and restrict our attention to $f_\infty^1(\mathbf{E}^M(\mathbf{x}))$ and $f_\infty^2(\mathbf{E}^M(\mathbf{x}))$. To illustrate the ideas, the level curves given by $f_\infty^1(\mathbf{E}^M(\mathbf{x})) = 4$ are plotted in Figure 8.2. The lined regions indicate where $f_\infty^1(\mathbf{E}^M(\mathbf{x})) > 4$ and $f_\infty^1(\mathbf{E}^M(\mathbf{x})) < 4$ outside these. For any square C_0 that doesn't intersect the lined regions, Proposition 2.2 implies that

$$\lim_{\varepsilon_k\to 0} P_1^{\varepsilon_k}(t, C_0, \omega) = 0 \quad \text{for } t > 4 \tag{8.19}$$

for almost every realization ω. In this way it is seen that the lined regions provide an asymptotically exact bound on the set where $|\mathbf{E}^{\varepsilon_k}(\mathbf{x},\omega)|^2 > 4$ in the α dielectric.

The level curves given by $f_\infty^2(\mathbf{E}^M(\mathbf{x})) = 1$ are plotted in Figure 8.3. The lined regions indicate where $f_\infty^2(\mathbf{E}^M(\mathbf{x})) > 1$ and $f_\infty^1(\mathbf{E}^M(\mathbf{x})) < 1$ outside these. For any square C_0 that doesn't intersect the lined regions, Proposition 2.2 implies that

$$\lim_{\varepsilon_k\to 0} P_2^{\varepsilon_k}(t, C_0, \omega) = 0 \quad \text{for } t > 1 \tag{8.20}$$

for almost every realization ω. It is seen as before that the lined regions provide an asymptotically exact bound on the set where $|\mathbf{E}^{\varepsilon_k}(\mathbf{x},\omega)|^2 > 1$ in the β dielectric.

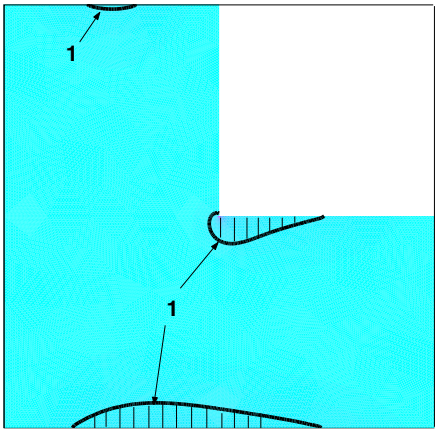


FIG. 8.3. *Distribution of the electric field intensity in the β dielectric.*

Appendix. Here we establish the inequality stated in (3.20) given by

$$\begin{aligned} &\|\tilde{\chi}_i(T(\mathbf{y})\omega)|(I+\underline{\tilde{\mathbf{G}}}(T(\mathbf{y})\omega))\mathbf{E}^M(\mathbf{x})|^2\|_{L^\infty(R^3)} \\ &\quad\le \|\tilde{\chi}_i(\omega)|(I+\underline{\tilde{\mathbf{G}}}(\omega))\mathbf{E}^M(\mathbf{x})|^2\|_{L^\infty(\Omega)} \end{aligned} \tag{A.1}$$

for almost every ω. To establish (A.1) put $\alpha = \|\tilde{\chi}_i(\omega)|(I+\underline{\tilde{\mathbf{G}}}(\omega))\mathbf{E}^M(\mathbf{x})|^2\|_{L^\infty(\Omega)}$ and introduce the set $\mathcal{G} = \{\omega : \tilde{\chi}_i(\omega)|(I+\underline{\tilde{\mathbf{G}}}(\omega))\mathbf{E}^M(\mathbf{x})|^2 \le \alpha\}$. From Lemma 7.1 of [8] there exists a set $\mathcal{G}_1 \subset \mathcal{G}$ for which $\mathcal{P}(\mathcal{G}_1) = 1$, and for any fixed ω in $\mathcal{G}_1$, one has that $T(\mathbf{y})\omega$ is in $\mathcal{G}$ for almost every $\mathbf{y}$ in R^3, and (A.1) follows.

REFERENCES

[1] A. BENSOUSSAN, J. L. LIONS, AND G. PAPANICOLAOU, *Asymptotic Analysis for Periodic Structures*, Stud. Math. Appl. 5, North-Holland, Amsterdam, 1978.
[2] A. BOURGREAT, A. MIKELIĆ, AND S. WRIGHT, *Stochastic two-scale convergence in the mean and applications*, J. Reine Angew. Math., 456 (1994), pp. 19–51.
[3] L. BREIMAN, *Probability*, Addison-Wesley, Reading, MA, 1968.
[4] H. CHENG AND S. TORQUATO, *Electric field fluctuations in random dielectric composites*, Phys. Rev. B, 56 (1997), pp. 8060–8068.
[5] K. GOLDEN AND G. PAPANICOLAOU, *Bounds for effective parameters of heterogeneous media by analytic continuation*, Comm. Math. Phys., 90 (1983), pp. 473–491.
[6] F. JOHN AND L. NIRENBERG, *On functions of bounded mean oscillation*, Comm. Pure Appl. Math., 14 (1961), pp. 415–426.
[7] D. JEULIN, *Random structure models for composite media and fracture statistics*, in Advances in Mathematical Modeling of Composite Materials, K. Z. Markov, ed., Advances in Mathematics for Applied Sciences 15, World Scientific, Singapore, 1994, pp. 239–289.
[8] V. V. JIKOV, S. M. KOZLOV, AND O. A. OLEINIK, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, Heidelberg, London, New York, 1994.
[9] A. KELLY AND N. H. MACMILLAN, *Strong Solids*, Monographs on the Physics and Chemistry of Materials, Clarendon Press, Oxford, 1986.
[10] S. M. KOZLOV, *Averaging of random structures*, Dokl. Akad. Nauk SSSR, 241 (1978), pp. 1016–1019 (in Russian); Soviet Math. Dokl., 19 (1978), pp. 950–954 (in English).
[11] R. LIPTON, *Assessment of the local stress state through macroscopic variables*, R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci., 361 (2003), pp. 921–946.
[12] R. LIPTON, *Relaxation through homogenization for optimal design problems with gradient constraints*, J. Optim. Theory Appl., 114 (2002), pp. 27–53.
[13] R. LIPTON, *Stress constrained G closure and relaxation of structural design problems*, Quart. Appl. Math., 62 (2004), pp. 295–321.

[14] R. Lipton, *Bounds on the distribution of extreme values for the stress in composite materials*, J. Mech. Phys. Solids, 52 (2004), pp. 1053–1069.
[15] G. W. Milton, *The Theory of Composites*, Cambridge University Press, Cambridge, UK, 2002.
[16] F. Murat and L. Tartar, *H convergence*, in Topics in the Mathematical Modelling of Composite Materials, A. V. Cherkaev and R. V. Kohn, eds., Birkhäuser, Boston, 1997, pp. 21–43.
[17] G. Papanicolaou and S. R. S. Varadhan, *Boundary value problems with rapidly oscillating random coefficients*, Colloquia Mathematica Sociietatis János Bolyai, 27 (1982), pp. 835–873.
[18] E. Sanchez-Palencia, *Non-Homogeneous Media and Vibration Theory*, Springer-Verlag, Berlin, 1980.
[19] S. Spagnolo, *Convergence in energy for elliptic operators*, in Proceedings of the Third Symposium on Numerical Solutions of Partial Differential Equations, B. Hubbard, ed., Academic Press, New York, 1976, pp. 469–498.

SIAM J. APPL. MATH.
Vol. 65, No. 2, pp. 494–520

PARAMETRIC RESONANCE IN IMMERSED ELASTIC BOUNDARIES*

RICARDO CORTEZ†, CHARLES S. PESKIN‡,
JOHN M. STOCKIE§, AND DOUGLAS VARELA¶

Abstract. In this paper, we investigate the stability of a fluid-structure interaction problem in which a flexible elastic membrane immersed in a fluid is excited via periodic variations in the elastic stiffness parameter. This model can be viewed as a prototype for active biological tissues such as the basilar membrane in the inner ear, or heart muscle fibers immersed in blood. Problems such as this, in which the system is subjected to internal forcing through a parameter, can give rise to "parametric resonance." We formulate the equations of motion in two dimensions using the immersed boundary formulation. Assuming small amplitude motions, we can apply Floquet theory to the linearized equations and derive an eigenvalue problem whose solution defines the marginal stability boundaries in parameter space. The eigenvalue equation is solved numerically to determine values of fiber stiffness and fluid viscosity for which the problem is linearly unstable. We present direct numerical simulations of the fluid-structure interaction problem (using the immersed boundary method) that verify the existence of the parametric resonances suggested by our analysis.

Key words. immersed boundary, fluid-structure interaction, parametric resonance

AMS subject classifications. 35B34, 35B35, 74F10, 76D05

DOI. 10.1137/S003613990342534X

1. Introduction.

Fluid-structure interaction problems abound in industrial applications as well as in many natural phenomena. Owing to the potentially complex, time-varying geometry and the nonlinear interactions that arise between fluid and solid, such problems represent a major challenge to both mathematical modelers and computational scientists.

This paper deals with a model for fluid-structure interaction known as the immersed boundary (IB) formulation [20], which has proven particularly effective for dealing with complex problems in biological fluid mechanics. In this formulation, the immersed structure is treated as an elastic surface or interwoven mesh of elastic fibers that exert a singular force on a surrounding viscous fluid, while also moving with the velocity of adjacent fluid particles. The associated immersed boundary method has been used to solve a wide range of problems such as blood flow in the heart and arteries [2, 21], swimming microorganisms [8], biofilms [7], and insect flight [16]. More recently, the IB method has also been extended to a much wider range of nonbiological problems involving flow past solid cylinders [12], flapping filaments [29], parachutes [10], and suspensions of flexible particles [25].

*Received by the editors March 31, 2003; accepted for publication (in revised form) February 4, 2004; published electronically December 16, 2004. This work was partially supported by NSF grants DMS-0094179 (Cortez) and DMS-9980069 (Peskin), NSERC grant RGP-238776-01 (Stockie), and the Center for Computational Science at Tulane and Xavier Universities grant DE-FG02-01ER63119 (Cortez).

http://www.siam.org/journals/siap/65-2/42534.html

†Department of Mathematics, Tulane University, 6823 St. Charles Ave., New Orleans, LA 70118 (cortez@math.tulane.edu).

‡Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012 (peskin@cims.nyu.edu).

§Department of Mathematics, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada (stockie@math.sfu.ca).

¶Department of Engineering and Applied Science, California Institute of Technology, Pasadena, CA 91125 (vladimir@its.caltech.edu).

While a significant body of literature has developed around numerical simulations using the IB method (as well as related methods such as the blob projection method [4]), comparatively little work has been done on analyzing the stability behavior of solutions to the underlying equations of motion. Beyer and LeVeque [3] were the first to perform an analysis of the IB formulation in one spatial dimension. Stockie and Wetton [26] investigated the linear stability of a flat fiber in two dimensions which is subjected to a small sinusoidal perturbation, and presented asymptotic results on the frequency and rate of decay for the resulting oscillations. Cortez and Varela [5] performed a nonlinear analysis for a perturbed circular elastic membrane immersed in an inviscid fluid.

In all of this previous work, the immersed boundary was assumed to be passive, moving along with the flow and generating forces only in response to elastic deformation. However, there are many problems in which the immersed boundary is an active material, generating time-dependent forces to drive its motion. Examples include beating heart muscle fibers and flagellated cells, both of which apply periodic forces to direct their motion and that of the surrounding fluid.

It is then natural to ask whether periodic forcing at various frequencies will give rise to resonance. In the case of immersed boundaries, the applied force is *internal*, in the sense that the system is forced through a periodic variation in a system parameter (namely, the elastic properties of the solid material) rather than through an external body force. In contrast to externally forced systems, these internally or parametrically forced systems are known to exhibit parametric resonance, in which the response for linear systems (i.e., assuming small amplitude motions) can become unbounded even in the presence of viscous damping. Once nonlinearities are taken into account, however, the response remains bounded, but the system can still exhibit large-amplitude motion when forced at certain resonant frequencies. Analyses of parametric resonance have appeared for various fluid flows involving surface or interfacial waves wherein the system is forced through gravitational modulation [11, 14, 17] or time-dependent heating [15]. However, to our knowledge there has been no analysis performed on fluid-structure problems that captures the two-way interaction between a fluid and an immersed, flexible structure. In a recent study by Wang [28], an analysis is performed of flutter and buckling instabilities in a paper making headbox, in which a long, flexible vane is driven to vibrate under the influence of periodic forcing from turbulent fluid jets. However, the author assumes a given fluid velocity and that the solid structure has no influence on the fluid motion.

In this paper, we perform an analysis of parametric resonance in a two-dimensional, circular, elastic membrane immersed in Navier–Stokes flow, which is driven by a time-periodic variation in its elastic stiffness parameter. The elastic membrane is under tension due to the incompressible fluid it encloses, and the changes in the elastic stiffness parameter are equivalent to changes in the tension of the membrane that mimic the contraction of biological fibers. The novelty of this work is that it includes the full, two-way interaction between the fluid and the membrane. In sections 2 and 3, we present the linearized equations of motion for the IB formulation and reduce them to a suitable nondimensional form. In sections 4, 5, and 6, we perform a Floquet analysis of the problem and derive a system of equations that relate the amplitude, frequency, and wavenumber of the internal forcing to the physical parameters in the problem. The natural modes for the unforced system are derived in section 7 and compared to the asymptotic expressions for the natural modes in a flat fiber that were derived in [26]. We also compare the natural modes to those observed in computations using the blob projection method, an alternate method for solving the IB problem that

affords higher spatial accuracy. In section 8 we consider the periodically forced problem and determine the conditions under which parametric resonance can occur. The results are verified using numerical simulations with the IB method.

2. Problem definition. Consider an elastic fiber immersed in a two-dimensional, viscous, incompressible fluid of infinite extent. The fiber is a closed loop with resting length of zero, so that in the absence of fluid the fiber would shrink to a point. However, because an incompressible fluid occupies the region inside the fiber, the equilibrium state is a circle of constant radius R, in which the pressure drop across the fiber is balanced by the tension force. Our aim is to investigate a *parametric excitation* of the fiber, in which the elastic stiffness is varied periodically in time.

The equations of motion for the fluid and immersed boundary can be written in terms of the vorticity $\xi(\boldsymbol{x},t)$ and stream function $\psi(\boldsymbol{x},t)$ as [19, p. 1]:

$$\rho\left(\xi_t + \boldsymbol{u}\cdot\nabla\xi\right) = \mu\nabla^2\xi + \hat{\boldsymbol{z}}\cdot\nabla\times\boldsymbol{f}, \tag{2.1a}$$

$$\boldsymbol{u} = -\hat{\boldsymbol{z}}\times\nabla\psi, \tag{2.1b}$$

$$-\nabla^2\psi = \xi, \tag{2.1c}$$

$$\boldsymbol{f}(\boldsymbol{x},t) = \int_0^{2\pi} (K\boldsymbol{X}_s)_s\,\delta(\boldsymbol{x}-\boldsymbol{X})\,ds, \tag{2.1d}$$

$$\boldsymbol{X}_t = \boldsymbol{u}(\boldsymbol{X},t), \tag{2.1e}$$

where $\boldsymbol{x} = (x,y)$ and $\hat{\boldsymbol{z}} = (0,0,1)$. The fluid velocity $\boldsymbol{u}(\boldsymbol{x},t)$ and fiber force $\boldsymbol{f}(\boldsymbol{x},t)$ are functions of position and time, whereas the fiber position $\boldsymbol{X}(s,t) = (X^r(s,t), X^\theta(s,t))$ is a function of time and the Lagrangian fiber parameter $0 \le s \le 2\pi$ (which is the same as the angle θ when the fiber's reference state is a circle). We assume in this paper that the stiffness is a periodic function of time,

$$K = K(s,t) = K_c(1 + 2\tau\sin\omega_o t), \tag{2.2}$$

which has no dependence on the spatial variable s.

We point out that s is a Lagrangian parameter and is not necessarily proportional to arclength since the fiber is allowed to stretch and shrink tangentially. The force density $(K\boldsymbol{X}_s)_s$ thus includes components normal and tangential to the fiber. If we identify the tension $T(s) = K\|\boldsymbol{X}_s\|$ and the unit tangent vector $\hat{\tau}(s) = \boldsymbol{X}_s/\|\boldsymbol{X}_s\|$, we may write the force density term as

$$(K\boldsymbol{X}_s)_s = (T(s)\hat{\tau}(s))_s = T_s(s)\hat{\tau}(s) + T(s)\hat{\tau}_s(s),$$

which shows explicitly that the force resolves itself into normal and tangential components. We also remark that spatial dependence in K can be easily incorporated into our analysis provided that the spatial variations are small (that is, on the order of the parameter ϵ introduced in the following section).

Our motivation comes from active muscle fibers that may contract and relax to generate fluid motion. For this reason we will focus on nonnegative values of the stiffness and we will be concerned mostly with the range $0 \le \tau \le \frac{1}{2}$ in (2.2). However, our analysis does not require this restriction and the results apply to values of $\tau > 1/2$ as well. This is important because there are instances in which a negative stiffness plays an important role as is the case of hair bundles in the bullfrog inner ear [9].

3. Nondimensionalization. We first simplify the problem by scaling the variables and forming dimensionless groups. For now, variables with a tilde will be dimensionless. Let

$$\begin{aligned} \boldsymbol{x} &= R\,\tilde{\boldsymbol{x}}, \\ \boldsymbol{u} &= U\,\tilde{\boldsymbol{u}} \qquad (U \text{ will be determined later}), \\ t &= R/U\,\tilde{t}, \\ K &= K_c\,\tilde{K}, \end{aligned}$$

which imply that

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{x},t) = \frac{K_c}{R}\,\tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}},\tilde{t}) &= \int_0^{2\pi} (\tilde{K}\tilde{\boldsymbol{X}}_s)_s\,\delta(\tilde{\boldsymbol{x}}-\tilde{\boldsymbol{X}})\,ds, \\ \xi = \frac{U}{R}\,\tilde{\xi}, \\ \psi = UR\,\tilde{\psi}. \end{aligned}$$

With these scalings the vorticity equation becomes

$$\tilde{\xi}_{\tilde{t}} + \tilde{\boldsymbol{u}}\cdot\tilde{\nabla}\,\tilde{\xi} = \left(\frac{\mu}{\rho U R}\right)\tilde{\nabla}^2\tilde{\xi} + \left(\frac{K_c}{\rho U^2}\right)\,(\hat{\boldsymbol{z}}\cdot\tilde{\nabla}\times\tilde{\boldsymbol{f}}).$$

In view of this and (2.2), we now define

$$U = R\omega_o, \qquad \nu = \frac{\mu}{\rho U R} = \frac{\mu}{\rho R^2\omega_o}, \qquad \kappa = \frac{K_c}{\rho U^2} = \frac{K_c}{\rho R^2\omega_o^2}, \tag{3.1}$$

so that the unperturbed fiber configuration is the unit circle ($\boldsymbol{X} = \hat{\boldsymbol{r}}(s)$) and we can write the dimensionless equations (omitting the tildes) as

$$\xi_t + \boldsymbol{u}\cdot\nabla\,\xi = \nu\nabla^2\xi + \kappa(\hat{\boldsymbol{z}}\cdot\nabla\times\boldsymbol{f}), \tag{3.2a}$$

$$\boldsymbol{u} = -\hat{\boldsymbol{z}}\times\nabla\psi, \tag{3.2b}$$

$$-\nabla^2\psi = \xi, \tag{3.2c}$$

$$\boldsymbol{f}(\boldsymbol{x},t) = \int_0^{2\pi} (K\boldsymbol{X}_s)_s\,\delta(\boldsymbol{x}-\boldsymbol{X})\,ds, \tag{3.2d}$$

$$\boldsymbol{X}_t = \boldsymbol{u}(\boldsymbol{X},t), \tag{3.2e}$$

with

$$K = (1 + 2\tau\sin t), \tag{3.3a}$$

$$\boldsymbol{X} = \hat{\boldsymbol{r}}(s) + \epsilon\boldsymbol{X}_1^{(1)}(s,t) + \cdots. \tag{3.3b}$$

Our aim is now to solve (3.2a)–(3.2e) with the only two remaining parameters, ν and κ, defined in (3.1). We point out that ν is the reciprocal of the Reynolds number and κ is the square of the natural frequency divided by the driving frequency.

4. Small-amplitude approximation. For the linear analysis, assume the fiber force can be written as

$$\boldsymbol{f} = \boldsymbol{f}^{(0)}(\boldsymbol{x},t) + \epsilon\boldsymbol{f}^{(1)}(\boldsymbol{x},t) + \cdots,$$

so that the term $(\hat{\boldsymbol{z}} \cdot \nabla \times \boldsymbol{f})$ in (3.2a) can be expanded as

$$(\hat{\boldsymbol{z}} \cdot \nabla \times \boldsymbol{f}) = (\hat{\boldsymbol{z}} \cdot \nabla \times \boldsymbol{f}^{(0)}) + \epsilon(\hat{\boldsymbol{z}} \cdot \nabla \times \boldsymbol{f}^{(1)}) + \cdots.$$

CLAIM 1. *Given the expansions in* (3.3a) *and* (3.3b),

$$(\hat{\boldsymbol{z}} \cdot \nabla \times \boldsymbol{f}^{(0)}) = 0 \qquad \textit{and}$$
$$(\hat{\boldsymbol{z}} \cdot \nabla \times \boldsymbol{f}^{(1)}) = K(t)\left(X^{\theta}_{ss} + X^{r}_{s}\right)\left(\frac{\delta(r-1)}{r}\right)_r - K(t)\left(X^{r}_{sss} - X^{\theta}_{ss}\right)\frac{\delta(r-1)}{r}.$$

The proof is found in Appendix B.

Since the leading-order term in the curl of the force is zero, the corresponding flow is simply the trivial solution $\xi^{(0)} = \psi^{(0)} = 0$. Therefore, we look for a series solution expanded in terms of powers of a small parameter ϵ:

$$\xi = \epsilon\, \xi^{(1)}(\boldsymbol{x}, t) + \cdots,$$
$$\psi = \epsilon\, \psi^{(1)}(\boldsymbol{x}, t) + \cdots.$$

In the remainder of this work, we assume ϵ is small and consider the $O(\epsilon)$ equations

$$\text{(4.1a)}\quad \xi_t = \nu\nabla^2\xi + \kappa K(t)\left(X^{\theta}_{ss} + X^{r}_{s}\right)\left[\frac{\delta(r-1)}{r}\right]_r - \kappa K(t)\left(X^{r}_{sss} - X^{\theta}_{ss}\right)\frac{\delta(r-1)}{r},$$
$$\text{(4.1b)}\quad \nabla^2\psi = -\xi,$$
$$\text{(4.1c)}\quad X^{r}_{t} = \psi_{\theta}\big|_{r=1},$$
$$\text{(4.1d)}\quad X^{\theta}_{t} = -\psi_{r}\big|_{r=1},$$

where we have omitted the superscript $^{(1)}$ on most variables because we will be working solely with these quantities from now on.

It is worthwhile mentioning that in the above expansions, the influence of the time-dependent stiffness coefficient is felt solely at first order in ϵ, and so we need only consider the forcing terms up to $O(\epsilon)$. The derivation can be easily extended to include the case where the stiffness is spatially dependent, for which the s-variation appears as an order ϵ perturbation of a time-dependent stiffness,

$$K = K(s,t) = K^{(0)}(t) + \epsilon K^{(1)}(s,t) + O(\epsilon^2).$$

When a Floquet-type expansion is assumed for $K^{(1)}(s,t)$ (refer to the next section), correction terms appear in the $O(\epsilon)$ equations, and these corrections complicate the analysis somewhat by coupling together the various spatial modes. We do not consider a spatially dependent stiffness in this paper.

It is quite striking that a spatially homogeneous $K(t)$ can have any interesting effects in this problem. This is because such $K(t)$ would have no effect at all on the equilibrium state in which the fiber is a circle. Starting with that equilibrium state as initial data and imposing time-periodic $K(t)$ will produce only an oscillation of the pressure and no motion of the fiber at all. But if instead we start with an initial condition that is arbitrarily close but not exactly equal to a circular equilibrium state, we can nevertheless build up a large-amplitude and spatially inhomogeneous oscillation of the fiber through the application of spatially homogeneous but time-periodic $K(t)$ at the right driving frequency. This is actually typical of parametric

resonance. A pendulum that is initially at rest hanging straight down cannot be made to swing by imposing periodic changes in its length, but a pendulum that is initially swinging ever so slightly can indeed be made to swing violently by the application of precisely such changes in length, as every child who swings in the playground knows [6].

5. Floquet analysis. Since we are not simply looking for natural modes but rather are investigating the response of the system to a periodic forcing with a given frequency, we must look for a solution in the form of an infinite series. This is a very general approach in Floquet theory [18], and our derivation parallels the analysis of Kumar and Tuckerman [11] for a horizontal fluid interface consisting of two fluids with different densities. It is worth mentioning that other approaches have also been used to study parametric resonances. For example, Semler and Païdoussis [23] analyze the case of a tubular beam filled with fluid whose velocity contains sinusoidal fluctuations. They use separation of variables and nonlinear dynamics techniques to solve the perturbation equations for small amplitude fluctuations. We will see that our approach will result in a system of equations in block form that is very similar to the equations derived in [23].

We assume that the unknown functions can be written as series of the following form:

$$\begin{aligned}
\xi(r,\theta,t) &= e^{ip\theta}\sum_{n=-\infty}^{\infty}\xi_n(r)e^{(\gamma+in)t},\\
\psi(r,\theta,t) &= e^{ip\theta}\sum_{n=-\infty}^{\infty}\psi_n(r)e^{(\gamma+in)t},\\
X^r(\theta,t) &= e^{ip\theta}\sum_{n=-\infty}^{\infty}X^r_n e^{(\gamma+in)t},\\
X^\theta(\theta,t) &= e^{ip\theta}\sum_{n=-\infty}^{\infty}X^\theta_n e^{(\gamma+in)t},
\end{aligned}$$

where $i=\sqrt{-1}$, p is an integer with $p>1$, and $\gamma=\alpha+i\beta$ with α and β real. To simplify the notation, we define $\mathcal{E}_n \doteq e^{(\gamma+in)t+ip\theta}$ and then write the equation for the vorticity from (4.1a) as

$$\begin{aligned}
\sum_{n=-\infty}^{\infty}\left((\gamma+in)+\frac{\nu p^2}{r^2}\right)\xi_n\mathcal{E}_n \;=\; \sum_{n=-\infty}^{\infty}\Bigg\{&\frac{\nu}{r}(r\xi_n')' + \kappa K(-p^2X_n^\theta + ipX_n^r)\left(\frac{\delta(r-1)}{r}\right)'\\
&+\kappa K(ip^3X_n^r - p^2X_n^\theta)\,\frac{\delta(r-1)}{r}\Bigg\}\mathcal{E}_n. \qquad (5.1)
\end{aligned}$$

Except for the stiffness $K(t)$, all of the t- and θ-dependence is now encompassed by the factor $\mathcal{E}_n$. Note that ξ_n and ψ_n are functions of r, while X_n^r and X_n^θ are constants. The primes in the equations denote derivatives with respect to r. If we take advantage of the fact that the stiffness from (3.3a) can be rewritten as

$$K(t) = 1 - i\tau e^{it} + i\tau e^{-it} \qquad (5.2)$$

and then rearrange terms in the series in (5.1) and divide by $\mathcal{E}_n$, we obtain

$$-\frac{1}{r}(r\xi_n')' + \left(\frac{(\gamma+in)}{\nu} + \frac{p^2}{r^2}\right)\xi_n \tag{5.3}$$
$$= \frac{\kappa}{\nu}\left(\frac{\delta(r-1)}{r}\right)'\left[-p^2\left(X_n^\theta - i\tau X_{n-1}^\theta + i\tau X_{n+1}^\theta\right) + ip\left(X_n^r - i\tau X_{n-1}^r + i\tau X_{n+1}^r\right)\right]$$
$$+ \frac{\kappa}{\nu}\frac{\delta(r-1)}{r}\left[-p^2\left(X_n^\theta - i\tau X_{n-1}^\theta + i\tau X_{n+1}^\theta\right) + ip^3\left(X_n^r - i\tau X_{n-1}^r + i\tau X_{n+1}^r\right)\right]$$

for all integer values of n. The equations for the remaining unknowns in (4.1b)–(4.1d) do not involve $K(t)$, and so for the nth coefficient functions we have

$$-\frac{1}{r}(r\psi_n')' + \frac{p^2}{r^2}\psi_n = \xi_n,$$
$$(\gamma+in)X_n^r = ip\psi_n(1),$$
$$(\gamma+in)X_n^\theta = -\psi_n'(1).$$

A more useful formulation of the problem is in terms of jump conditions across the fiber. The delta function terms in (5.3) are nonzero only on the fiber, and so the following two equations hold on either side of the fiber:

$$-\frac{1}{r}(r\xi_n')' + \left(\frac{(\gamma+in)}{\nu} + \frac{p^2}{r^2}\right)\xi_n = 0, \tag{5.4a}$$
$$-\frac{1}{r}(r\psi_n')' + \frac{p^2}{r^2}\psi_n = \xi_n. \tag{5.4b}$$

The fiber evolution equations hold on the fiber (at $r = 1$):

$$(\gamma+in)X_n^r = ip\psi_n(1), \tag{5.4c}$$
$$(\gamma+in)X_n^\theta = -\psi_n'(1). \tag{5.4d}$$

At the same time, the jump conditions connect the solutions on either side of the fiber (see Appendix C):

$$[\![\xi_n]\!] = \frac{\kappa}{\nu}\left[p^2\left(X_n^\theta - i\tau X_{n-1}^\theta + i\tau X_{n+1}^\theta\right) - ip\left(X_n^r - i\tau X_{n-1}^r + i\tau X_{n+1}^r\right)\right], \tag{5.4e}$$
$$[\![\xi_n']\!] = -\frac{i\kappa p(p^2-1)}{\nu}\left(X_n^r - i\tau X_{n-1}^r + i\tau X_{n+1}^r\right), \tag{5.4f}$$
$$[\![\psi_n]\!] = 0, \tag{5.4g}$$
$$[\![\psi_n']\!] = 0, \tag{5.4h}$$

where we denote the jump in a quantity q by $[\![q]\!] = q|_{r=1^+} - q|_{r=1^-}$. These jump conditions were found in the same way as in [13, 22], by integrating the equations for vorticity across the fiber. We point out that this system of equations cannot be reduced to a Mathieu equation (see [11]).

6. Solution of the vorticity and stream function equations. The equations (5.4) from the previous section can be solved explicitly for each integer value of n, both inside and outside the fiber. Due to (5.4c) and (5.4d), the character of the solution will be different depending on whether the quantity $(\gamma + in)$ is zero. This will happen for $n = 0$ whenever $\gamma = 0$. We therefore solve the equations considering two cases.

6.1. Case $(\gamma+in) \neq 0$. If we make the change of variables $z^2 = -r^2(\gamma+in)/\nu$, then (5.4a) turns into the Bessel equation

$$z^2\xi_n''(z) + z\xi_n'(z) + [z^2 - p^2]\xi_n(z) = 0.$$

The appropriate solution, rewritten in terms of r, is [27]

$$\xi_n(r) = \begin{cases} b_n J_p(i\Omega_n r) & \text{if } r < 1 \qquad \text{(inner)}, \\ a_n H_p(i\Omega_n r) & \text{if } r > 1 \qquad \text{(outer)}, \end{cases}$$

where $\Omega_n \doteq \sqrt{(\gamma+in)/\nu}$ is chosen as the square root with positive real part, $J_p(z)$ is the Bessel function of the first kind, $H_p(z) = J_p(z) + iY_p(z)$ is the Hankel function of the first kind, and a_n and b_n are arbitrary constants (see Appendix A).

The corresponding stream function is given by (see Appendix D)

$$\psi_n(r) = \frac{1}{2ip\Omega_n} \begin{cases} b_n r\Big[J_{p-1}(i\Omega_n r) + J_{p+1}(i\Omega_n r)\Big] \\ \qquad + r^p\Big[a_n H_{p-1}(i\Omega_n) - b_n J_{p-1}(i\Omega_n)\Big] \text{ if } 0 \le r < 1, \\ a_n r\Big[H_{p-1}(i\Omega_n r) + H_{p+1}(i\Omega_n r)\Big] \\ \qquad - r^{-p}\Big[a_n H_{p+1}(i\Omega_n) - b_n J_{p+1}(i\Omega_n)\Big] \text{ if } r > 1. \end{cases}$$

One can easily check that $\psi_n(r)$ and $\psi_n'(r)$ are continuous across the boundary and that

$$\psi_n(1) = \frac{1}{2ip\Omega_n}\Big[a_n H_{p-1}(i\Omega_n) + b_n J_{p+1}(i\Omega_n)\Big],$$
$$\psi_n'(1) = \frac{1}{2i\Omega_n}\Big[a_n H_{p-1}(i\Omega_n) - b_n J_{p+1}(i\Omega_n)\Big].$$

Turning now to (5.4c) and (5.4d), we find the fiber position to be

$$X_n^r = \frac{1}{2\nu\Omega_n^3}\Big[a_n H_{p-1}(i\Omega_n) + b_n J_{p+1}(i\Omega_n)\Big],$$
$$X_n^\theta = \frac{-1}{2i\nu\Omega_n^3}\Big[a_n H_{p-1}(i\Omega_n) - b_n J_{p+1}(i\Omega_n)\Big].$$

We can solve for a_n and b_n from these equations to find

$$a_n = \frac{\nu\Omega_n^3}{H_{p-1}(i\Omega_n)}\left(X_n^r - iX_n^\theta\right),$$
$$b_n = \frac{\nu\Omega_n^3}{J_{p+1}(i\Omega_n)}\left(X_n^r + iX_n^\theta\right).$$

We now substitute these expressions into the jump conditions (5.4e)–(5.4f) to get the following system of equations:

$$\begin{aligned} 0 = {} & i\left\{\phi\ \Omega_n^3\left[\frac{H_p(i\Omega_n)}{H_{p-1}(i\Omega_n)} - \frac{J_p(i\Omega_n)}{J_{p+1}(i\Omega_n)}\right] + ip\right\} X_n^r \\ & + \left\{\phi\ \Omega_n^3\left[\frac{H_p(i\Omega_n)}{H_{p-1}(i\Omega_n)} + \frac{J_p(i\Omega_n)}{J_{p+1}(i\Omega_n)}\right] - ip^2\right\} X_n^\theta \\ & + i\tau p\left(X_{n-1}^r - X_{n+1}^r\right) - \tau p^2\left(X_{n-1}^\theta - X_{n+1}^\theta\right), \end{aligned} \tag{6.1}$$

$$
\begin{aligned}
0 = i\left\{\phi\ \Omega_n^4\left[2 - \frac{H_{p+1}(i\Omega_n)}{H_{p-1}(i\Omega_n)} - \frac{J_{p-1}(i\Omega_n)}{J_{p+1}(i\Omega_n)}\right] + 2p(p^2-1)\right\} X_n^r \\
- \phi\ \Omega_n^4\left[\frac{H_{p+1}(i\Omega_n)}{H_{p-1}(i\Omega_n)} - \frac{J_{p-1}(i\Omega_n)}{J_{p+1}(i\Omega_n)}\right] X_n^\theta \\
+ 2\tau p(p^2-1)\left(X_{n-1}^r - X_{n+1}^r\right),
\end{aligned} \tag{6.2}
$$

where we define ϕ as the grouping $\phi = \nu^2/\kappa$. It is these last two equations, (6.1) and (6.2), that determine the coefficients X_n^r and X_n^θ in the case when $(\gamma + in) \neq 0$.

6.2. Case $(\gamma + in) = 0$. For this case, the vorticity equation becomes

$$r^2\xi_n''(r) + r\xi_n'(r) - p^2\xi_n(r) = 0,$$

whose appropriate solution is

$$\xi_o(r) = \begin{cases} b_o r^p & \text{if } r < 1 \qquad \text{(inner)}, \\ a_o r^{-p} & \text{if } r > 1 \qquad \text{(outer)}, \end{cases}$$

while the corresponding solution for the stream function is

$$\psi_o(r) = \begin{cases} \dfrac{1}{4p}\left(b_o - \dfrac{a_o}{1-p}\right) r^p - \dfrac{b_o}{4(1+p)} r^{2+p} & \text{if } r < 1, \\ \dfrac{1}{4p}\left(\dfrac{b_o}{1+p} - a_o\right) r^{-p} - \dfrac{a_o}{4(1-p)} r^{2-p} & \text{if } r > 1. \end{cases}$$

One can easily check that $\psi_o(r)$ and $\psi_o'(r)$ are continuous across the boundary and that

$$\psi_o(1) = \frac{1}{4p}\left[\frac{a_o}{p-1} + \frac{b_o}{p+1}\right],$$
$$\psi_o'(1) = \frac{1}{4}\left[\frac{a_o}{p-1} - \frac{b_o}{p+1}\right].$$

Turning now to (5.4c) and (5.4d), we find that in this case, the above equations reduce to $\psi_o(1) = \psi_o'(1) = 0$, which implies that $a_o = b_o = 0$. This means that when $(\gamma + in) = 0$ we obtain the trivial solution $\xi_o(r) \equiv 0$ and $\psi_o(r) \equiv 0$. The jump conditions (5.4e) and (5.4f) then give the system of equations

$$0 = ip^2\left(X_n^\theta - i\tau X_{n-1}^\theta + i\tau X_{n+1}^\theta\right) + p\left(X_n^r - i\tau X_{n-1}^r + i\tau X_{n+1}^r\right), \tag{6.3}$$

$$0 = ip(p^2-1)\left(X_n^r - i\tau X_{n-1}^r + i\tau X_{n+1}^r\right), \tag{6.4}$$

which are used instead of (6.1)–(6.2) for the case $(\gamma + in) = 0$.

7. Natural modes for the unforced fiber. When there is no forcing applied to the fiber, then $\tau = 0$ and there is no need to look for a solution in the form of an infinite series. Instead, we can consider a single mode by taking $n = 0$, for which the problem reduces to the following single pair of equations for X^r and X^θ:

$$
\begin{aligned}
0 = i\left\{\phi\ \Omega_o^3\left[\frac{H_p(i\Omega_o)}{H_{p-1}(i\Omega_o)} - \frac{J_p(i\Omega_o)}{J_{p+1}(i\Omega_o)}\right] + ip\right\} X^r \\
+ \left\{\phi\ \Omega_o^3\left[\frac{H_p(i\Omega_o)}{H_{p-1}(i\Omega_o)} + \frac{J_p(i\Omega_o)}{J_{p+1}(i\Omega_o)}\right] - ip^2\right\} X^\theta,
\end{aligned} \tag{7.1}
$$

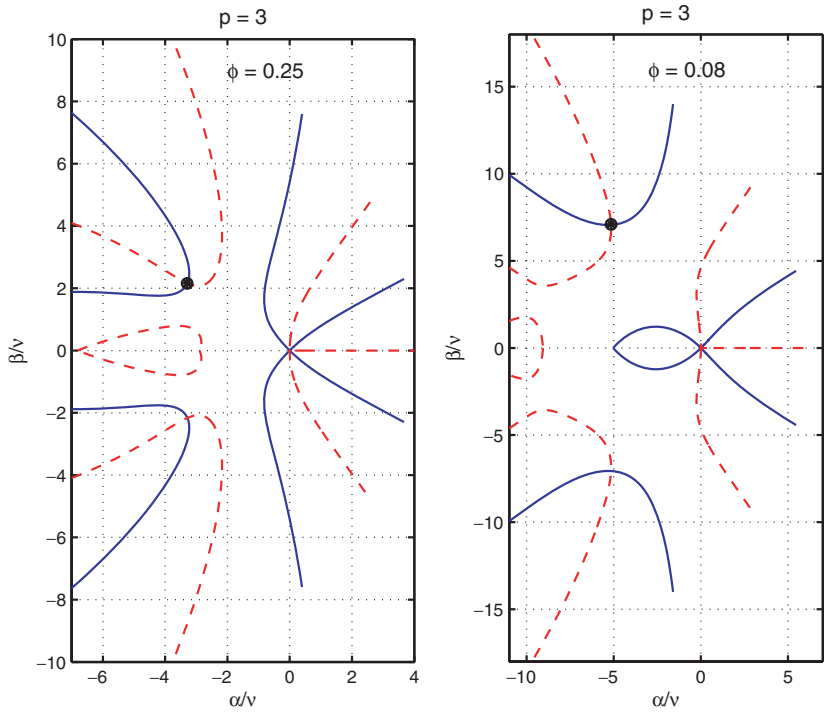


FIG. 7.1. *Zero-level contours of $\mathcal{D}(\Omega_o;\phi,p)$ for $p=3$ and two values of ϕ. The solid lines correspond to the real part and the dashed lines to the imaginary parts. Intersection points represent the natural modes of oscillation of the fiber.*

$$0 = i\left\{\phi\,\Omega_o^4\left[2-\frac{H_{p+1}(i\Omega_o)}{H_{p-1}(i\Omega_o)}-\frac{J_{p-1}(i\Omega_o)}{J_{p+1}(i\Omega_o)}\right]+2p(p^2-1)\right\}X^r \\ -\phi\,\Omega_o^4\left[\frac{H_{p+1}(i\Omega_o)}{H_{p-1}(i\Omega_o)}-\frac{J_{p-1}(i\Omega_o)}{J_{p+1}(i\Omega_o)}\right]X^\theta, \tag{7.2}$$

where $\Omega_o^2=\gamma/\nu=(\alpha/\nu)+i(\beta/\nu)$ and $\phi=\nu^2/\kappa$. Consequently, the above equations can be rewritten as a 2×2, homogeneous, linear system

$$\mathcal{M}\begin{pmatrix}X^r\\X^\theta\end{pmatrix}=0,$$

which has a nontrivial solution only if $\det(\mathcal{M})=0$. For notational convenience we define the function $\mathcal{D}(\Omega_o;\phi,p)=\det(\mathcal{M})$. The roots of $\mathcal{D}(\Omega_o;\phi,p)=0$, which is a *dispersion relation*, correspond to the natural modes of oscillation of the immersed boundary for a given set of parameters ϕ and p.

From the left plot of Figure 7.1, which corresponds to $p=3$ and $\phi=0.25$, a possible value of the natural response is $(\gamma/\nu)=(\alpha/\nu,\beta/\nu)\approx(-3.29,2.16)$. The right plot of Figure 7.1 shows that for $\phi=0.08$ the root is $(\alpha/\nu,\beta/\nu)\approx(-5.15,7.1)$ instead. These roots are in dimensionless form, and the dimensional frequency γ is obtained by multiplying these roots by $\mu/\rho R^2$. Note that in both cases, the natural mode of oscillation is stable (i.e., α is negative), as would be expected from an unforced immersed fiber (see [26]).

7.1. Collapse of the curves. The condition $\mathcal{D}(\Omega_o; \phi, p) = 0$ defines implicitly the root

$$\frac{\gamma}{\nu} = F(\phi, p) \tag{7.3}$$

as a function of the parameters. This suggests that the same root can be obtained with different values of damping ν and stiffness κ, provided the ratio $\phi = \nu^2/\kappa$ is held constant.

We corroborated this with the numerical solution of the nonlinear equations (3.2a)–(3.2e) for different combinations of (ν, κ) that yield the same value of ϕ. The blob projection method described in [4] was used in these calculations, which solves the IB problem using high-order regularized delta functions and fourth-order finite differences and time stepping. We have chosen to use this approach (rather than the IB method described in section 9) since the blob projection method affords higher accuracy, particularly when ϕ is small.

The initial condition of the fiber was the unit circle augmented by a single p-mode perturbation with amplitude ϵ, so that the fiber configuration can be written in polar coordinates as

$$r(\theta, t) = 1 + \epsilon B(t) \cos(p\theta) + O(\epsilon^2).$$

After each run, the amplitude $B(t)$ was assumed to have the form $B(t) = e^{\alpha t}[\cos(\beta t) + (\alpha/\beta)\sin(\beta t)]$ because this function has zero slope at $t = 0$ and it oscillates with decaying amplitude. The values of (α, β) can then be estimated numerically using least squares. These values were then compared with the roots of (7.3). This procedure was repeated for several values of ϕ with perturbed modes $p = 2$, 3, and 4, and the results are summarized in Figure 7.2.

The analytical results (plotted as solid lines) correspond very well with the numerical calculations (plotted as points). Two simulations corresponding to different combinations of ν and κ are performed for every ϕ. The observed values of α and β corresponding to a given ϕ are nearly identical, so that the points cannot be distinguished visually in Figure 7.2. Consequently, ϕ is an appropriate parameter for characterizing the fiber oscillations.

The correspondence between the analytical and computed results is very good for most values of the parameters, except for relatively large values of ϕ. This is mostly due to the fact that the blob projection method is implemented with periodic boundary conditions in a box about twice as large as the fiber's diameter. For large values of ϕ, the diffusion is significant over the time scale of the runs, and the dynamics are affected by the periodicity. Nonetheless, the results clearly show that different values of (ν, κ) that yield the same ϕ give the same root.

7.2. The small viscosity limit. We consider the limit of small viscosity by assuming a fixed value of the stiffness κ and small value of ν. It will be convenient to define a natural response frequency by $\omega_N \doteq \sqrt{p(p^2-1)/2}$, which is displayed in the frequency plot in Figure 7.2 as a dashed line for comparison with the analytical and computed results discussed in the preceding section. In the small viscosity case, (7.3), whose roots represent the dispersion relation, can be expanded as a series in powers of ν. The results show that the first few terms in the expansion of the natural response of the fiber are

$$\alpha = -\frac{p}{2\sqrt{2}}\omega_N^{1/2}\kappa^{1/4}\sqrt{\nu} + \frac{(p^3 - 3p^2 - p + 1)}{4}\nu + \cdots, \tag{7.4}$$

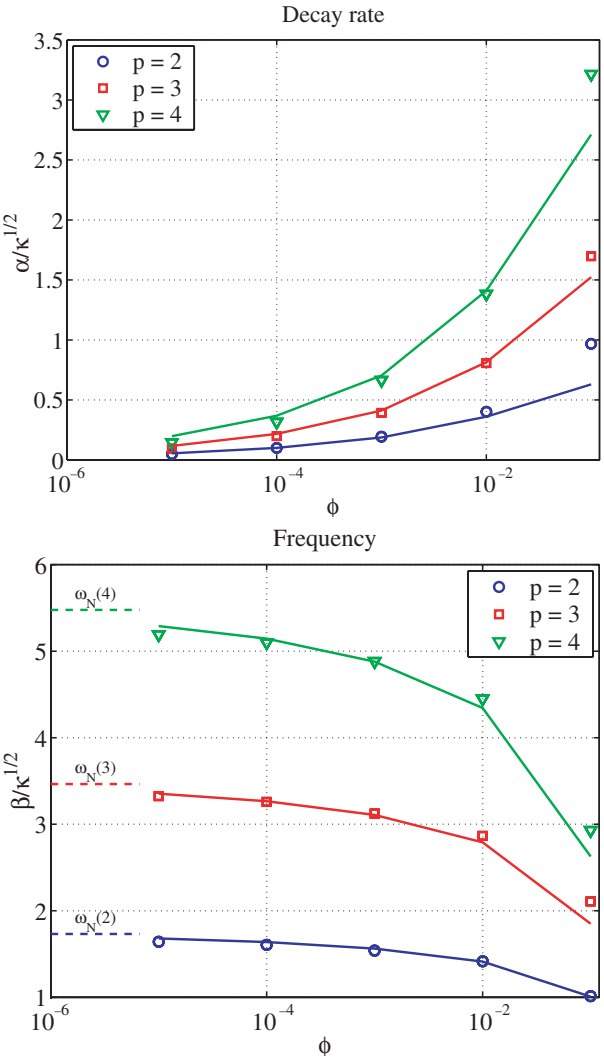


FIG. 7.2. *Roots $\gamma = \alpha + i\beta$ of the dispersion relation $\mathcal{D}(\Omega_o; \phi, p) = 0$, plotted as scaled values of α and β versus ϕ. The analytical results are displayed as solid lines, while the numerical approximations are plotted as points, with multiple points corresponding to simulations done using different values of κ and ν corresponding to the same ϕ. The results are presented for initial fiber perturbations corresponding to three different modes: $p = 2$, 3, and 4. The zero-viscosity ($\phi = 0$) limit, $\omega_N(p) = \sqrt{p(p^2-1)/2}$, is also shown for comparison.*

$$\beta = \omega_N\sqrt{\kappa} - \frac{p}{2\sqrt{2}}\omega_N^{3/4}\kappa^{1/4}\sqrt{\nu} - \frac{p(4p^3 - 39p^2 - 4p + 8)}{64\sqrt{2}}\omega_N^{1/4}\kappa^{-1/4}\nu^{3/2} + \cdots. \tag{7.5}$$

In the limit as $\nu \to 0$, these become $\alpha = 0$ and $\beta = \omega_N\sqrt{\kappa}$, which is precisely the linear dispersion relation found by other methods in [5] for a closed membrane in Euler flow.

It is also possible to compare the above expressions to the asymptotic expansions derived in [26] for an unforced, horizontal fiber immersed in a viscous fluid. The normal mode of oscillation for the flat fiber initialized with a p-mode obeys the following

to leading order:

$$\alpha \sim -2^{-7/4}\,\rho^{-3/4}\,\mu^{1/2}\,K_c^{1/4}\,p^{7/4}\,R^{-3/2} = -p\left(\frac{p^3}{2^7}\right)^{1/4}\left(\frac{\mu}{\rho R^2}\right)^{1/2}\left(\frac{K_c}{\rho R^2}\right)^{1/4},$$

$$\beta \sim 2^{-1/2}\,\rho^{-1/2}\,K_c^{1/2}\,p^{3/2}\,R^{-1} = \left(\frac{p^3}{2}\right)^{1/2}\left(\frac{K_c}{\rho R^2}\right)^{1/2}.$$

Substituting the nondimensional variables from (3.1) into (7.4) and (7.5), it is possible to show that the parameter dependence in the first term in each of α and β is identical to that in the expressions above, provided that we take $\omega_o \equiv 1$. We note that the natural frequency of the flat fiber is different from ω_N above due to the differences in the geometry of the problems. Furthermore, the linear theory for the circular fiber excludes the $p = 1$ mode, since to leading order this perturbation results only in a translation of the fiber.

8. Stability of the periodically forced fiber. In Floquet stability theory, the standard approach is to determine solutions with $\mathrm{Re}(\gamma) = \alpha = 0$, which correspond to the boundary of the stability region. In general, we expect the coefficients in (6.1) and (6.2) to decrease in magnitude as n increases, and so it is reasonable to truncate the series expansions at some finite number of terms, say, $-N \le n \le N$ (this is an assumption that will be checked later on). Equations (6.1) and (6.2) are written for $n = \pm 1, \pm 2, \ldots, \pm N$, and (6.3) and (6.4) are written for $n = 0$, which results in a linear system of dimension $(4N + 4) \times (4N + 4)$ for the unknown coefficients X_n^r and X_n^θ.

8.1. The reality condition. In general, the coefficients arising from the solution to the linear system are complex-valued, but the position of the fiber must be a real quantity. We therefore need to impose additional constraints to guarantee that the Floquet expansion

$$X^r(t) = \sum_n X_n^r e^{int}$$

is real. This will also allow the series to be written for positive index only. Consequently,

$$\overline{X}^r(t) = \sum_n \overline{X}_n^r e^{-int} = \sum_n \overline{X}_{-n}^r e^{int} = \sum_n X_n^r e^{int} = X^r(t),$$

which implies that

$$X_{-n}^r = \overline{X}_n^r \quad \text{for every } n.$$

This is the called the *reality condition* (see [11]), and it allows us to consider (6.1) and (6.2) for strictly positive values of n, thereby eliminating the coefficients for $n < 0$. The condition also implies that $X_o^r = \overline{X}_o^r$, for use in (6.3) and (6.4) with $n = 0$. The same reality condition applies to X_n^θ.

8.2. Formulation as an eigenvalue problem. Equations (6.1)–(6.2) now take the form

$$0 = A_n X_n^r + B_n X_n^\theta + \tau\left[C_n X_{n-1}^r + D_n X_{n-1}^\theta + E_n X_{n+1}^r + F_n X_{n+1}^\theta\right].$$

By setting the real part and the imaginary part of the equation to zero independently, we can write the final system as

$$(\mathcal{D} + \tau\, \mathcal{C})\boldsymbol{v} = 0,$$

where $\mathcal{D}$ and $\mathcal{C}$ are *real-valued* matrices, each of dimension $(4N+4)\times(4N+4)$. The vector is $\boldsymbol{v} = [v_0, v_1, v_2, \ldots, v_N]^T$, where each component has four elements, $v_n = [\mathrm{Re}(X_n^r), \mathrm{Im}(X_n^r), \mathrm{Re}(X_n^\theta), \mathrm{Im}(X_n^\theta)]$, with the difference now being that the reality condition ensures that all solution components are real values. The matrices $\mathcal{D}$ and $\mathcal{C}$ have the following block form:

$$\mathcal{D} = \begin{pmatrix} D_0 & 0 & & \ldots & 0 \\ 0 & D_1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & & 0 & D_{N-1} & 0 \\ 0 & \ldots & & 0 & D_N \end{pmatrix},$$

$$\mathcal{C} = \begin{pmatrix} 0 & C_{01} & & \ldots & 0 \\ C_{10} & 0 & C_{12} & \ldots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & & C_{N-1,N-2} & 0 & C_{N-1,N} \\ 0 & \ldots & & C_{N,N-1} & 0 \end{pmatrix},$$

where each element shown is a 4×4 matrix. The first row of each matrix corresponds to $n = 0$ and is derived from (6.3)–(6.4).

An eigenvalue problem is formed by rewriting the system as

$$-\mathcal{D}^{-1}\mathcal{C}\boldsymbol{v} = \frac{1}{\tau}\,\boldsymbol{v}, \tag{8.1}$$

where the eigenvalue $1/\tau$ must be real. Since the matrices have real entries, all eigenvalues are either real or occur in complex conjugate pairs.

For a given value of the wavenumber p, as well as parameters ν and κ from (3.1), the eigenvalue equation (8.1) yields a sequence of values for the forcing amplitude τ. In practice, we have found that choosing $N = 60$ terms in the series expansions is sufficient to ensure that the neglected coefficients are small (that is, the computed eigenvalues do not change appreciably when N is taken any larger than 60). The choices $\gamma = 0$ or $\gamma = \frac{1}{2}i$ are known as the harmonic and subharmonic cases, respectively, and any complex value of τ is discarded. These are the two cases that correspond to real Floquet multipliers, $e^{2\pi\gamma} = e^{2\pi(\alpha+i\beta)}$. When $0 < \beta < \frac{1}{2}$ on the other hand, the Floquet multipliers are complex and always correspond to solutions that are damped; hence, they are of no interest in our stability analysis.

The resulting harmonic and subharmonic eigenvalues correspond to physical modes of oscillation of the fiber which are *marginally stable*. We can then vary the wave number p, and produce a plot of each real value of τ, which traces out the boundary of the stability region of the linearized problem in parameter space. Furthermore, if we concentrate on the range $\tau \le \frac{1}{2}$, we need only consider stability boundaries that drop below the curve $\tau = \frac{1}{2}$.

Figures 8.1–8.3 depict the stability regions for various parameter values, with both harmonic (H) and subharmonic (S) modes shown. The stability boundaries separate

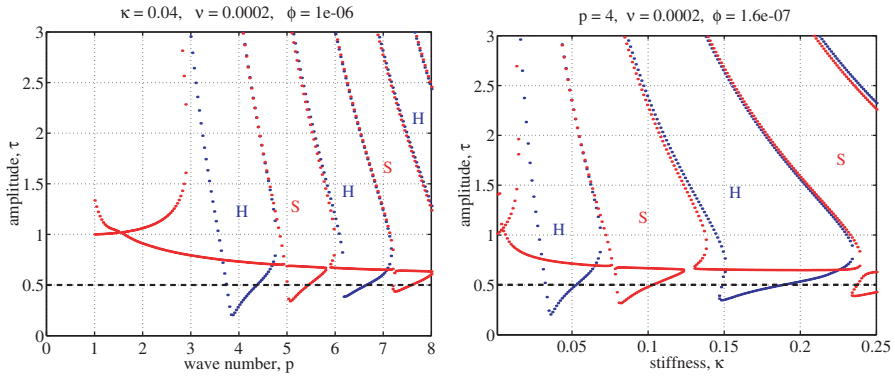


FIG. 8.1. *Stability diagrams, depicting plots of the fiber amplitude τ as a function of (left) wave number p, holding $\kappa = 0.04$; and (right) stiffness κ, holding $p = 4$. The other parameter values are $\nu = 0.0002$ and $\alpha = 0$. The curves traced out by the individual points represent the stability boundaries, and the regions above and inside each "tongue" correspond to unstable oscillations of the linearized IB problem. Regions of instability corresponding to harmonic modes (with $\gamma = 0$) are denoted "H" and are drawn with blue points, while the subharmonic modes (with $\gamma = \frac{1}{2}i$) are denoted "S" and are drawn with red points. The dashed horizontal line corresponds to $\tau = \frac{1}{2}$, and only portions of the tongues lying below this line correspond to oscillations with positive stiffness.*

parameter space into regions where the solution is stable and regions where it is unstable. Because of their distinctive shape, the regions of instability are usually referred to as "tongues." The unstable tongues alternate between harmonic and subharmonic modes, moving from left to right, with no overlap between the successive tongues.

Figure 8.1 shows two views of the stability regions for $\nu = 0.0002$: the first in the p, τ-plane with κ held constant at 0.04, and the second in the κ, τ-plane with $p = 4$. In both views, the tongue-like structure of the stability regions is apparent. The first unstable mode occurring for $\kappa = 0.04$ is a $p = 4$ mode, which corresponds to the left-most tongue that falls below the line $\tau = \frac{1}{2}$. Only tongues that correspond to integer values of the angular wavenumber p are physical. Notice in the right plot that increasing the stiffness has a stabilizing influence in the sense that the tongues migrate upward as κ is increased and therefore require higher-amplitude forcing in order to generate parametric resonance.

Figure 8.2 demonstrates more clearly the influence of changes in the stiffness parameter, by depicting the stability boundaries in the p, τ-plane for $\kappa = 0.02$, 0.04, and 0.08. As κ is increased (corresponding to a stiffer fiber), the regions of instability move towards the left, causing the wavenumber of the first unstable mode to decrease, periodically moving in and out of the "physical" regime.

The effect of changes in viscosity for constant stiffness is investigated in Figure 8.3, where plots for $\nu = 0.00005$, 0.0002, and 0.001 are given for $\kappa = 0.04$. As viscosity increases, the unstable regions migrate vertically upwards, so that the minimum value of τ corresponding to a linear instability increases, and some modes that were unstable no longer lead to parametric resonance. As a result, larger-amplitude forcing is necessary to cause onset of parametric resonance in fibers with larger viscosity. If ν is taken large enough that all tongues lift above the $\tau = \frac{1}{2}$ line, then the system is no longer subject to parametric resonance. These results are evidence of the fact that viscosity has a stabilizing influence on the system.

In the limit as $\nu \to 0$, the tongues extend down to the p-axis, where they touch the

line $\tau = 0$, corresponding to unforced oscillations of the immersed fiber. Therefore, the points where the tongues touch down in the $\nu = 0$ limit represent the natural modes of oscillation which were discussed in detail in section 7.

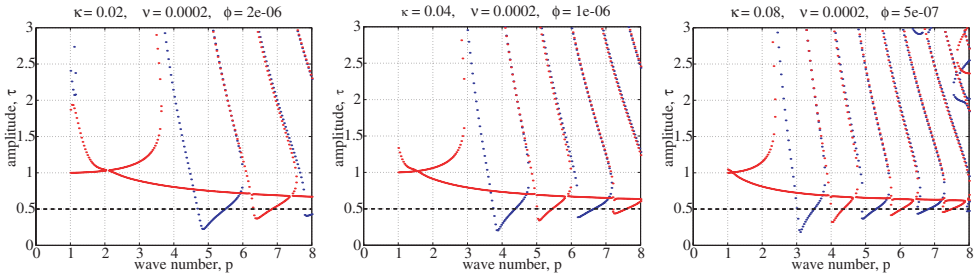


FIG. 8.2. *A series of three plots showing the impact of changes in the stiffness on the stability boundaries in the* p, τ*-plane, with viscosity* $\nu = 0.0002$ *and three different values of stiffness:* $\kappa = 0.02$ *(left),* 0.04 *(middle), and* 0.08 *(right).*

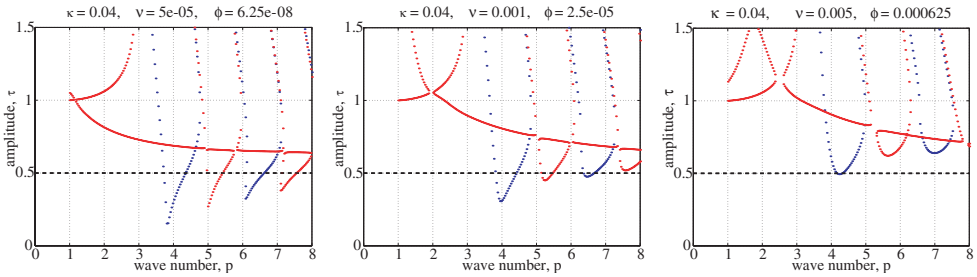


FIG. 8.3. *A series of three plots showing the impact of changes in viscosity in the stability boundaries in the* p, τ*-plane, with stiffness* $\kappa = 0.04$ *and three different values of viscosity:* $\nu =$ 0.00005 *(left),* 0.001 *(middle), and* 0.005 *(right).*

9. Comparison with IB computations. In this section, we use several immersed boundary computations to demonstrate the validity of the preceding Floquet analysis. The numerical method is based on a straightforward discretization of the IB equations in terms of velocity and pressure variables. An alternating direction implicit (ADI) approach is used to apply the convection, diffusion, and forcing terms to obtain an intermediate velocity field. The resulting velocity is then made divergence-free through the use of a split-step pressure projection procedure. The standard cosine approximation to the Dirac delta function is employed [20], which is smoothed over a square box with side length of four fluid grid points. The resulting method is second-order accurate in space, except for the approximate delta function interpolation which limits the spatial accuracy to first order. The method is explicit and has first-order accuracy in time. There are many variants of the IB method that increase both spatial and temporal accuracy, but we have chosen instead to demonstrate the presence of parametric resonance using this simplest and most common implementation. For complete details of the numerical technique, refer to [20] or [24].

All computations were performed on an immersed boundary whose rest configuration is a circle, immersed in a periodic box of dimension 2.5 times the size of the circular boundary. The analysis strictly applies only to an infinite fluid domain, but we found that this domain size was large enough in practice to avoid significant

TABLE 9.1
Parameters for the two resonant cases, one with a first unstable mode with angular wavenumber $p = 2$, and the other with $p = 4$.

Case I	Case II
$\kappa = 0.5$ $\nu = 0.004$ $\phi = 3.2 \times 10^{-5}$	$\kappa = 0.04$ $\nu = 0.00056$ $\phi = 7.84 \times 10^{-6}$
$\rho = 1$ $\mu = 0.4$ $R = 0.2$ $K_c = 125000$ $\omega_o = 2500$ $\tau = 0.45$ $p = 2$ unstable	$\rho = 1$ $\mu = 0.5$ $R = 1$ $K_c = 40000$ $\omega_o = 900$ $\tau = 0.45$ $p = 4$ unstable

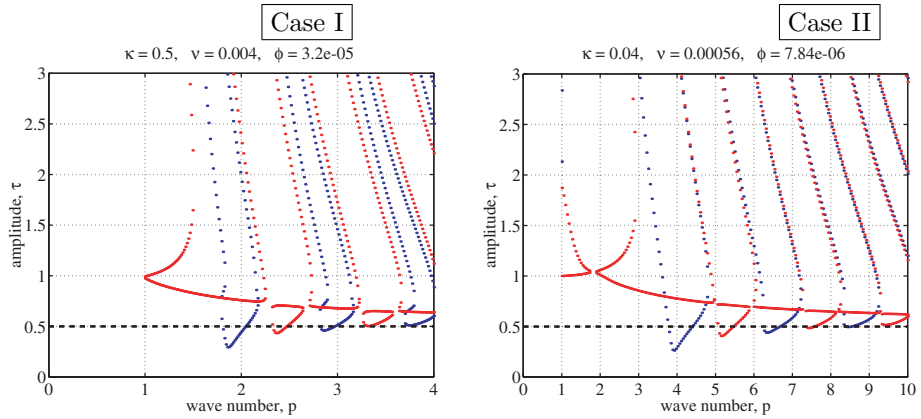


FIG. 9.1. *Stability diagrams for Case* I *($\kappa = 0.5$, $\nu = 0.004$) and Case* II *($\kappa = 0.04$, $\nu = 0.00056$).*

interference from neighboring periodic copies. The fluid domain is divided into a 64×64 grid with 192 immersed boundary points, ensuring that there is significant resolution of the boundary within each fluid grid cell. Finer grid computations were also performed with a 128×128 fluid grid and 384 fiber points to validate the results. The time step was selected to be well within the stability restriction imposed by the explicit method.

We chose to focus on two specific sets of parameters for which the stability plots suggested different resonant p-modes. The parameters are listed in Table 9.1, and the corresponding stability regions are displayed in Figure 9.1. In Case I, the first unstable tongue corresponds to a harmonic mode of oscillation at $p = 2$, while the lowest-wavenumber unstable mode for the second case is $p = 4$ (also harmonic).

We next present numerical evidence that supports the existence of these two instances of parametric resonance. In both cases, the immersed boundary is initially in the shape of a circle of radius R with a radial perturbation of the form $r = R(1 + 0.05\cos(p\theta))$. The results of a given simulation are reported as plots of $\hat{r}_p(t)$, which represents the amplitude of the p-mode of oscillation in the fiber versus time, and is calculated as follows:

1. we convert the (x, y) position of each point on the immersed fiber to radial coordinates;

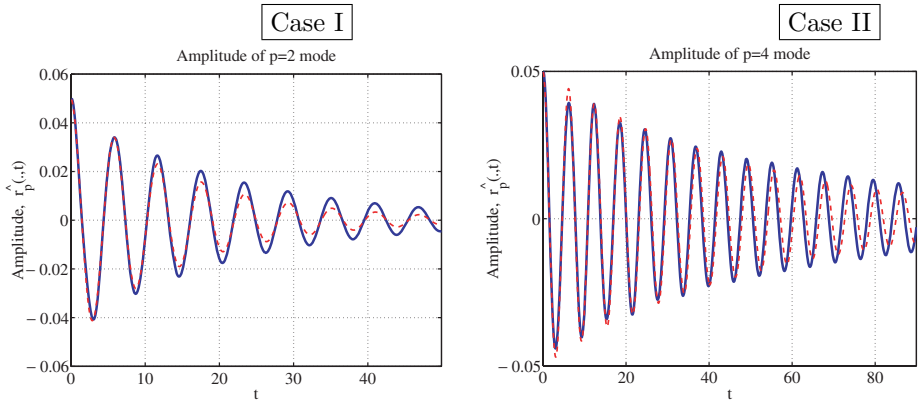


FIG. 9.2. *Amplitude of the unforced immersed boundary for Case* I *($p = 2$) and Case* II *($p = 4$). The dashed curve is an approximate fit to the computed results using a function of the form $e^{\alpha t}\cos(\beta t)$.*

2. using cubic splines, we interpolate the points representing the fiber positions onto a set of points that are equally spaced in θ;
3. we calculate $\hat{r}_p(t) = FFT_\theta(r(\theta, t))$, the fast Fourier transform of the radius in θ, which then yields the amplitude of the desired p-mode.

First, we present plots of the unforced solution (with $\tau = 0$) in Figure 9.2 that represent the natural mode of oscillation for the fiber in each case. The rate of decay and frequency of the natural modes of oscillation for the unforced fiber are as follows:

$$\text{Case I:} \quad \alpha = -0.066, \qquad \beta = 1.071,$$
$$\text{Case II:} \quad \alpha = -0.020, \qquad \beta = 1.016.$$

When the immersed boundary is then forced internally at a frequency equal to the resonant frequency suggested by the plots in Figure 9.1 (i.e., $\omega_o = 2500$ in Case I and $\omega_o = 900$ in Case II), amplitude of oscillation grows far beyond the initial amplitude of 0.05 cm. The motion never actually becomes unstable, but the fiber instead exhibits sustained, large-amplitude oscillations (see Figure 9.3 and compare to Figure 9.2).

In order to verify that this behavior is truly arising from a parametric resonance, we consider changes to either κ or ν that move the resonant tongue in the eigenvalue plot outside the range of parameters being considered. The results are summarized in Figures 9.4 and 9.5. the As the viscosity is increased, the oscillations either decrease in amplitude or decay in time, though not at as rapid a rate as for the unforced case. Because an increase in viscosity acts to raise the "tongues" in the eigenvalue plot, this is precisely the behavior we would expect.

Alternately, if we increase or decrease the value of κ in relation to the resonant value, the resulting oscillations are pictured in Figure 9.5. The amplitude of oscillation for the resonant mode has a maximum value close to the resonant value of κ, which represents the movement of the resonant tongues either to the left or to the right as κ is increased or decreased, respectively.

The effect of varying the stiffness perturbation amplitude τ is displayed in Figure 9.6 for Case II. For $\tau \leq 0.40$, there is no longer a sustained oscillation in the fiber, and as τ is reduced the amplitude of the fiber motion decreases.

At this point, it is important to emphasize that in our numerical simulations, resonance is indicated by sustained, large-amplitude motions, rather than any actual

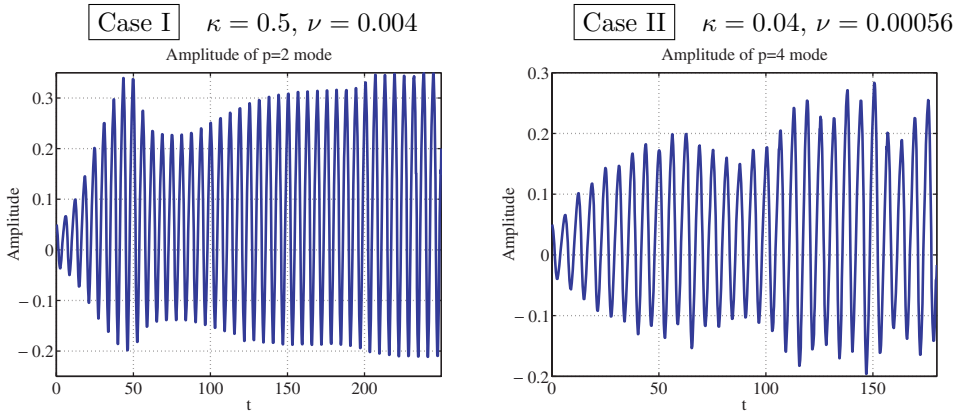


FIG. 9.3. *Amplitude of the resonant p-mode, when the immersed boundary is forced at the resonant frequency ($p = 2$ for Case* I *and $p = 4$ for Case* II*). Both cases display a sustained, large-amplitude oscillation.*

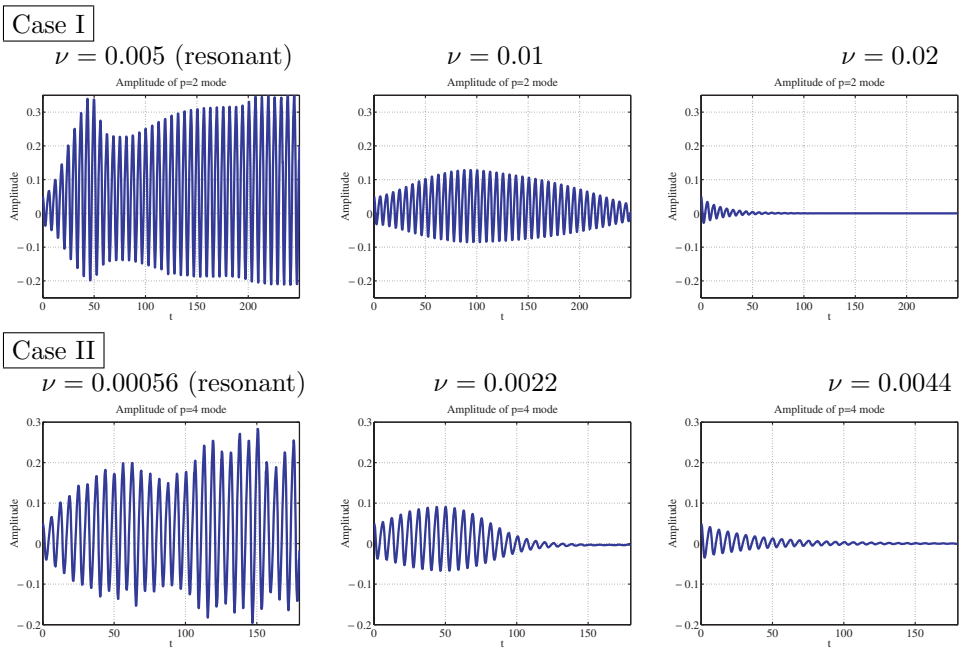


FIG. 9.4. *Amplitude of the resonant p-mode in Cases* I *and* II*, as ν is increased. The stabilizing influence of viscosity is exhibited by the inability of the fiber to sustain large-amplitude oscillations for even small increases in the viscosity.*

instability (i.e., unbounded oscillations). This discrepancy arises from both numerical errors and simplifications to the model, namely:

- the numerical scheme is only first order in space and time and introduces a significant level of artificial viscosity;
- although the fiber force term in the linearized Navier–Stokes equations was also linearized in the forgoing analysis, it is actually nonlinear, which acts to

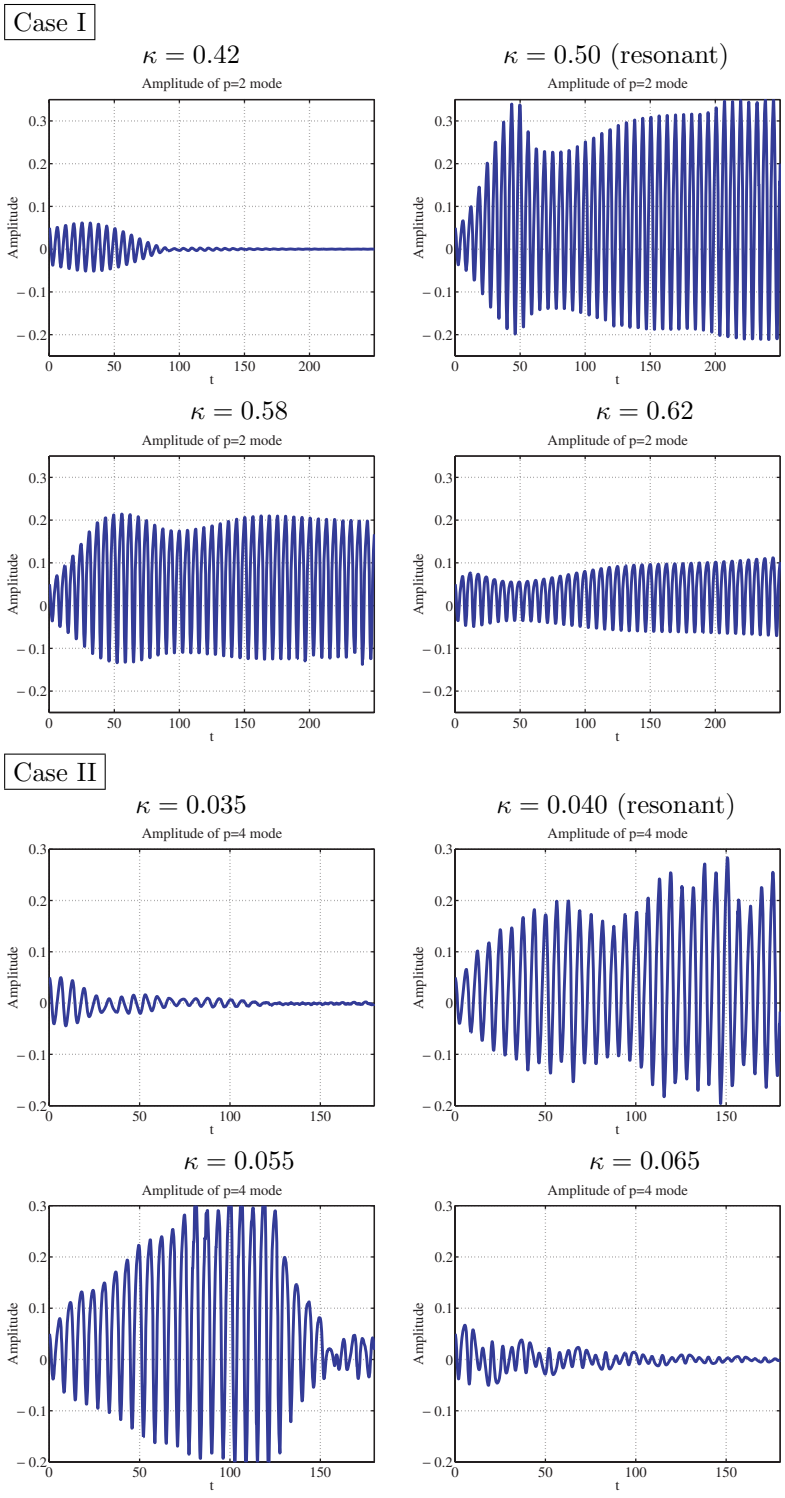


FIG. 9.5. *Amplitude of the resonant p-mode in Cases* I *and* II, *as* κ *is varied. When the stiffness is taken either smaller or larger than the resonant value, the amplitude of the oscillations decreases to the point that they can no longer be sustained.*

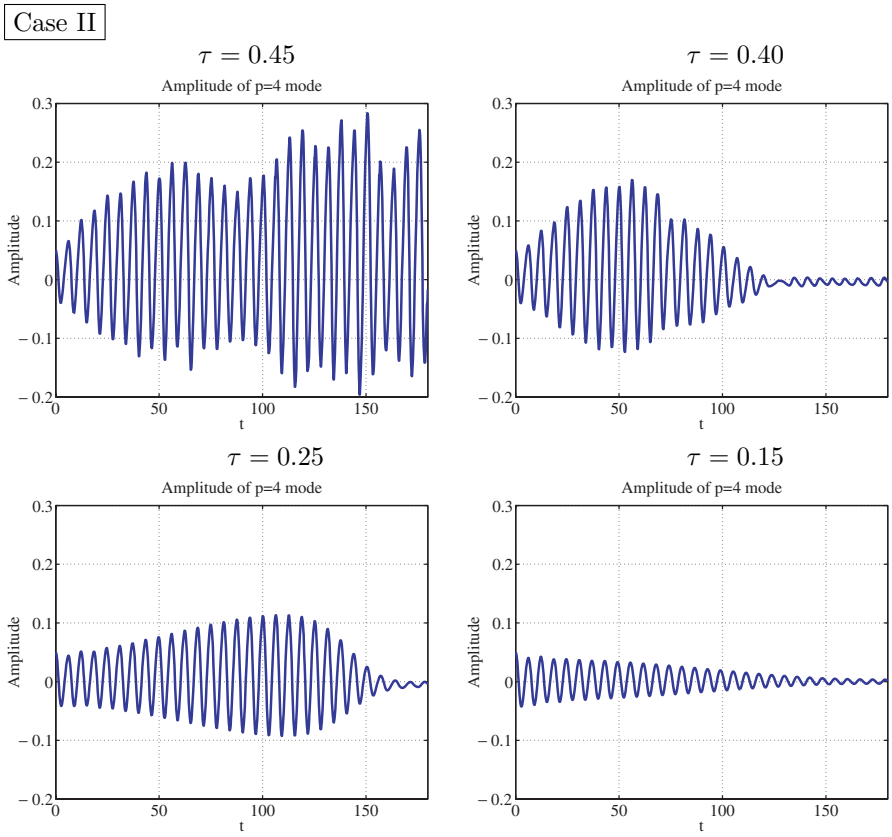


FIG. 9.6. *Amplitude of the resonant p-mode in Case* II*, as the forcing amplitude, τ, is varied. As the forcing is reduced, the amplitude of the fiber oscillations also diminishes.*

stabilize the numerical results;

- the analysis assumed an infinite fluid domain, while our numerical simulations use periodic boundary conditions. Discrepancies owing to interference from periodic copies of the immersed fiber are therefore unavoidable, though we have attempted to choose the size of our computational domain large enough so that these errors are minimized.

A further symptom of these errors is the fact that the stability boundaries in parameter space demonstrated in the simulations are not nearly as sharp, or located in exactly the same locations, as indicated by the plots in section 8. Nonetheless, the correspondence between analytical and numerical results is still quite convincing evidence of the presence of parametric resonance in the linearized IB problem.

The final set of results indicates what transpires when a mode other than the resonant mode is excited initially. We restrict ourselves to Case II and initialize the fiber position with modes having wavenumber $p = 2$ or $p = 3$, while still forcing the stiffness through a $p = 4$ mode. The other parameters remain the same as in the resonant case. Figure 9.7 shows that in both cases, the given 2- and 3-modes do not grow; however, energy transfers over time into the resonant $p = 4$ mode which eventually dominates the fiber oscillation.

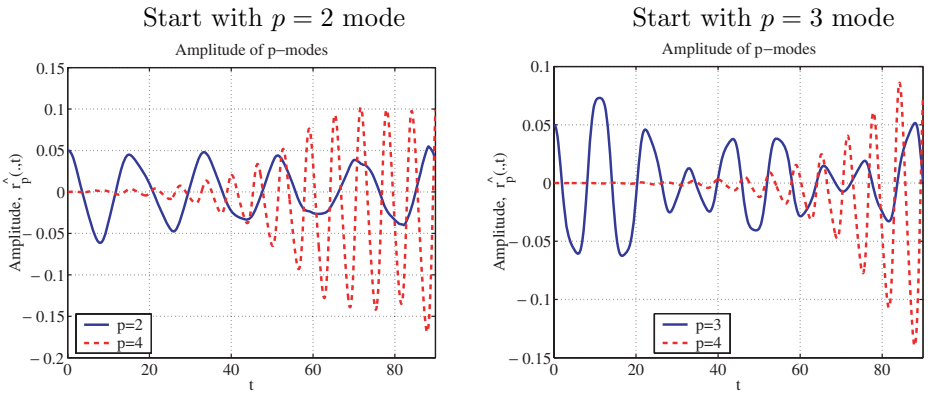


FIG. 9.7. *The amplitude of various fiber modes in Case* II *when wavenumbers* $p = 2$ *and* $p = 3$ *are excited initially instead of the* $p = 4$ *mode. Numerical errors give rise to phase-shifted oscillations which, when combined linearly with the initial conditions, perturb the* $p = 4$ *mode and hence feed energy into the resonant mode over time.*

10. Conclusions. Floquet analysis has proven to be an extremely useful tool in examining parametric resonance and the stability behavior of a wide variety of flows involving interfaces and fluid-structure interaction. In this paper, we study parametric resonances arising from an elastic membrane immersed in a viscous fluid in two dimensions, which is driven by periodic variations in the stiffness parameter of the elastic material. The underlying mathematical model, known as the immersed boundary formulation, captures not only the flow-induced deformations of the elastic membrane but also the influence of the immersed structure on the surrounding fluid flow. To our knowledge, this is the first study of its kind that captures this two-way interaction between fluid and fiber in a parametrically forced system.

Our Floquet analysis leads to an eigenvalue problem that can be solved in order to determine the stability boundaries in parameter space that separate the regions in which the motion is stable from those in which it is unstable. Using asymptotic expansions of the resulting solutions, we demonstrate that our results are consistent with previous analyses of unforced immersed fibers. The decay rates and frequencies of oscillation for the forced system are also shown to match closely those found in full numerical simulations of the fluid-fiber system for small-wavenumber perturbations. We also present numerical results that verify the existence of resonances in parametrically forced immersed boundaries.

This study opens the door for several avenues of further investigation. First of all, while we have demonstrated the existence of parametric resonances numerically in periodically forced immersed fibers, we have yet to find a biological system in which the parameters lie within the unstable regime. In the heart, for example, the muscle fiber stiffness appears to be too small to lead to parametric resonance, according to our analysis. However, we intend to investigate other biological systems with different parameter ranges to determine if resonances are possible.

There are also several natural extensions to the analysis that would allow us to investigate much more interesting fiber dynamics. For example, introducing a spatial dependence in the stiffness, $K(s,t)$, would better mimic biological systems in which an active fiber is pulsed via a wave of contraction that travels around the fiber. However, this form of the stiffness complicates the Floquet analysis significantly by coupling the various fiber modes, and hence would require an extension of our analytical technique.

We also intend to investigate the use of optimal control to see whether it is possible to eliminate parametric resonances by introducing an additional periodic forcing term in the system.

Appendix A. Some useful Bessel function formulas. The following identities are taken from [27] and [1]:

$$J_p(z) = \frac{z}{2p}(J_{p-1}(z) + J_{p+1}(z)), \tag{A.1}$$

$$J_p'(z) = \frac{1}{2}(J_{p-1}(z) - J_{p+1}(z)), \tag{A.2}$$

$$\int z^{p+1} J_p(az)dz = \frac{z^{p+1}}{a} J_{p+1}(az), \tag{A.3}$$

$$\int z^{1-p} J_p(az)dz = \frac{-z^{1-p}}{a} J_{p-1}(az). \tag{A.4}$$

Equations (A.1)–(A.4) are written for the Bessel function of the first kind, $J_p(z)$, but are also valid for the various other Bessel functions, $Y_p(z)$, $H_p(z)$, etc.

Appendix B. Proof of Claim 1.

CLAIM 1. *Given the expansions in* (3.3a) *and* (3.3b)*,*

$$(\hat{\boldsymbol{z}} \cdot \nabla \times \boldsymbol{f}^{(0)}) = 0 \qquad \textit{and}$$
$$(\hat{\boldsymbol{z}} \cdot \nabla \times \boldsymbol{f}^{(1)}) = K\left(X^{\theta}_{ss} + X^{r}_{s}\right)\left(\frac{\delta(r-1)}{r}\right)_r - K\left(X^{r}_{sss} - X^{\theta}_{ss}\right)\frac{\delta(r-1)}{r}.$$

Proof. Since

$$\begin{aligned}(\hat{\boldsymbol{z}} \cdot \nabla \times \boldsymbol{f}) &= -\hat{\boldsymbol{z}} \cdot \int_0^{2\pi} (K\boldsymbol{X}_s)_s \times \nabla\delta(\boldsymbol{x} - \boldsymbol{X})\, ds \\ &= -\int_0^{2\pi} [\hat{\boldsymbol{z}} \times (K\boldsymbol{X}_s)_s] \cdot \nabla\delta(\boldsymbol{x} - \boldsymbol{X})\, ds,\end{aligned}$$

we have that (3.3a)–(3.3b) imply that

$$\begin{aligned}(\hat{\boldsymbol{z}} \cdot \nabla \times \boldsymbol{f}^{(0)}) &= -\int_0^{2\pi} [\hat{\boldsymbol{z}} \times (K(t)\hat{\boldsymbol{r}}_s)_s] \cdot \nabla\delta(\boldsymbol{x} - \boldsymbol{X}(s,t))\, ds \\ &= -K(t)\int_0^{2\pi} [\hat{\boldsymbol{z}} \times \hat{\boldsymbol{r}}_{ss}] \cdot \nabla\delta(\boldsymbol{x} - \boldsymbol{X}(s,t))\, ds \\ &= -K(t)\int_0^{2\pi} \frac{d}{ds}\delta(\boldsymbol{x} - \boldsymbol{X}(s,t))\, ds \; = 0.\end{aligned}$$

We also have that

$$\begin{aligned}(\hat{\boldsymbol{z}} \cdot \nabla \times \boldsymbol{f}^{(1)}) &= -\int_0^{2\pi} \left[\hat{\boldsymbol{z}} \times (K\boldsymbol{X}^{(1)}_s)_s\right] \cdot \nabla\delta(\boldsymbol{x} - \boldsymbol{X}^{(0)})\, ds \\ &\quad + \int_0^{2\pi} \left\{\left[\hat{\boldsymbol{z}} \times (K\boldsymbol{X}^{(0)}_s)_s\right] \cdot \nabla\right\}\left(\boldsymbol{X}^{(1)} \cdot \nabla\right)\delta(\boldsymbol{x} - \boldsymbol{X}^{(0)})\, ds \\ &= I_1 + I_2.\end{aligned} \tag{B.1}$$

The second term can be simplified:

$$\begin{aligned} I_2 &= K\int_0^{2\pi}\left(\boldsymbol{X}^{(1)}\cdot\nabla\right)\left\{\left[\hat{\boldsymbol{z}}\times\boldsymbol{X}_{ss}^{(0)}\right]\cdot\nabla\right\}\delta(\boldsymbol{x}-\boldsymbol{X}^{(0)})\,ds \\ &= K\int_0^{2\pi}\left(\boldsymbol{X}^{(1)}\cdot\nabla\right)\frac{d}{ds}\delta(\boldsymbol{x}-\boldsymbol{X}^{(0)})\,ds \\ &= -K\int_0^{2\pi}\boldsymbol{X}_s^{(1)}\cdot\nabla\delta(\boldsymbol{x}-\boldsymbol{X}^{(0)})\,ds, \end{aligned}$$

so that so far we have that (B.1) is

$$(\hat{\boldsymbol{z}}\cdot\nabla\times\boldsymbol{f}^{(1)}) = -K\int_0^{2\pi}\left[\hat{\boldsymbol{z}}\times\boldsymbol{X}_{ss}^{(1)}+\boldsymbol{X}_s^{(1)}\right]\cdot\nabla\delta(\boldsymbol{x}-\boldsymbol{X}^{(0)})\,ds.$$

Now it is convenient to write $\boldsymbol{X}^{(1)}$ in polar coordinates

$$\boldsymbol{X}^{(1)} = X^r(s,t)\hat{\boldsymbol{r}}(s)+X^\theta(s,t)\hat{\boldsymbol{\theta}}(s),$$

so that

$$\hat{\boldsymbol{z}}\times\boldsymbol{X}_{ss}^{(1)}+\boldsymbol{X}_s^{(1)} = -\left(X_{ss}^\theta+X_s^r\right)\hat{\boldsymbol{r}}+\left(X_{ss}^r-X_s^\theta\right)\hat{\boldsymbol{\theta}}.$$

Now we can write

$$\begin{aligned} (\hat{\boldsymbol{z}}\cdot\nabla\times\boldsymbol{f}^{(1)}) &= K\int_0^{2\pi}\left[X_{ss}^\theta+X_s^r\right]\hat{\boldsymbol{r}}\cdot\nabla\delta(\boldsymbol{x}-\boldsymbol{X}^{(0)})\,ds \\ &\quad - K\int_0^{2\pi}\left[X_{ss}^r-X_s^\theta\right]\hat{\boldsymbol{\theta}}\cdot\nabla\delta(\boldsymbol{x}-\boldsymbol{X}^{(0)})\,ds. \end{aligned}$$

The last term can be integrated by parts and we arrive at

$$\begin{aligned} (\hat{\boldsymbol{z}}\cdot\nabla\times\boldsymbol{f}^{(1)}) &= \int_0^{2\pi}K\left(X_{ss}^\theta+X_s^r\right)\hat{\boldsymbol{r}}\cdot\nabla\delta(\boldsymbol{x}-\boldsymbol{X}^{(0)})\,ds \\ &\quad - \int_0^{2\pi}K\left(X_{sss}^r-X_{ss}^\theta\right)\delta(\boldsymbol{x}-\boldsymbol{X}^{(0)})\,ds. \end{aligned}$$

If we now write

$$\delta(\boldsymbol{x}-\boldsymbol{X}^{(0)}) = \frac{\delta(r-1)\delta(\theta-s)}{r},$$

we get

$$\begin{aligned} (\hat{\boldsymbol{z}}\cdot\nabla\times\boldsymbol{f}^{(1)}) &= \left[K\left(X_{ss}^\theta+X_s^r\right)\right](\theta,t)\left(\frac{\delta(r-1)}{r}\right)_r \\ &\quad - \left[K\left(X_{sss}^r-X_{ss}^\theta\right)\right](\theta,t)\,\frac{\delta(r-1)}{r}. \qquad \square \end{aligned}$$

Appendix C. Jump conditions. The derivation of the jump conditions for vorticity will make use of the following result.

CLAIM 2. *In the sense of distributions*

$$r\left(\frac{\delta(r-R)}{r}\right)' = \delta'(r-R) - \frac{\delta(r-R)}{R}.$$

Proof. Let $\phi(r)$ be a smooth function. Then

$$\begin{aligned}\int \phi(r)\, r\left(\frac{\delta(r-R)}{r}\right)' dr &= -\int \left(r\phi(r)\right)_r \left(\frac{\delta(r-R)}{r}\right) dr \\ &= -\int \left(\phi'(r) + \frac{\phi(r)}{r}\right)\delta(r-R)\, dr \\ &= -\phi'(R) - \frac{\phi(R)}{R}.\end{aligned}$$

The result follows. ∎

The jump conditions for an equation of the form

$$-\frac{1}{r}(r\xi')' + \left(\frac{(\gamma+in)}{\nu} + \frac{p^2}{r^2}\right)\xi = A\left(\frac{\delta(r-1)}{r}\right)' + B\left(\frac{\delta(r-1)}{r}\right),$$

where A and B are independent of r, can be derived as follows. We first multiply the equation by r:

$$-(r\xi')' + \left(\frac{(\gamma+in)}{\nu}r + \frac{p^2}{r}\right)\xi = Ar\left(\frac{\delta(r-1)}{r}\right)' + B\delta(r-1),$$

and use the claim to write it as

$$-(r\xi')' + \left(\frac{(\gamma+in)}{\nu}r + \frac{p^2}{r}\right)\xi = A\delta'(r-1) + (B-A)\delta(r-1).$$

We now integrate from $1-\epsilon$ to some point r to get

$$-\big[r\xi'(r) - (1-\epsilon)\xi'(1-\epsilon)\big] + \int_{1-\epsilon}^{r}\left(\frac{(\gamma+in)}{\nu}q + \frac{p^2}{q}\right)\xi(q)\, dq = A\delta(r-1) + (B-A)H(r-1). \tag{C.1}$$

We can use (C.1) in two ways. First, set $r = 1+\epsilon$ and take the limit $\epsilon \to 0$:

$$-[\![\xi']\!] = (B-A). \tag{C.2}$$

Second, we divide (C.1) by r, integrate from $1-\epsilon$ to $1+\epsilon$, and let $\epsilon \to 0$ to get

$$-[\![\xi]\!] = A. \tag{C.3}$$

Equations (C.2)–(C.3) are the two jump conditions.

Appendix D. Solution of the stream function equation. In this section we describe the method for finding the solution of the stream function equation

$$-\frac{1}{r}\left(r\psi_r\right)_r + \frac{p^2}{r^2}\psi = \xi(r),$$

where $\xi(r)$ is given by

$$\xi(r) = \begin{cases} bJ_p(i\Omega_n r) & \text{if } r < 1 \qquad \text{(inner)}, \\ aH_p(i\Omega_n r) & \text{if } r > 1 \qquad \text{(outer)}. \end{cases}$$

The general solution may be written as

$$\psi(r) = \int_0^\infty \xi(r)\mathcal{G}(r,z)dz.$$

Here, the Green's function $\mathcal{G}(r,z)$ is the solution of

$$-\frac{1}{r}\left(r\mathcal{G}_r\right)_r + \frac{p^2}{r^2}\mathcal{G} = \delta(r-z). \tag{D.1}$$

Multiplying this equation by r, integrating from $z-\epsilon$ to $z+\epsilon$, and taking the limit $\epsilon \to 0$, we find that $\mathcal{G}(r,z)$ satisfies the jump conditions

$$[\![\mathcal{G}]\!] = 0 \qquad \text{and} \qquad [\![\mathcal{G}_r]\!] = -1.$$

Since solutions of equation (D.1) are of the form r^p and r^{-p}, we have that

$$\mathcal{G}(r,z) = \frac{z}{2p}\begin{cases} \dfrac{r^p}{z^p} & \text{if } r < z, \\ \dfrac{z^p}{r^p} & \text{if } r > z. \end{cases}$$

The stream function is then found by piecewise integration. For the interior solution, $r < 1$, we obtain

$$\begin{aligned} \psi(r) &= \frac{b}{2pr^p}\int_0^r J_p(i\Omega_n r')(z)^{p+1}\,dz + \frac{br^p}{2p}\int_r^1 J_p(i\Omega_n r')(z)^{1-p}\,dz \\ &\quad + \frac{ar^p}{2p}\int_1^\infty H_p(i\Omega_n r')(z)^{1-p}\,dz \\ &= \frac{br}{2ip\Omega_n}\Big[J_{p-1}(i\Omega_n r) + J_{p+1}(i\Omega_n r)\Big] + \frac{r^p}{2ip\Omega_n}\Big[aH_{p-1}(i\Omega_n) - bJ_{p-1}(i\Omega_n)\Big], \end{aligned}$$

making use of the identities (A.3) and (A.4). The solution for $r > 1$ is found in the same way.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, John Wiley and Sons, New York, 1972.

[2] K. M. Arthurs, L. C. Moore, C. S. Peskin, E. B. Pitman, and H. E. Layton, *Modeling arteriolar flow and mass transport using the immersed boundary method*, J. Comput. Phys., 147 (1998), pp. 402–440.

[3] R. P. Beyer and R. J. LeVeque, *Analysis of a one-dimensional model for the immersed boundary method*, SIAM J. Numer. Anal., 29 (1992), pp. 332–364.

[4] R. Cortez and M. Minion, *The blob projection method for immersed boundary problems*, J. Comput. Phys., 161 (2000), pp. 428–453.

[5] R. Cortez and D. A. Varela, *The dynamics of an elastic membrane using the impulse method*, J. Comput. Phys., 138 (1997), pp. 224–247.

[6] S. M. Curry, *How children swing*, Amer. J. Phys., 44 (1976), pp. 924–926.

[7] R. Dillon and L. Fauci, *Microscale model of bacterial and biofilm dynamics in porous media*, Biotechnol. and Bioengrg., 68 (2000), pp. 536–547.
[8] L. J. Fauci, *A computational model of the fluid dynamics of undulator and flagellar swimming*, Amer. Zool., 36 (1996), pp. 599–607.
[9] P. Martin, A. D. Mehta, and A. J. Hudspeth, *Negative hair-bundle stiffness betrays a mechanism for mechanical amplification by the hair cell*, Proc. Nat. Acad. Sci. USA, 97 (2000), pp. 12026–12031.
[10] Y. Kim and C. S. Peskin, 2-*D parachute simulation by the immersed boundary method*, manuscript, Courant Institute of Mathematical Sciences, New York University, New York, 2001.
[11] K. Kumar and L. S. Tuckerman, *Parametric instability of the interface between two fluids*, J. Fluid Mech., 279 (1994), pp. 49–68.
[12] M.-C. Lai, *Simulations of the Flow Past an Array of Circular Cylinders as a Test of the Immersed Boundary Method*, Ph.D. thesis, New York University, New York, 1998.
[13] M.-C. Lai and Z. Li, *A remark on jump conditions for the three-dimensional Navier-Stokes equations involving an immersed moving membrane*, Appl. Math. Lett., 14 (2001), pp. 149–154.
[14] O. Lioubashevski, J. Fineberg, and L. S. Tuckerman, *Scaling of the transition to parametrically driven surface waves in highly dissipative systems*, Phys. Rev. E (3), 55 (1997), pp. 3832–3835.
[15] G. McKay, *Onset of double-diffusive convection in a saturated porous layer with time-periodic surface heating*, Contin. Mech. Thermodyn., 10 (1998), pp. 241–251.
[16] L. A. Miller, *A mathematical model of insect flight: The immersed boundary method with fling*, Amer. Zool., 40 (2000), p. 1133.
[17] H. W. Müller, H. Wittmer, C. Wagner, J. Albers, and K. Knorr, *Analytic stability theory for Faraday waves and the observation of the harmonic surface response*, Phys. Rev. Lett., 78 (1997), pp. 2357–2360.
[18] A. H. Nayfeh and D. T. Mook, *Nonlinear Oscillations*, John Wiley and Sons, New York, 1979.
[19] C. S. Peskin, *Linearized Immersed Boundary*, unpublished notes.
[20] C. S. Peskin, *The immersed boundary method*, in Acta Numerica, Acta Numer. 11, Cambridge University Press, Cambridge, UK, 2002, pp. 1–39.
[21] C. S. Peskin and D. M. McQueen, *A three-dimensional computational model for blood flow in the heart.* I. *Immersed elastic fibers in a viscous incompressible fluid*, J. Comput. Phys., 81 (1989), pp. 372–405.
[22] C. S. Peskin and B. F. Printz, *Improved volume conservation in the computation of flows with immersed boundaries*, J. Comput. Phys., 105 (1993), pp. 33–46.
[23] C. Semler and M. P. Païdoussis, *Nonlinear analysis of the parametric resonances of a planar fluid-conveying cantilevered pipe*, J. Fluids Struct., 10 (1996), pp. 787–825.
[24] J. M. Stockie, *Analysis and Computation of Immersed Boundaries, with Application to Pulp Fibres*, Ph.D. thesis, Institute of Applied Mathematics, University of British Columbia, Vancouver, BC, Canada, 1997; also available online from http://www.iam.ubc.ca/theses/stockie/stockie.html.
[25] J. M. Stockie and S. I. Green, *Simulating the motion of pulp fibers using the immersed boundary method*, J. Comput. Phys., 147 (1998), pp. 147–165.
[26] J. M. Stockie and B. T. R. Wetton, *Stability analysis for the immersed fiber problem*, SIAM J. Appl. Math., 55 (1995), pp. 1577–1591.
[27] C. J. Tranter, *Bessel Functions with Some Physical Applications*, Hart Publishing, New York, 1968.
[28] X. Wang, *Instability analysis of some fluid-structure interaction problems*, Comput. & Fluids, 32 (2003), pp. 121–138.
[29] L. Zhu and C. S. Peskin, *Simulation of a flapping flexible filament in a flowing soap film by the immersed boundary method*, J. Comput. Phys., 179 (2002), pp. 452–468.

SIAM J. APPL. MATH.
Vol. 65, No. 2, pp. 521–549

DYNAMICS IN A ROD MODEL OF SOLID FLAME WAVES*

J. H. PARK†, A. BAYLISS†, AND B. J. MATKOWSKY†

Abstract. We consider gasless solid fuel combustion in a cylinder of radius $\tilde{R}$, associated with the SHS (self-propagating high-temperature synthesis) process, which employs combustion waves to synthesize materials. A powder mixture of reactants is cold pressed into a sample, typically a cylinder, and ignited at one end. A high-temperature combustion wave then propagates along the sample, converting the unreacted powder into the desired product. It has recently become popular to model systems governed by partial differential equations as an array of interacting oscillators. Here, we extend this approach by considering an array of interacting rods, each of which supports propagating waves. Thus, we employ an array of interacting one-dimensional (1D) rods connected via heat transfer. The heat transfer terms correspond to a discretization of the transverse Laplacian.

Both the full 3D model and the rod model allow for a uniformly propagating planar combustion wave. The dispersion relation for this solution is determined for both models and shown to be equivalent when certain parameters in the two models are identified. The rod model is able to describe a number of the features of the 3D model, thus allowing numerical simulations with significantly reduced computational resources. In this paper we consider a rod model consisting of an outer ring of three rods equally spaced along the ring, together with an axial rod. Clearly, this limits the 3D modes of wave propagation which can be described, and the results below have to be considered within this basic limitation. The 3/1 model admits analogues of spin and radial modes which are known to exist experimentally and as solutions of the 3D model. We propose that the new modes of solution behavior that we find are also related to modes of the 3D model.

We determine solution behavior as a function of R, the nondimensionalized cylindrical radius. We consider three cases characterized by the Zeldovich number $Z = N(1-\sigma)/2$, where N is a nondimensionalized activation energy and σ is the ratio of the unburned to the burned temperature. For Z sufficiently small, the uniformly propagating planar solution is stable to planar, i.e., rod independent, perturbations. In this case, analysis of the dispersion relation predicts that for sufficiently small and sufficiently large R, only the uniformly propagating solution is stable. Spin modes, seen in experiments as hot spots spinning periodically around the cylinder as the wave propagates, occur for small R, and radial modes, periodically oscillating solutions independent of the cylindrical angle, occur for larger R. We find analogues of these modes for the rod model and describe the transition from spin to radial modes via a family of quasiperiodic (QP) modes, manifested by periodic variations in the temperature and rotation speed of the spot.

For a larger value of Z, the uniformly propagating solution is unstable to planar perturbations. In this case, we find that for sufficiently small values of R, the singly periodic pulsating planar (PP) solution is the only stable solution. For larger values of R, there is bistability between PP solutions and spin solutions. As R increases further, both solutions lose stability, and the only stable solution is a QP mode. This mode is a combination of spin and radial behavior. Unlike the smaller Z case, stable radial solutions are not found. As R increases, the two generator frequencies of the QP solutions converge to the dominant frequency of the PP solution. In the time domain the solution in each rod is approximately the PP solution on intermediate time scales, with a long time envelope corresponding to the small difference between the two generator frequencies. There is a phase difference between adjacent outer rods which is essentially constant over intermediate time scales, but which varies over longer time scales. The intermediate time scale increases as R increases. Thus, we find QP spin-like behavior which on intermediate time scales appears as a spinning manifestation of the PP solution. We refer to this behavior as spinning PP (SPP) behavior.

We also consider a yet larger value of Z. We find a PP solution which is now period doubled. In addition, we find spin and QP spin solutions as before. We also find an interval in R where spinning-type solutions reverse direction in a periodic or QP fashion.

Key words. combustion wave dynamics, self-propagating high-temperature synthesis, solid flames

*Received by the editors February 4, 2004, accepted for publication (in revised form) June 10, 2004; published electronically January 5, 2005. This research was supported by NSF grants DMS02-02485, DMS00-72491, and NSF-IGERT grant DGE99-87577.

http://www.siam.org/journals/siap/65-2/60378.html

†Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL 60208 (jangpark@northwestern.edu, a-bayliss@northwestern.edu, b-matkowsky@northwestern.edu).

DOI. 10.1137/040603784

1. Introduction. We consider modes of solid flame propagation associated with the SHS (self-propagating high-temperature synthesis) process of materials synthesis. In this process reactants are ground into a powder, cold pressed into a solid sample, typically a cylinder, and ignited at one end. Synthesis ensues as a high-temperature self-sustaining combustion wave propagates through the sample, converting reactants to products. When gas plays no significant role in the process, the resulting gasless combustion wave is referred to as a "solid flame" [12]. The SHS process enjoys a number of advantages over conventional technology, in which the sample is placed in a furnace and "baked" until it is "well done." The advantages include (i) simpler equipment; (ii) significantly shorter synthesis times; (iii) greater economy, since the internal energy of the chemical reactions is employed rather than the costly external energy of the furnace; (iv) greater product purity, due to volatile impurities being burned off by the very high temperatures of the propagating combustion wave; and (v) no intrinsic limit on the size of the sample to be synthesized, as exists in conventional technology.

In many instances, synthesis does not proceed in a uniform manner. Rather, nonuniform structures appear in the synthesized samples, corresponding to nonuniformly propagating combustion waves. A variety of dynamical modes of propagation have been found both from analysis and computations (see, e.g., [1, 2, 7, 8, 10, 12]), including planar pulsating (PP) modes, in which waves independent of the radial and angular coordinates, $\tilde{r}$ and ψ, exhibit time periodic oscillations; spinning modes, where one or more hot spots rotate periodically around the cylindrical axis as the wave propagates, so that a helical pattern is visible on the cylindrical surface; and radial modes, where a periodically oscillating wave depends on $\tilde{r}$ but not ψ, i.e., there are radial pulsations as the wave propagates axially. We note that it is difficult to visually determine the internal structure of spin and radial waves, as only the surface of the sample is visible. The internal structure can only be ascertained from a dissection of the synthesized product in controlled experiments. Thus, it is useful to obtain information on these modes from computations.

We consider solid flame waves in a cylindrical geometry. A full simulation of this problem would require three-dimensional (3D) computations. In order to reduce the required computational resources, we introduce a simplified model consisting of an array of 1D rods aligned along the cylindrical axis and connected to each other via heat transfer. The model simulates the 3D sample, but with coarse angular and radial grids. It has recently become popular to model systems governed by partial differential equations using an array of interacting oscillators. Here, we extend this approach by considering an array of interacting rods, each of which supports propagating waves. Our model generalizes the model introduced in [1], where surface spin modes were studied via a discrete number of interacting layers on the surface of the cylinder. In particular, spin modes were identified as pulsating solutions in each layer with a constant phase difference between adjacent layers. In this paper we consider a 3/1 model in which three equally spaced outer rods are located on the surface of the cylinder and an axial rod is located on the axis. All rods are connected to one another via heat transfer. This very simple rod model clearly limits the 3D modes that can be described. For example, it does not allow for a description of multiheaded spin solutions. However, it does allow an examination of the behavior of one-headed (one rotating hot spot) spins for large R, the nondimensional cylindrical radius. In addition, it allows the description of radial waves and of spin-like and radial-like quasiperiodic (QP) waves.

Since the activation energy of the reaction is typically large, the region between the burned and unburned material is generally very thin. Therefore, the reaction term is sometimes replaced by a δ-function of appropriate strength. This is referred to as the *reaction sheet approximation*. The resulting temperature profile is continuous, though its derivative is discontinuous at the reaction site (the site of the δ-function), which separates the burned from the unburned region. Here, we employ the distributed Arrhenius reaction term. We will use the descriptive term *front* to describe the thin region separating the burned and unburned regions.

A key parameter in describing solution behavior is the Zeldovich number $Z = N(1-\sigma)/2$, where N is a nondimensionalized activation energy and $\sigma = \tilde{T}_u/\tilde{T}_b$, where $\tilde{T}_u$ and $\tilde{T}_b$ are the unburned and adiabatic burned temperatures, respectively. Both the full 3D model and the rod model admit a uniformly propagating solution. In the rod model it is characterized by a uniformly propagating combustion wave in each rod, all propagating in phase. For all values of R the uniform solution loses stability to planar, i.e., rod independent, perturbations at a critical value Z_c, leading to a planar pulsating (PP) solution, characterized by a pulsating wave in each rod, all propagating in phase. In the reaction sheet approximation, $Z_c = 2+\sqrt{5}$ (see [11]). For Arrhenius kinetics as considered here, Z_c is slightly different. The dispersion relations for the uniformly propagating solution of the full 3D model in the reaction sheet approximation and of the rod model are shown to be equivalent when appropriate parameter identifications are made.

We consider solution behavior for selected values of Z, as R increases. The parameter R is technologically important, as the objective often is to synthesize large samples. We determine detailed solution behavior as a function of R for three values of Z, one less than Z_c and two greater than Z_c. The largest value of Z corresponds to the case when the PP solution is period doubled.

Preliminary results for $Z < Z_c$ were presented in [13]. It was shown that only uniformly propagating solutions are found for small values of R ($R < R_{sp}$). For $R > R_{sp}$, stability is transferred to a family of spin modes, characterized by periodically pulsating waves of period T in each rod, with a constant $(T/3)$ phase difference between adjacent outer rods. The spin modes lose stability above $R = R_s$. For large R ($R > R_r$) there exist radial modes, characterized by pulsating propagating waves in each rod, with the oscillations in the outer rods being in phase with each other but out of phase with the axial rod. Radial modes lose stability below $R = R_r > R_s$. The transitions at R_s and R_r are supercritical Hopf bifurcations. As R increases above R_s or decreases below R_r, a new frequency enters, and as R increases in $R_s < R < R_r$, a family of QP modes was found which continuously evolved from spin-like to radial-like character. The QP solutions are combinations of radial and spin modes, with the radial component entering with zero amplitude as the spin modes lose stability at $R = R_s$. Similarly, the spin component enters with zero amplitude when the radial mode loses stability at $R = R_r$. Depending on R, two different types of QP modes were found. For R near R_s the QP modes are characterized by spin behavior for the outer rods; i.e., the rods always fire in a fixed order, though there is a modulation in both the firing amplitude as well as the interval between successive firings. For R near R_r where radial behavior predominates, while there is still a modulation in the amplitude of the firings, the outer rods no longer fire in a fixed order. The former behavior can be described as spin-type QP behavior, while the latter are nonspin QP modes.

It is known that as Z increases the uniformly propagating solution becomes unstable and stable PP modes exist. As Z increases further, the PP solution undergoes

a bifurcation from singly periodic to doubly periodic ($2T$) [5]. We have considered two values of $Z > Z_c$, Z_T and Z_{2T}. For $Z = Z_T$ the PP solution is singly periodic, while for $Z = Z_{2T}$ the PP solution is period doubled. For $Z = Z_T$, stable PP solutions are found for $R < R_p$. Spin modes are found for $R_{sp} < R < R_s$, where the same notation is used for the spin mode boundary R_s, for both values of Z, though the numerical values are different. We note that $R_{sp} < R_p$, so that there is a region of bistability between the pulsating planar and spin modes. Spin modes lose stability at $R = R_s$ where stability is transferred to a family of QP solutions. No further transitions are found for the range of values of R considered. In particular, for $Z > Z_c$ we do not find stable radial solutions. Furthermore, all QP solutions found for this value of Z are spin-type QP solutions. However, examination of the frequency content of the QP solutions shows that their spectra are generated by two distinct generator frequencies. By examining the behavior associated with each of these frequencies, we show that the QP modes are combinations of spin and radial behavior as in the $Z < Z_c$ case. Calling these frequencies f_{sp} and f_r, we find that as R increases, the two frequencies coalesce, i.e., $f_{sp} - f_r \to 0$. Furthermore, the limiting frequency is the dominant frequency of the PP solution, f_p. We note that since the equations are autonomous, the term PP solution encompasses a family of solutions differing from each other by a constant phase shift in time. For large values of R, the QP solutions exhibit spin-like behavior, where for each rod the solution is approximately periodic over intermediate time scales which increase as R increases, with a period corresponding to the period of the PP solution and with a long period modulation in amplitude and phase corresponding to the difference frequency $f_d = \mid f_{sp} - f_r \mid$. The amplitude modulation decays to 0 as $R \to \infty$. There is a phase difference between adjacent outer rods which is essentially constant on intermediate time scales, but which varies over longer time scales. The phase modulation does not vanish as $R \to \infty$. The solution remains QP, though the quasiperiodicity is manifested only by a persistent modulation in the phase difference between temperature maxima of adjacent rods over long time scales. These time scales become infinite as $R \to \infty$. Thus, for large R we find an approximate spin mode where the behavior in each rod, over intermediate time scales, is close to that of an appropriately phase shifted PP solution. That is, the solution behaves as a multidimensional form, specifically a spinning form, of the 1D pulsating planar solution. We refer to this behavior as a spinning form of the PP (SPP) solution.

We have not traced the behavior for $Z = Z_{2T}$ in as much detail. Analogous to the case for $Z = Z_T$, we find that for small values of R only PP solutions are found, with the PP solution being $2T$. As R increases, we find periodic and QP spin modes. We note that the periodic spin modes are singly periodic. The dynamics of the spin modes do not appear to be affected by the period doubled nature of the pulsating planar solution, i.e., by the presence of the subharmonic in the PP mode. The QP spin mode is a combination of a spin and radial mode as is the case for smaller values of Z.

As R increases, we find QP spin modes where a subharmonic of one of the generator frequencies develops. These are spin modes which have the character of period doubled solutions; i.e., the front temperature exhibits a high temperature firing followed by a low temperature firing. There is also an interval of apparently chaotic behavior as R is increased further. Beyond a critical value of R we find periodic direction reversing modes which exhibit the character of both T and $2T$ solutions. In these modes the outer rods fire in a clockwise sequence and then a counterclockwise sequence. If the outer rods are numbered 2, 3, and 4, the rods fire in the sequence 324 followed by 423, with this behavior repeated periodically. Thus, the direction of spin

reverses periodically, and two of the rods (in this case rods 3 and 4) fire successively. The successive firings are at different amplitudes, while the intermediate rod (rod 2) always fires at the same amplitude. Thus, rod 2 and the axial rod exhibit singly periodic behavior, while the remaining two outer rods exhibit a period doubled behavior with firings at two different amplitudes within each cycle. Increasing R further leads to QP direction reversing modes.

Dynamical modes of solid fuel combustion were studied in [2] for the case that burning was confined to the surface. We showed that spin modes evolved into QP spins as R increased, as occurs in our rod model. The transition from spin modes to QP spin-type modes was accompanied by an increasing localization of the hot spots as R increased, thus leading to nearly planar behavior over a large fraction of the front on the cylindrical surface.

Three-dimensional computations of spinning modes are presented in [7, 8] for adiabatic combustion with $Z > Z_c$. Generally, the computational requirements for full 3D computations preclude the detailed, high resolution, studies presented here. Results were presented for one-headed spins, exhibiting both a rigid rotation of the hot spot, i.e., a traveling wave in the angular coordinate, and variable spot behavior where the spot exhibits a variability in brightness, size, and speed (possible QP behavior). Furthermore, the amplitude of the pulsations on the axis increased with R. These results are qualitatively similar to the results for the simplified rod model presented here. Furthermore, we identify the QP behavior as due to an interaction between spin and radial modes and demonstrate that axial pulsations increase as the amplitude of the radial component increases. This is a different mechanism for the transition to quasiperiodicity than that observed for surface combustion [2], where radial modes are not present. The mechanism identified for surface combustion was the increasing localization of the spot, leading to nearly uniform behavior over a large portion of the front, together with the fact that, for the parameters considered, the uniform solution is unstable. The spot extends radially into the sample, and even if it is localized on the surface, it will not necessarily be localized in the interior. Thus, the breakdown of periodic spins is due not to localization, but rather to the interaction of spin with radial modes. A preliminary analysis of the nature of the quasiperiodicity in three dimensions is presented in [13].

Since all the results presented in this paper are for the 3/1 rod model, it is still an open question as to which dynamical behaviors that we find are qualitatively similar to those which would be observed in the full 3D problem. While this model does not allow for a description of multiheaded spin modes, it does allow for a description of modes involving one hot spot. As shown in [2] for surface combustion, such modes may persist stably for relatively large values of R. Spin, radial, and QP spin modes which are combinations of spin and radial modes are found for models involving a larger number of rods, suggesting that this behavior is not an artifact of this particular rod model. In addition, SPP behavior is observed for more extensive rod models, thus suggesting that such behavior would also be observed in the continuous system. However, this has yet to be verified.

2. The mathematical model. We consider a model consisting of an array of 1D rods coupled via heat transfer. We denote dimensional quantities by $\tilde{\cdot}$. We also employ this notation for the mass fraction, even though it is nondimensional, in order to distinguish it from the normalized mass fraction employed in the nondimensional model. We consider a configuration with three rods symmetrically located on a cylindrical surface of radius $\tilde{R}$ and one rod located on the cylindrical axis. A similar model was employed in [1] to simulate the case when burning is confined to the surface of

the cylinder.

Let $\tilde{T}_i$ and $\tilde{Y}_i$ be the temperature and mass fraction, respectively, of a deficient component of the reaction in rod i. Suppose that rod 1 is located on the cylindrical axis, while rods 2, 3, and 4 are symmetrically located on $\tilde{r} = \tilde{R}$. For $i = 2, 3, 4$ we have

$$\frac{\partial \tilde{T}_i}{\partial \tilde{t}} = \tilde{\lambda}\frac{\partial^2 \tilde{T}_i}{\partial \tilde{z}^2} + \tilde{q}\tilde{W}(\tilde{T}_i, \tilde{Y}_i) - \tilde{\alpha}_\psi(\tilde{T}_i - \tilde{T}_{i+}) - \tilde{\alpha}_\psi(\tilde{T}_i - \tilde{T}_{i-}) - \tilde{\alpha}_{ri1}(\tilde{T}_i - \tilde{T}_1), \tag{1}$$

$$\frac{\partial \tilde{Y}_i}{\partial \tilde{t}} = -\tilde{W}(\tilde{T}_i, \tilde{Y}_i),$$

where $i+ = 3, 4, 2$ and $i- = 4, 2, 3$. Here, $\tilde{\lambda}$ is the thermal diffusivity and $\tilde{q} = \tilde{Q}/(\tilde{c}\tilde{\rho})$, where $\tilde{Q}$, $\tilde{c}$, and $\tilde{\rho}$ are the heat of reaction, specific heat, and solid density, respectively (all assumed to be constant and independent of i), is the scaled heat of reaction. The coefficients $\tilde{\alpha}_\psi$ and $\tilde{\alpha}_{ri1}$ represent heat transfer between the rods in the angular and radial directions (from rod i to the axial rod), respectively. The reaction rate $\tilde{W}$ is given by

$$\tilde{W}(\tilde{T}_i, \tilde{Y}_i) = \tilde{A}g(\tilde{T}_i)\tilde{Y}_i \exp\left(\frac{-\tilde{E}}{\tilde{R}_g \tilde{T}_i}\right), \tag{2}$$

where $\tilde{A}, \tilde{E}, \tilde{R}_g$ are the frequency factor, activation energy, and universal gas constant, respectively, and

$$g(\tilde{T}_i) = 0,\ \tilde{T}_i < \tilde{T}_{cut}, \qquad g(\tilde{T}_i) = 1,\ \tilde{T}_i > \tilde{T}_{cut},$$

where $\tilde{T}_{cut}$ is chosen to cut off the reaction far ahead of the combustion zone.

The equations for the axial rod $\tilde{T}_1$ are

$$\frac{\partial \tilde{T}_1}{\partial \tilde{t}} = \tilde{\lambda}\frac{\partial^2 \tilde{T}_1}{\partial \tilde{z}^2} + \tilde{q}\tilde{W}(\tilde{T}_1, \tilde{Y}_1) - \sum_{i=2}^{i=4} \tilde{\alpha}_{r1i}(\tilde{T}_1 - \tilde{T}_i), \quad \frac{\partial \tilde{Y}_1}{\partial \tilde{t}} = -W(\tilde{T}_1, \tilde{Y}_1), \tag{3}$$

where $\tilde{\alpha}_{r1i}$ represents heat transfer from the axial rod to the outer rod i.

For all rods the solutions satisfy the boundary conditions

$$\lim_{\tilde{z}\to-\infty} \tilde{T}_i = \tilde{T}_u, \qquad \lim_{\tilde{z}\to-\infty} \tilde{Y}_i = \tilde{Y}_u, \qquad \lim_{\tilde{z}\to\infty} \frac{\partial \tilde{T}_i}{\partial \tilde{z}} = 0, \tag{4}$$

where $\tilde{T}_u, \tilde{Y}_u$ are the unburned temperature and mass fraction, respectively. We note that $\tilde{T}_i \to \tilde{T}_b$ as $\tilde{z} \to \infty$, where $\tilde{T}_b$ is the adiabatic burned temperature; however, we employ the Neumann condition in our computations. Finally, we note that $\tilde{T}_b$ is derivable from thermodynamical considerations as $\tilde{T}_b = \tilde{T}_u + \tilde{q}\tilde{Y}_u$.

The heat transfer coefficients $\tilde{\alpha}_\psi, \tilde{\alpha}_{ri1}, \tilde{\alpha}_{r1i}$ correspond to a coarse grained approximation to the transverse Laplacian, i.e., the terms $\tilde{T}_{\psi\psi}/\tilde{r}^2$ and $(\tilde{r}\tilde{T}_{\tilde{r}})_{\tilde{r}}/\tilde{r}$, with periodicity in ψ assumed, thus establishing a relationship between the rod model and the fully 3D model. Approximating these expressions by finite differences leads to unique heat transfer terms. For example, interpreting rod i as a grid point and approximating the angular diffusion term at rod i gives

$$\frac{\tilde{T}_{\psi\psi}}{\tilde{r}^2} \simeq -\frac{\tilde{T}_i - \tilde{T}_{i+}}{(\Delta\psi\tilde{R})^2} - \frac{\tilde{T}_i - \tilde{T}_{i-}}{(\Delta\psi\tilde{R})^2}, \qquad \text{so that } \tilde{\alpha}_\psi = \frac{\tilde{\lambda}}{(\Delta\psi\tilde{R})^2},$$

where $\Delta\psi = 2\pi/3$ since we take three outer rods. For radial heat transfer, $\tilde{\alpha}_{ri1}$ can be determined by employing a finite difference approximation for radial diffusion together with the no flux boundary condition $\tilde{T}_r = 0$ at $\tilde{r} = \tilde{R}$, giving

$$\frac{(\tilde{r}\tilde{T}_{\tilde{r}})_{\tilde{r}}}{\tilde{r}} \simeq -2\frac{\tilde{T}_i - \tilde{T}_1}{(\Delta\tilde{r})^2}, \qquad \text{so that } \tilde{\alpha}_{ri1} = \frac{2\tilde{\lambda}}{(\Delta\tilde{r})^2},$$

where $\Delta\tilde{r} = \tilde{R}$. For heat transfer from the axis, we approximate the Laplacian by a suitably averaged five point difference in Cartesian coordinates with $\Delta\tilde{x} = \Delta\tilde{y} = \Delta\tilde{r}$ to get

$$\nabla^2\tilde{T} \simeq \frac{4}{3\Delta\tilde{r}^2}\sum_{i=2}^{i=4}(\tilde{T}_i - \tilde{T}_1), \qquad \text{so that } \tilde{\alpha}_{r1i} = \frac{4\tilde{\lambda}}{3(\Delta\tilde{r})^2},$$

where $\Delta\tilde{r} = \tilde{R}$.

Although this model is very coarse-grained, it allows a qualitative description of the fully 3D problem at a small fraction of the computational cost. We nondimensionalize as in [11] by introducing

$$Y_i = \frac{\tilde{Y}_i}{\tilde{Y}_u}, \qquad \Theta_i = \frac{\tilde{T}_i - \tilde{T}_u}{\tilde{T}_b - \tilde{T}_u}, \qquad t = \frac{\tilde{t}\tilde{U}^2}{\tilde{\lambda}}, \qquad z = \frac{\tilde{z}\tilde{U}}{\tilde{\lambda}},$$

$$\sigma = \frac{\tilde{T}_u}{\tilde{T}_b}, \qquad N = \frac{\tilde{E}}{\tilde{R}_g\tilde{T}_b},$$

where

$$\tilde{U}^2 = \frac{\tilde{\lambda}\tilde{A}}{2Z}\exp(-N),$$

$Z = N(1-\sigma)/2$ is the Zeldovich number, and $\tilde{U}$ is the velocity of the uniformly propagating front in the reaction sheet approximation [9]. Note that lengths are scaled by the size of the preheat zone. Finally, letting $\tilde{\alpha}$ (α) generically denote any of the dimensional (nondimensional) heat transfer coefficients, we have

$$\alpha = \tilde{\alpha}\frac{\tilde{\lambda}}{\tilde{U}^2}.$$

We will describe combustion waves propagating in the axial ($-z$) direction. We introduce the moving coordinate system [2, 11]

$$x = z - \phi(t),$$

where $\phi(t)$ is defined by $Y_1(\phi(t), t) = 0.5$. Here, the choice of the particular rod at which the mass fraction is fixed is arbitrary. Thus, ϕ_t is the approximate velocity of the wave, so that the transformation to the moving coordinate system enables us to localize the front to a neighborhood of $x = 0$.

In terms of the nondimensionalized quantities we have, for $i = 2, 3, 4$,

$$\frac{\partial\Theta_i}{\partial t} = \frac{\partial^2\Theta_i}{\partial x^2} + W(\Theta_i, Y_i) - \alpha_\psi(\Theta_i - \Theta_{i+}) - \alpha_\psi(\Theta_i - \Theta_{i-}) - \alpha_{ri1}(\Theta_i - \Theta_1), \tag{5}$$

$$\frac{\partial Y_i}{\partial t} = -W(\Theta_i, Y_i),$$

where

$$W(\Theta_i, Y_i) = 2Zg(\Theta_i)Y_i \exp\left(\frac{N(1-\sigma)(\Theta_i - 1)}{\sigma + (1-\sigma)\Theta_i}\right).$$

For the axial rod we have

$$\frac{\partial \Theta_1}{\partial t} = \frac{\partial^2 \Theta_1}{\partial x^2} + W(\Theta_1, Y_1) - \sum_{i=2}^{i=4} \alpha_{r1i}(\Theta_1 - \Theta_i), \quad \frac{\partial Y_1}{\partial t} = -W(\Theta_1, Y_1). \tag{6}$$

The boundary conditions are

$$\begin{gathered} Y_i \to 1, \quad \Theta_i \to 0 \ \text{ as } \ x \to -\infty, \\ \frac{\partial \Theta_i}{\partial x} \to 1 \ \text{ as } \ x \to \infty. \end{gathered} \tag{7}$$

These boundary conditions are specified at finite points far from the front which is located in the vicinity of $x = 0$. The computations presented here were obtained with the boundary conditions imposed at $x = \pm 12$. There is virtually no effect of further increasing the size of the computational domain.

Our numerical method employs an adaptive Chebyshev pseudospectral method, described in detail in [2, 3]. We solve the initial value problem, marching forward in time until steady state is achieved, so that we compute only stable solutions.

3. The uniformly propagating planar solution and its stability. As shown in [13], the system (5)–(7) admits a planar uniformly propagating solution. In the reaction sheet approximation this solution (see [11, 10]) is

$$\Theta_{unif} = \begin{cases} \exp(x), & x \le 0, \\ 1, & x \ge 0, \end{cases} \qquad Y = \begin{cases} 1, & x < 0, \\ 0, & x > 0, \end{cases} \qquad \phi(t) = -t + C,$$

where $\Theta_1 = \Theta_2 = \Theta_3 = \Theta_4 = \Theta_{unif}$ and C is an arbitrary constant. Note that for this solution the heat transfer between the rods vanishes. A linear stability analysis shows that three different types of instabilities can occur. The dispersion relation requires that the product of three terms vanishes, so that instability sets in when any of the factors vanishes. In terms of Z the three stability boundaries are given by

$$Z = 2 + \sqrt{5}, \tag{8}$$

$$Z = \frac{4 + 12(3\beta + 2)\alpha_R + \sqrt{\left[4 + 12(3\beta + 2)\alpha_R\right]^2 + 4\left[1 + 4(3\beta + 2)\alpha_R\right]^3}}{2[1 + 4(3\beta + 2)\alpha_R]}, \tag{9}$$

$$Z = \frac{4 + 12(6)\alpha_R + \sqrt{\left[4 + 12(6)\alpha_R\right]^2 + 4\left[1 + 4(6)\alpha_R\right]^3}}{2[1 + 4(6)\alpha_R]}, \tag{10}$$

where $\beta = 1/(\Delta\psi)^2 = 9/(4\pi^2)$, $\alpha_R = 1/R^2$, and $R = \tilde{R}U/\tilde{\lambda}$ is the nondimensional cylindrical radius. The eigenvectors indicate that (8) corresponds to a PP solution where all rods pulsate in phase and describes the well-known PP stability boundary, while (9) corresponds to a one-headed spin mode (the outer rods exhibit identical pulsations with a phase difference of $T/3$ where T is the period), and (10) represents

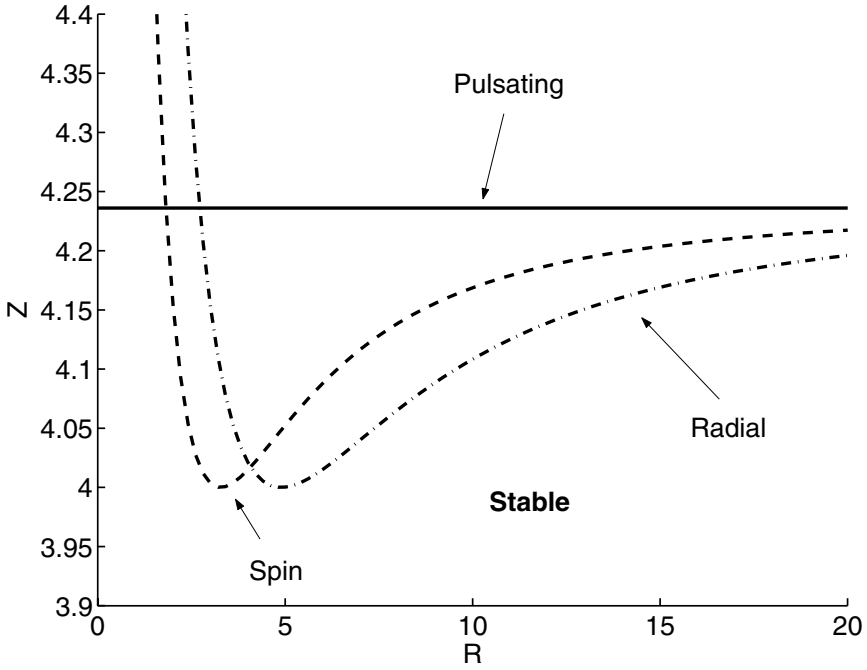


FIG. 1. *Stability boundaries for the uniform solution.*

a radial mode where the outer rods pulsate in phase, but are out of phase with the axial rod, the analogue of 3D radial modes where the solution depends on r but not on ψ. The three stability boundaries are shown in Figure 1.

The dispersion relation for the nondimensionalized continuous 3D problem is

$$Z_{3D} = \frac{4 + 12k^2 + \sqrt{\left[4 + 12k^2\right]^2 + 4\left[1 + 4k^2\right]^3}}{2[1 + 4k^2]}, \quad k = \frac{\xi_m^{(n)}}{R}. \tag{11}$$

Here, $\xi_m^{(n)}$ is the mth root of $J_n'(\xi) = 0$, where J_n is the Bessel function of order n, with n the angular wavenumber [10].

Since in our model we consider rods located at only three angular and two radial locations, only wave numbers $n = 0, 1$ and $m = 1, 2$ are relevant for comparison between the full model and our rod model. Since $\xi_1^{(0)} = 0$ $(m = 1, n = 0)$, the dispersion relations for the pulsating solutions are identical for the two models. The spin $(m = 1, n = 1)$ and radial $(m = 2, n = 0)$ modes are equivalent to the rod model if we identify k^2 with $(3\beta + 2)\alpha_R$ and $6\alpha_R$, respectively. This equivalence serves as a partial justification for the assumption that the behavior of the rod model is qualitatively similar to that of the full 3D model.

4. Results. We first consider results for a value of Z below the pulsating stability boundary Z_c, e.g., $Z \simeq 4.1$, for which the linear stability analysis suggests the sequence of transitions uniform → spin → transition region → radial → uniform as R increases. However, Figure 1 is based on a δ-function reaction term, while the results presented here are for the Arrhenius reaction term, leading to a slightly different dispersion

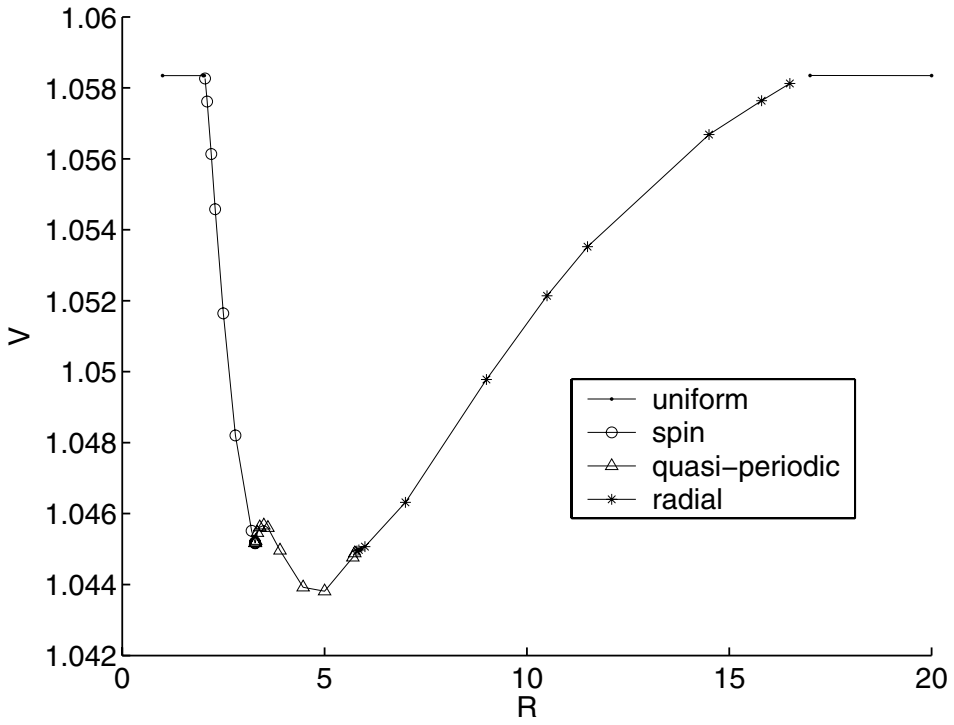


FIG. 2. *Mean speed V for different solution branches for $Z = 3.916$.*

relation. Our computations are for $Z = 3.916$, $N = 17.7$, $\sigma = .5575$, conditions which exhibit behavior analogous to $Z = 4.1$ in Figure 1.

In order to visualize the solution, we compute the temperature as a function of time at the flame front $\Theta_{Fi}(t)$, in each of the rods. Since we do not employ the reaction sheet model, there is not, strictly speaking, a front separating the burned and unburned regions. Nevertheless, we can determine the spatial location at which the reaction term is maximal and use it as an approximate flame front location. Computing this quantity as a function of time and then doing a least squares fit enables us to compute the mean flame speed V. The temperature on the front, $\Theta_{Fi}(t)$ can then be obtained by evaluating the Chebyshev polynomial approximation of Θ at this location [1]. Typically, the time history of $\Theta_{Fi}(t)$ will be oscillatory (though not necessarily periodic). Peaks in $\Theta_{Fi}(t)$ can be interpreted as a firing of the rod. As an analogue of the continuous 3D problem, a peak in $\Theta_{Fi}(t)$ can also be interpreted as a hot spot passing over the angle corresponding to the rod's location. Since this paper deals with the rod model, we will primarily refer to the firing of each rod in describing the dynamics of the solutions that we find.

Preliminary results for $Z = 3.916$ were presented in [13]. These results are summarized in Figure 2, which shows the mean front speed V as a function of R for the different solution branches that we have found. The spin modes are connected to the radial modes via a family of solutions exhibiting QP dynamics where the behavior continuously varies from spin-like (for smaller R) to near-radial (for larger R). There is no indication of bistability.

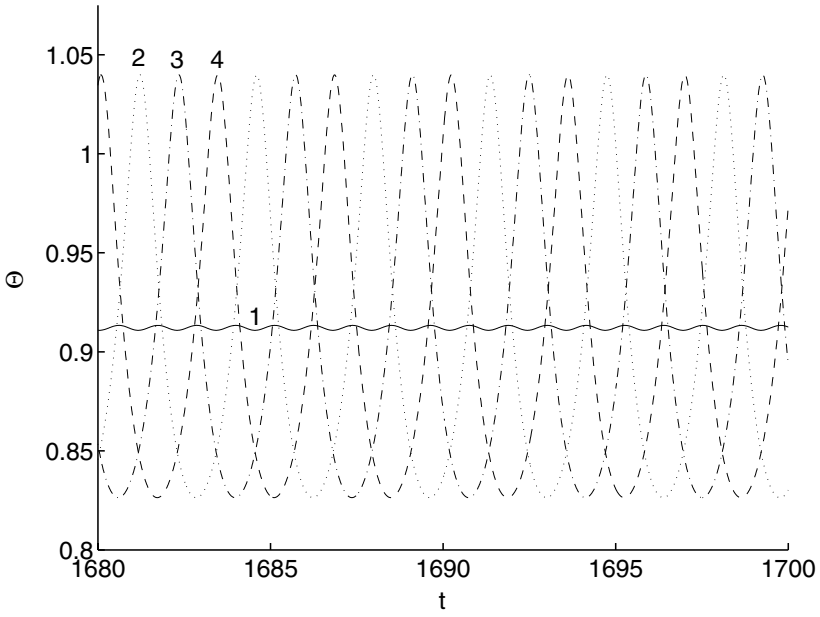


FIG. 3. $\Theta_{Fi}(t)$ $(i = 1, 2, 3, 4)$ *for spin solution for* $Z = 3.916$ *and* $R = 3.2$.

For the rod model a spin mode is represented by pulsating waves that are identical in each of the three outer rods except for the constant $T/3$ phase shift, with $\Theta_{Fi}(t)$ periodic of period T for $i = 2, 3, 4$. This is exactly what would be observed by examining the temperature of a spin mode at three angular locations separated by $2\pi/3$. When $\Theta_{Fi}(t)$ attains a maximum in time, one can think of a rotating hot spot passing over the spatial location corresponding to the rod, i.e., for the three outer rods the angles $k2\pi/3$ $(k = 0, 1, 2)$. Alternatively, in terms of the rod model, one can think of such a temperature maximum as a "firing" of the rod. We note that while the linear stability analysis predicts uniform temperature on the axis, there is a low level periodic temperature pulsation due to nonlinear effects neglected in the linear analysis. This axial pulsation will grow as R increases along the spin branch. Near the onset of spin modes the axial rod fires at a period $1/3$ that of the outer rods; i.e., it fires with the firing of each outer rod though with a slight phase shift, behavior clearly dependent on the choice of three outer rods and one axial rod in the model.

For the spin modes, the firing of the outer rods is synchronized. The rods fire in a fixed order with a constant time interval between firings. Note that the reverse firing order can also occur (corresponding to clockwise and counterclockwise rotating spots), with the order depending on initial conditions. The axial rod also fires periodically, but at a significantly lower level than the surface rods. For larger values of R, radial behavior is found. For a radial mode all outer rods fire simultaneously; i.e., there is no angular behavior to the firing, while the axial rod fires out of phase with the outer rods. Spin and radial modes were found in [13] and are illustrated in Figures 3 and 4, respectively, where the numbers 1–4 refer to the flame front temperatures in rods 1–4, respectively.

Near $R = 3.285$, the spin modes lose stability, and a transition to QP behavior is

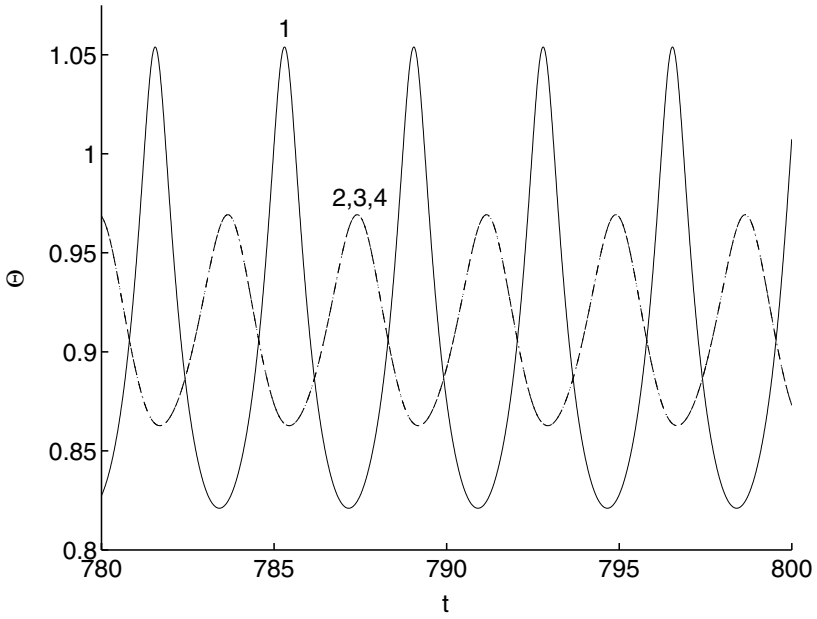


FIG. 4. $\Theta_{Fi}(t)$ $(i = 1, 2, 3, 4)$ *for radial solution for* $Z = 3.916$ *and* $R = 5.8$.

observed. Near the transition point the solution exhibits approximate spin behavior, in that the firing order is preserved. However, the firings are no longer synchronized; the phase shift between firings becomes periodic, corresponding to continuous spin modes in which adjacent hot spots periodically approach and then withdraw from each other, as seen in [2].

The QP modes are best described by analyzing the frequency spectrum of $\Theta_{Fi}(t)$. The spin and radial solutions are periodic in time. Thus, their frequency spectrum is composed of a single dominant frequency and its harmonics. The QP modes have two dominant frequencies which generate all other frequencies in the spectrum via linear combinations. Analysis of the frequency spectrum for the QP solutions indicates that the two generator frequencies correspond to spin and radial behavior, respectively. Thus, the QP modes are combinations of spin and radial modes. At the transition from spin to QP behavior a new frequency, corresponding to radial behavior, enters with zero amplitude. As R increases, the amplitude of the radial spectral component increases while the amplitude of the spin component decreases, so that at the transition from QP to radial behavior, the amplitude of the spin component vanishes.

This behavior was described in [13] and can be seen in Figure 5, where we exhibit the dominant frequencies for spin (S), QP, and radial (R) modes as a function of R. For the spin and radial modes there is only one dominant frequency. For the QP modes there are two dominant frequencies. The figure clearly shows that in the QP regime the two generator frequencies emanate continuously from the spin and radial frequencies, respectively. For $R \leq R_a$ (indicated by point a in the figure) the frequency corresponding to the largest spectral amplitude is the spin frequency, while for $R \geq R_b$ (indicated by point b in the figure), the frequency corresponding to the largest spectral amplitude is the radial frequency. The most dominant frequency

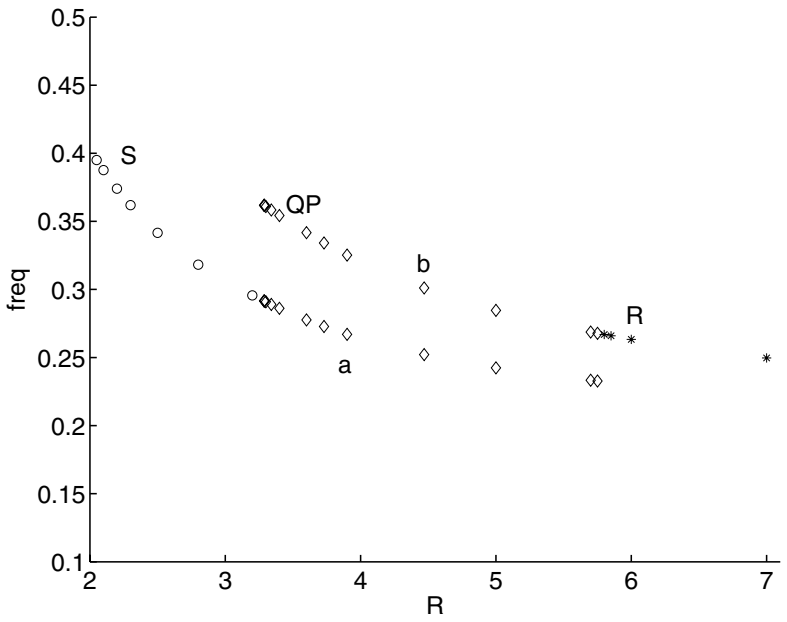


FIG. 5. *Principal frequency components for* $\Theta_{F2}(t)$ *for* $Z = 3.916$ *and* $R = 3.3$.

shifts from the spin frequency to the radial frequency as R increases. Thus, at least in terms of spectral content, for small R the QP modes take on a spin character, while for larger values of R the QP modes take on a radial character.

The time domain behavior, described in [13], is illustrated in the time domain in Figure 6, where we consider a QP mode near the spin-QP transition. For this figure, $\Theta_{Fi}(t)$ was Fourier transformed, the amplitudes of all frequencies except the secondary generator frequency and its harmonics were set to zero, and the resulting spectral representation was transformed back to the time domain. The resulting time domain behavior shown in the figure clearly exhibits radial behavior, including the simultaneous firings of the outer rods and the constant phase shift (near π) between the outer rods and the axial rod. A similar analysis near the transition from QP to radial behavior (Figure 7) indicates that the new frequency corresponds to spin behavior. We note that the solution shown in Figure 7 corresponds to a nonspin QP mode, where the firing order of the outer rods is not fixed. Such modes are connected to the transition to radial behavior. These modes are not found for larger values of Z since we cannot find stable radial modes for values of Z above the pulsating stability boundary.

Behavior near the spin-QP transition was discussed in [13]. Flame front temperatures for a QP mode near the spin-QP transition point are shown in Figure 8. The behavior is close to that of the spin modes. The rods fire in the same firing order; however, the time interval between successive firings is no longer constant, and the temperature maxima exhibit a nonconstant envelope. A QP mode near the QP-radial transition is shown in Figure 9. The outer rods fire nearly simultaneously; however, there is an amplitude modulation, and the time delay between successive firings is nonzero and nonconstant. Up to $R \simeq 3.7$, the outer rods fire in a fixed firing order (that of the spin modes). As R increases, the outer rods fire closer together in time,

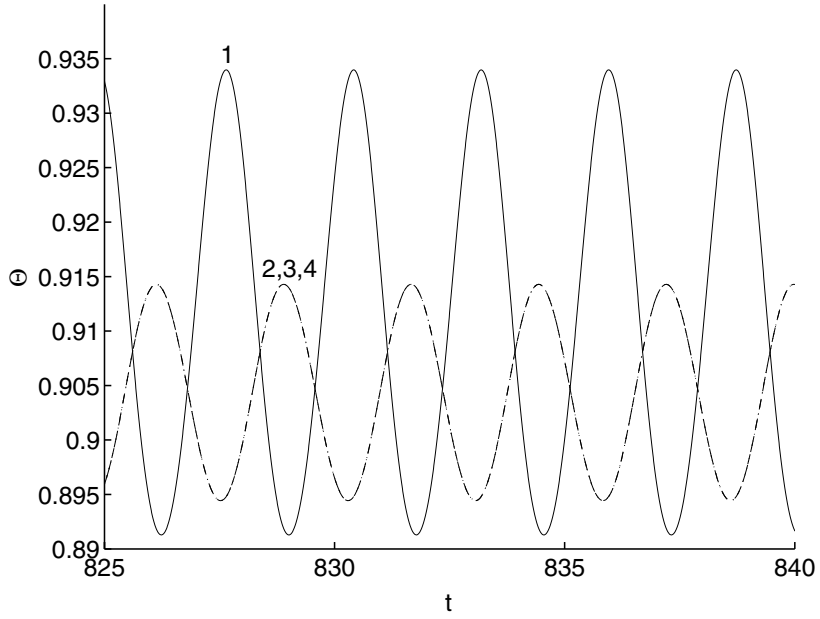


FIG. 6. *Radial component of* $\Theta_{Fi}(t)$ $(i = 1, 2, 3, 4)$ *for QP mode for* $Z = 3.916$ *and* $R = 3.3$.

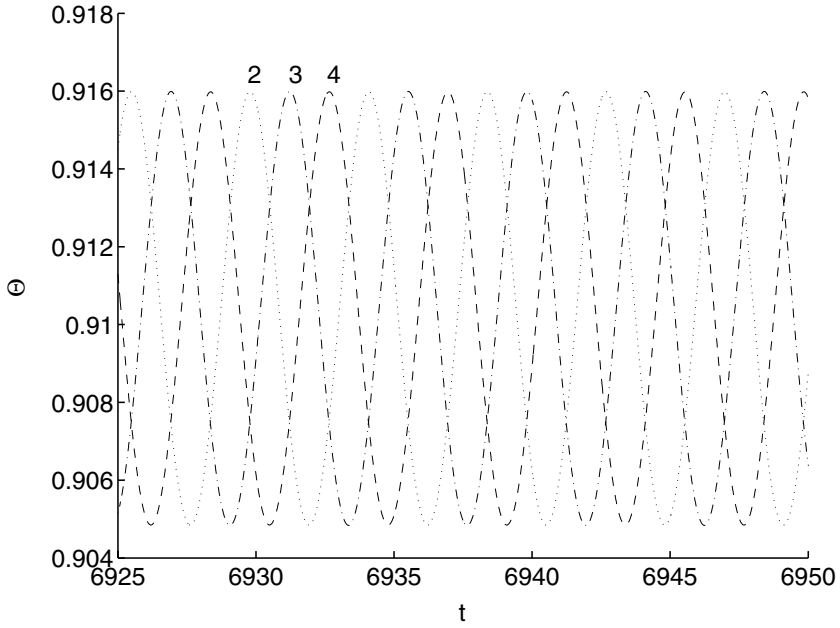


FIG. 7. *Spinning component of* $\Theta_{Fi}(t)$ $(i = 2, 3, 4)$ *for QP mode for* $Z = 3.916$ *and* $R = 5.75$.

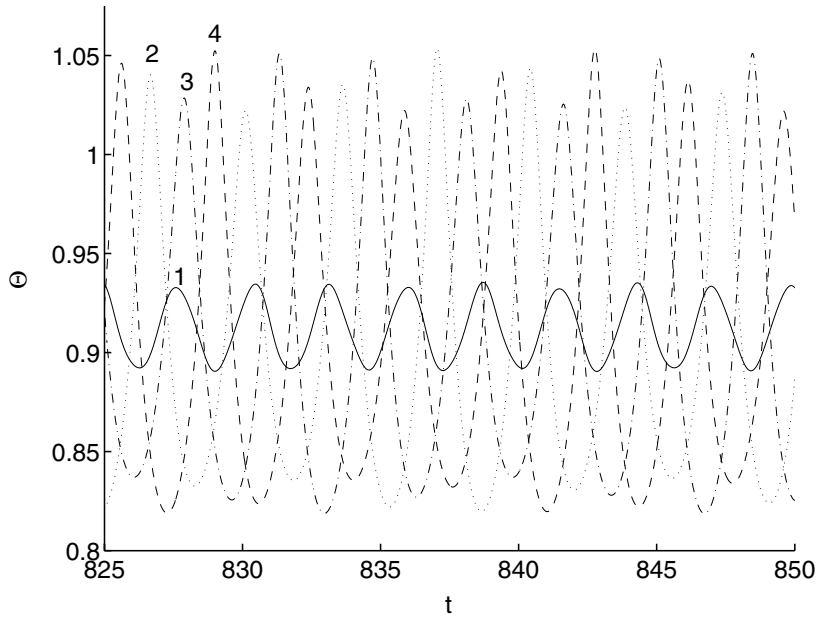


FIG. 8. $\Theta_{Fi}(t)$ $(i = 1, 2, 3, 4)$ *for QP solution for* $Z = 3.916$ *and* $R = 3.3$.

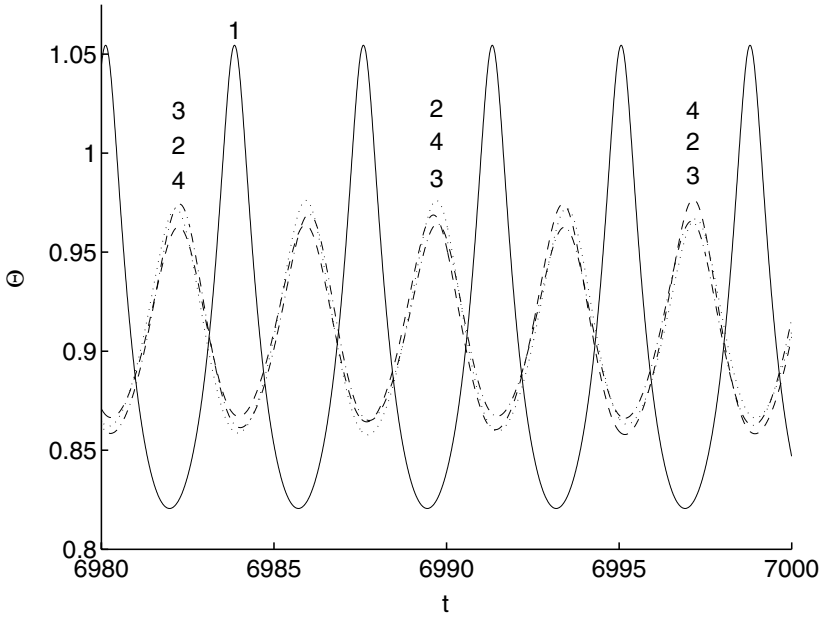


FIG. 9. $\Theta_{Fi}(t)$ $(i = 1, 2, 3, 4)$ *for QP solution for* $Z = 3.916$ *and* $R = 5.75$.

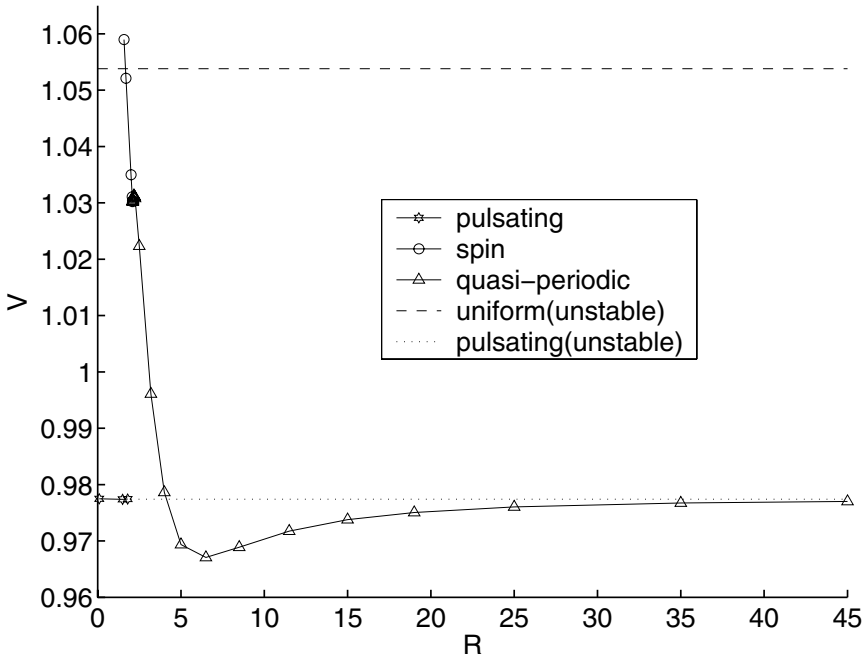


FIG. 10. *Mean speed for different solution branches for* $Z = 4.3$.

with the firing order no longer fixed, as the QP solutions take on more of a radial mode (simultaneous firing) character. In addition, the amplitude of the axial rod grows. Such a solution is shown in Figure 9. Near the transition to radial behavior, the firing of the outer rods occurs nearly simultaneously. Firings occur in groups of three, separated by a time interval close to the period of the radial solution.

Thus, for Z below the pulsating stability boundary there are three branches of nonuniform solutions, a spin branch for small values of R, a radial branch for larger values of R, and a connecting QP branch, where the behavior of the solution continuously changes from spin-like to radial-like as R increases. There is no indication of bistability.

We next consider behavior for Z above the pulsating stability boundary so that the uniformly propagating solution is unstable for all values of R. Specifically, we consider $Z = Z_T = 4.3$. A summary of our results is shown in Figure 10, where we plot V for the different solution branches as a function of R. While the PP solution, involving no heat transfer between the rods, exists for all values of R, we find that it is stable only for small values of R ($R < R_p \simeq 2$). Stable spin modes develop at $R = R_{sp} < R_p$, where $R_{sp} \simeq 1.58$. Since $R_{sp} < R_p$, there is a region of bistability between the spin and PP solutions. Examples of these modes in the bistable region are shown in Figures 11 and 12, where we plot $\Theta_{Fi}(t)$ for a spin (Figure 11) and PP (Figure 12) solution for $R = 1.7$. For $R > R_p$ only spin and QP spin-type modes are found.

Spin modes lose stability at $R = R_{QP} \simeq 2.0994$, and QP spin-type modes develop. This is accompanied by an increase in the amplitude of the axial pulsation. An example of such a QP spin mode is shown in Figure 13, where we plot $\Theta_{Fi}(t)$ for all rods for $R = 2.102$. The frequency spectrum of $\Theta_{F2}(t)$, shown in Figure 14, exhibits

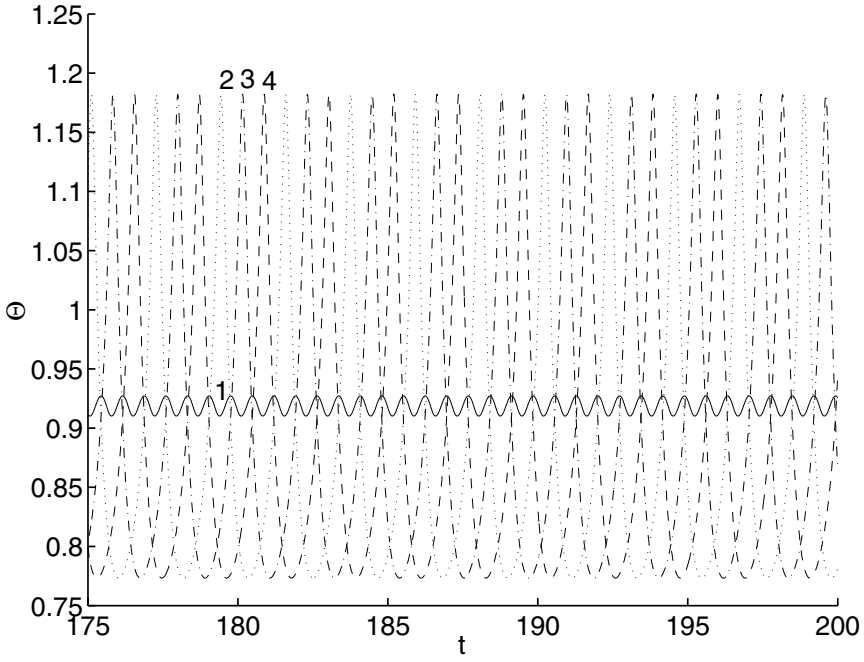


FIG. 11. $\Theta_{Fi}(t)$ $(i = 1, 2, 3, 4)$ *for spin solution for* $Z = 4.3$ *and* $R = 1.7$.

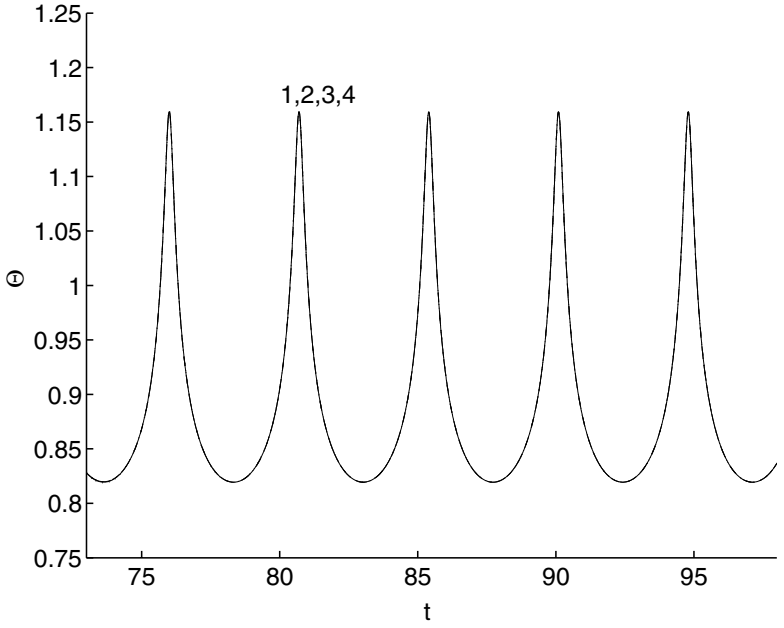


FIG. 12. $\Theta_{Fi}(t)$ $(i = 1, 2, 3, 4)$ *for PP solution for* $Z = 4.3$ *and* $R = 1.7$.

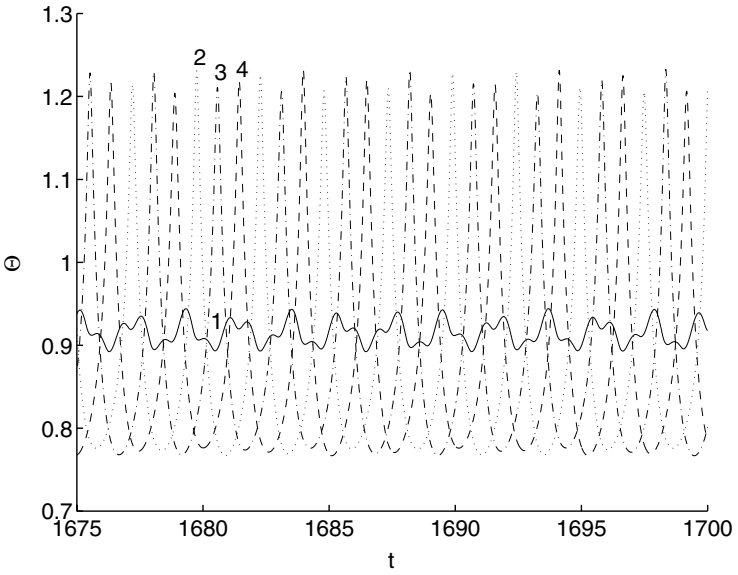


FIG. 13. $\Theta_{Fi}(t)$ $(i = 1, 2, 3, 4)$ *for QP solution for* $Z = 4.3$ *and* $R = 2.102$.

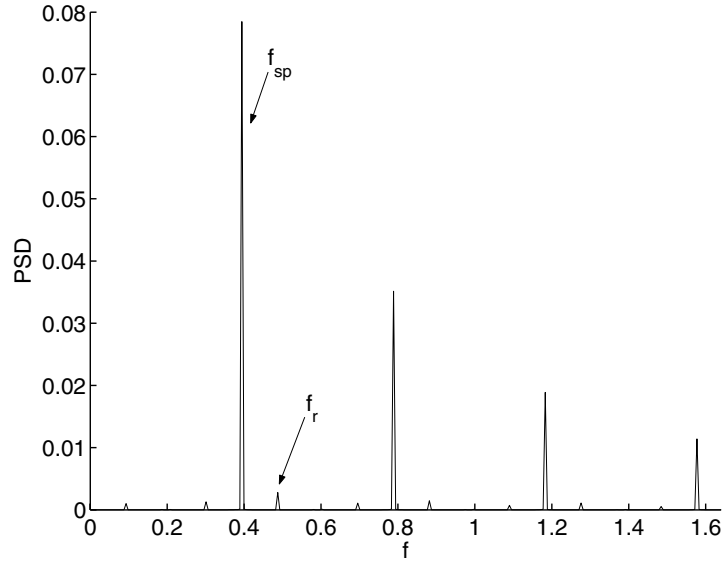


FIG. 14. *Frequency spectrum of* $\Theta_{F2}(t)$ *for QP solution for* $Z = 4.3$ *and* $R = 2.102$.

two clear generator frequencies together with linear combinations. Reconstructing time domain behavior corresponding to each of the two generator frequencies, denoted by f_{sp} and f_r in the figure, as described above, shows that they correspond to spin and radial behavior, respectively. This is shown in Figures 15 and 16.

We next consider the behavior of the QP modes as R increases. Unlike the previous case, we do not find stable radial solutions. For all values of $R > R_{QP}$ that we have considered the only stable solution that we find is the QP spin-type mode,

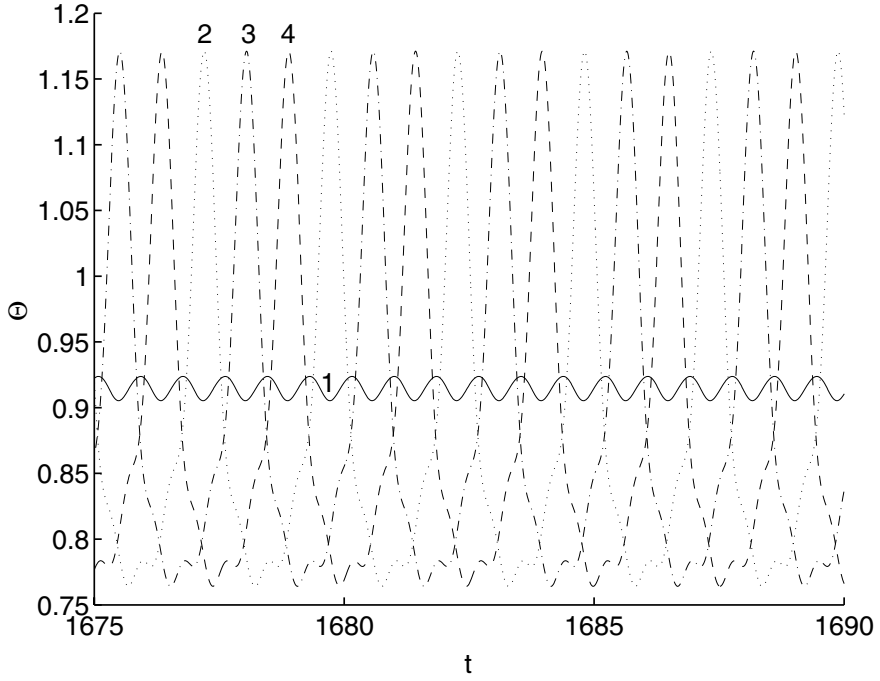


FIG. 15. *Spin component of QP solution for* $Z = 4.3$ *and* $R = 2.102$.

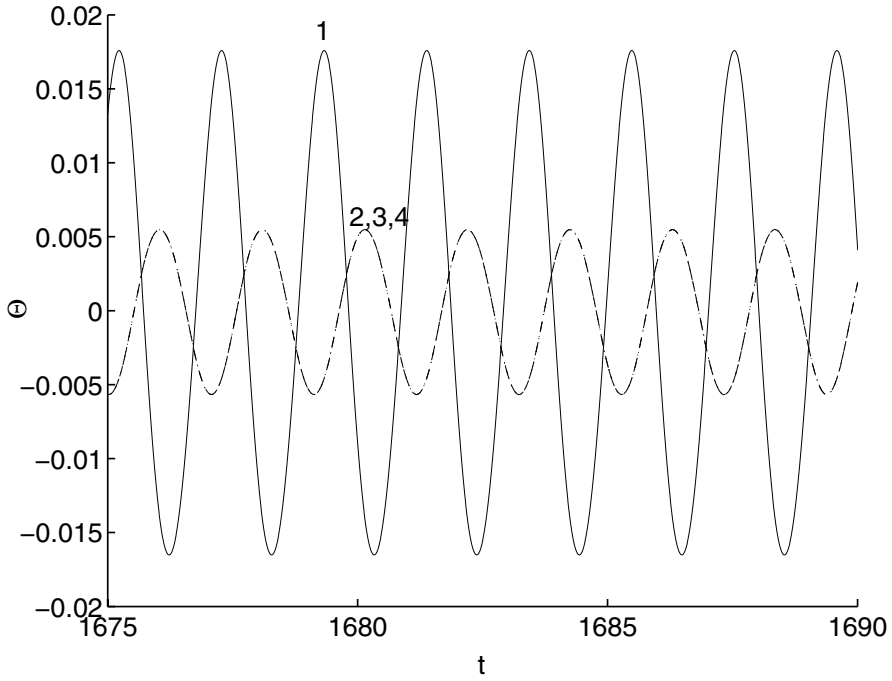


FIG. 16. *Radial component of QP mode for* $Z = 4.3$ *and* $R = 2.102$.

though like all QP spin modes it has a radial component. Furthermore, there is no breakdown in the firing order as occurred previously when there was a transition to radial behavior. For all values of R that we consider, the outer rods fire in a fixed order, characteristic of a spin mode.

As R increases, the QP modes are characterized by an asymptotic coalescence of the dominant spin and radial frequencies to the dominant frequency of the PP solution, f_p. In addition, the front temperatures in the time domain approach that of the PP solution, appropriately shifted in time. The solution varies over three distinct time scales. Over short time scales, of the order of the period T of the PP solution, the QP solution is essentially identical to the PP solution (appropriately phase shifted). Over intermediate time scales, which increase with R, the QP solutions are characterized as SPP solutions; i.e., the time history at each rod is approximately that of the PP solution appropriately phase shifted. Furthermore, adjacent outer rods are phase shifted with respect to each other, by a shift which is nearly constant in time, though the shift is different for different pairs of rods. Thus, over intermediate time scales the solution behaves as a spinning form of the PP solution, with a nonconstant rotation rate. We refer to this behavior as SPP behavior.

In addition, there is a longer time scale determined by the inverse of the difference frequency $f_d = \mid f_{sp} - f_r \mid$. Since $f_d \to 0$ as $R \to \infty$, this time scale grows as R increases. On this time scale there is a phase and an amplitude modulation. The amplitude modulation decreases as R increases and appears to vanish as $R \to \infty$. However, there is a phase variation between any two outer rods which appears to persist as R increases. Thus, for large R the front temperature appears to be QP, but without a noticeable amplitude modulation.

An indication of the SPP behavior can be seen from Figure 10, where V asymptotically approaches the mean speed of the PP solution as R increases. Note that our computations indicate that the PP solution is unstable for large values of R. The only stable solutions that we find are the SPP modes.

The frequency coalescence is shown in Figure 17 where we plot the two generator frequencies for the QP spin mode as a function of R. As R increases, the two generator frequencies converge to the dominant frequency of the PP solution, f_p. We note that even for large R, reconstruction of the time history employing only f_{sp} (f_r) and harmonics shows that these frequencies still correspond to spin (radial) behavior. Thus, for large values of R, spin and radial behavior occur with frequencies which converge to the frequency of the PP solution.

Figure 17 shows only the frequencies corresponding to the largest amplitudes of the spectrum. However, there is convergence in the time domain as well. The front temperature varies over three different time scales. Over time scales comparable to the period T of the PP solution (short time scale), Θ_{Fi} are almost identical to the PP solution (appropriately phase shifted). This behavior is maintained over intermediate time scales, of the order of $10T$, for the values of R considered here. Furthermore, there is a nearly constant phase difference between any two outer rods (though not necessarily $T/3$ as would be expected for a purely spinning mode). Thus, over the intermediate time scale, the QP spin mode behaves as an SPP solution, thus establishing a connection between spin modes and the PP solution which is stable only for small R.

Finally, there is a long time scale in which there is a modulation in the phase difference between two successive firings as well as in the amplitude of the pulsation. This time scale is proportional to f_d^{-1} and thus approaches ∞ as $R \to \infty$. The am-

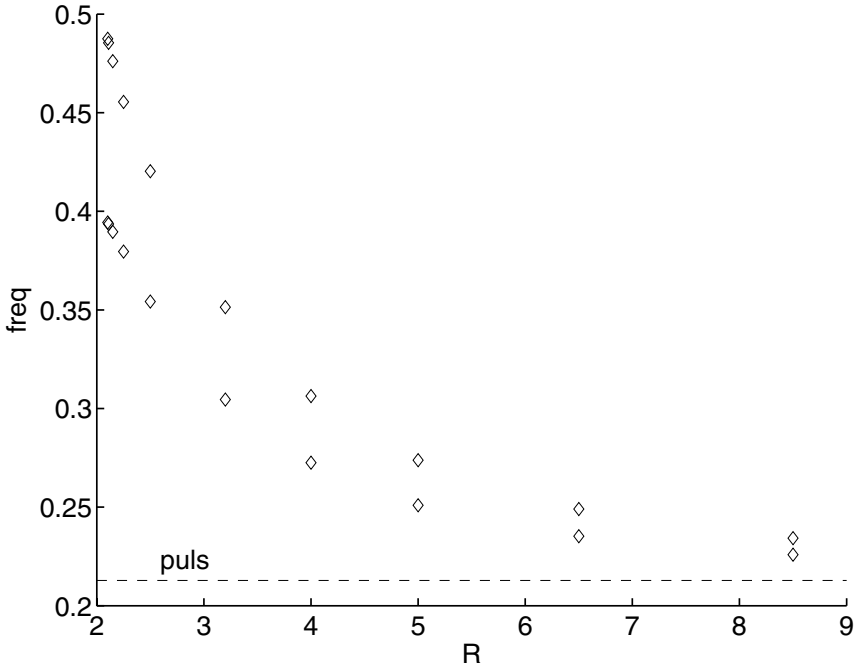


FIG. 17. *Generator frequencies for* $\Theta_{F2}(t)$.

plitude modulation appears to decay to 0 as $R \to \infty$. However, the phase modulation persists and does not decay to 0, even for large values of R.

We first illustrate behavior for intermediate values of R, where there are indications of SPP behavior, although it is not pronounced. In Figure 18 we plot $\Theta_{Fi}(t)$ and the PP solution (appropriately phase shifted) for $R = 19$. The data is plotted over approximately 100 time units. The amplitude and phase modulation can be seen from the figures, although for each rod the QP temperature is close to that of the PP solution (appropriately phase shifted).

In Figure 19 we plot the same quantities for $R = 35$. In this case there are essentially no visible differences between the QP and the PP solution over this time scale, and the QP solution can be considered as an SPP form of the solution. The intermediate time scale on which this occurs increases with R. The phase modulation is illustrated in Figures 20 and 21, where we plot $\Theta_{F2}(t)$ and $\Theta_{F3}(t)$ over roughly three PP periods. Figures 20 and 21 are separated by roughly 1000 units of time, a time interval where the phase modulation is evident. We note that in Figure 20 the time difference between the firings of rods 2 and 3 is approximately one unit, while in Figure 21 the time difference is approximately three units. There is virtually no change in amplitude between the pulsations in the two figures. Furthermore, all time histories in these two figures would be indistinguishable from appropriately phase shifted PP solutions. Thus, the quasiperiodicity of the QP mode for large R is manifested by a long time phase modulation between adjacent rods of a time signal that exhibits essentially no amplitude modulation and is essentially identical to that of the PP solution. The persistence of the phase modulation is further shown

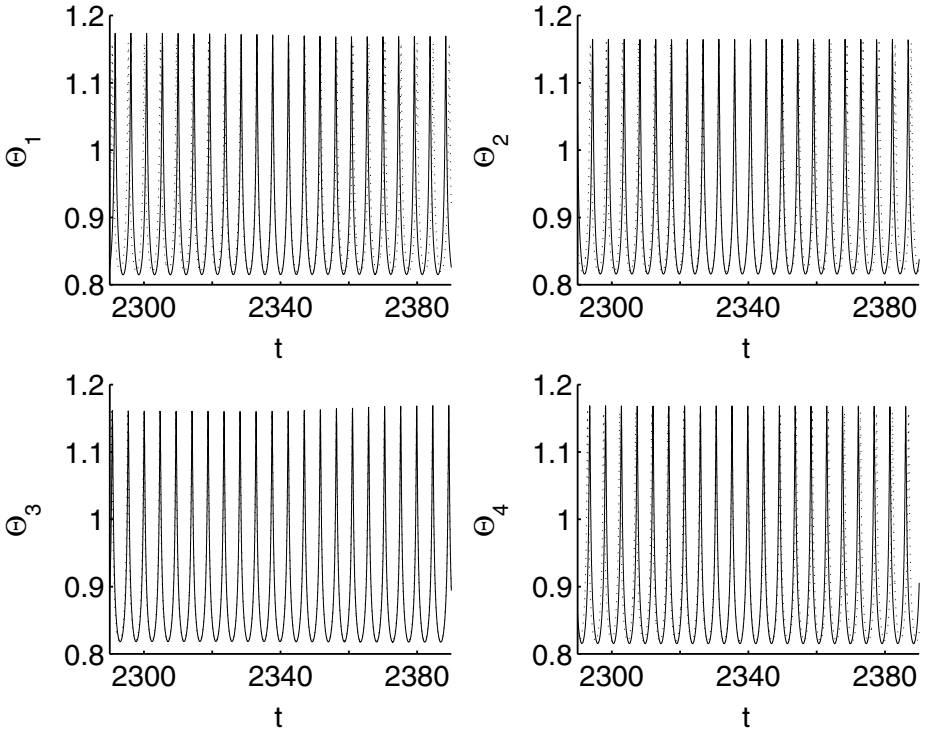


FIG. 18. $\Theta_{Fi}(t)$ $(i = 1, 2, 3, 4)$ *(solid) and an appropriately phase shifted PP solution (dotted) for* $Z = 4.3$ *and* $R = 19$.

in Figure 22, where the maximum and minimum times between the firings of rods 2 and 3 is plotted against R. We note that these extrema asymptotically approach nonzero values (approximately 2.8 and 0.8, respectively). Though we have described the firings of rods 2 and 3, we note that the same behavior is observed for the phase modulation between any pair of adjacent outer rods.

Thus, these results suggest that at least for one-headed spin modes, the behavior for large sample sizes is dominated by a dynamical transformation of the PP solution, which manifests itself as a QP spin mode, with the outer rods exhibiting approximately PP behavior with a phase shift between neighboring rods.

We next describe results for $Z = Z_{2T} = 4.6$. For this value of Z, the pulsating planar solution is period doubled, exhibiting both large amplitude and small amplitude bursts within one cycle. Analogous to the $Z = Z_T$ case, for small values of R ($R \leq 1.3$) the only stable solution that we find is the $2T$ pulsating planar solution; i.e., the subharmonic is now present for the PP solution. For $1.3 < R < 1.5$ we find a region of bistability with both PP solutions and spin solutions. As R increases, we find that the PP solution becomes unstable and there is a branch of spin solutions ($1.3 \leq R \leq 1.7$). Increasing R further, we find a transition to QP spin modes ($1.75 \leq R \leq 4$). These solutions are very similar to those found for $Z = Z_T$ (see Figures 11 and 13) and will not be described further, except to note that the behavior for these modes is singly periodic as opposed to the $2T$ PP solution. Thus, for this range of R there is no indication in the spin modes that the PP solution is $2T$. This is similar to the behavior found for surface combustion in [2].

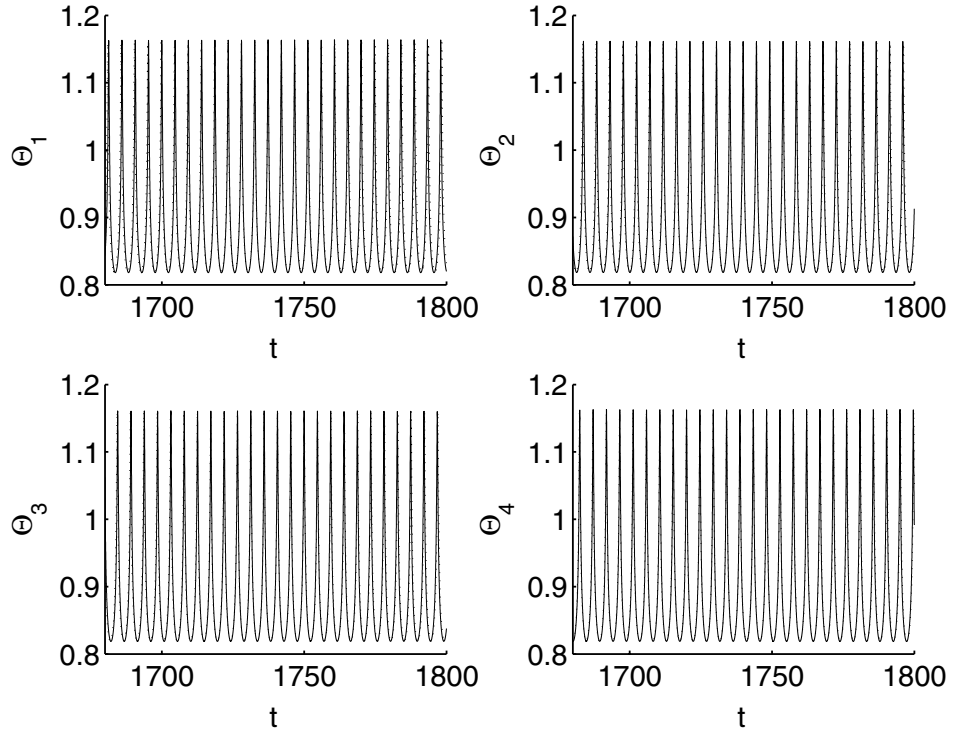


FIG. 19. $\Theta_{Fi}(t)$ $(i = 1, 2, 3, 4)$ *(solid) and an appropriately phase shifted PP solution (dotted) for* $Z = 4.3$ *and* $R = 35$.

As R increases, subharmonics of the generator frequencies develop, giving the solution something of the character of a period doubled solution ($R \simeq 4.8$). We also find intervals where the solution is apparently chaotic, ($5 \leq R \leq 5.6$) and ($6 \leq R \leq 6.78$).

For $6.8 \leq R \leq 9.1$ we find direction reversing modes where two rods exhibit T behavior while the other two exhibit $2T$ behavior, and we observe a reversal of the spin direction around the cylinder. Note that the case of four coupled oscillators with the property that two of them exhibit period $2T$ behavior, though out of phase with each other by T, while the remaining two oscillators exhibit period T behavior, was discussed in [6], though the relation of such a mode to direction reversing behavior was not discussed. In order to describe these solutions we recall the numbering convention whereby rod 1 is located on the axis and rods 2, 3, and 4 are on the cylindrical surface. For the direction reversing modes the axial rod is singly periodic with period T, say. The surface rods are also periodic, but two of them, rods 3 and 4, exhibit period doubled behavior due to a direction reversal of the firing order. Note that rods 3 and 4 are out of phase by T.

In order to understand the dynamics of this mode, suppose for concreteness that rod 2 has period T while rods 3 and 4 have periods $2T$. Consider a sequence of firings commencing with a firing of rod 3, and suppose that rod 2 is the next rod to fire. An ordinary spin mode would then fire in the order 324, 324, ..., thus describing a specific direction to the spin around the cylinder, i.e., clockwise or counterclockwise. The period would be the time interval between two consecutive firings of any of the

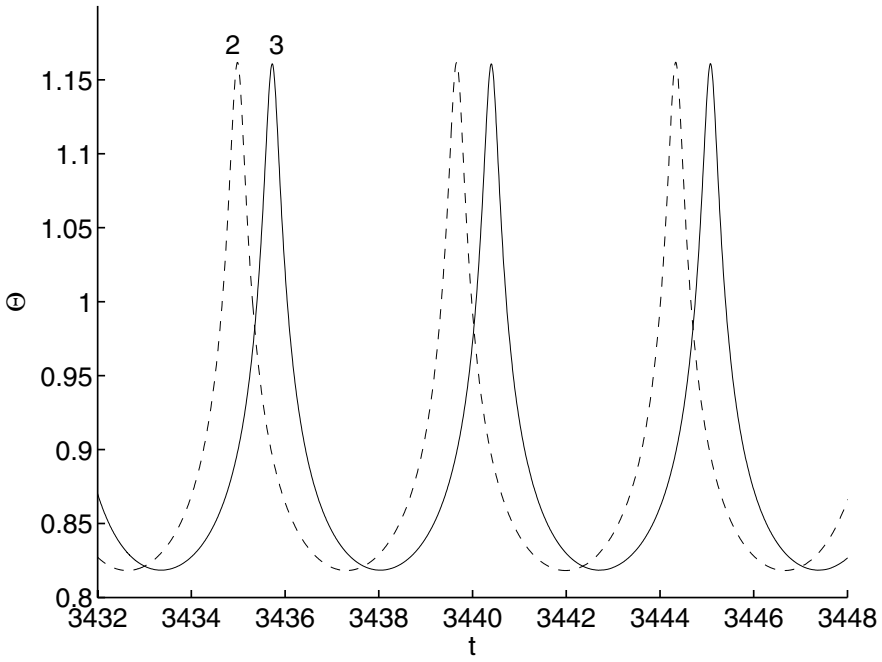


FIG. 20. $\Theta_{F2}(t)$ *and* $\Theta_{F3}(t)$ *plotted for small time interval near* $t = 3440$ *for* $Z = 4.3$ *and* $R = 35$.

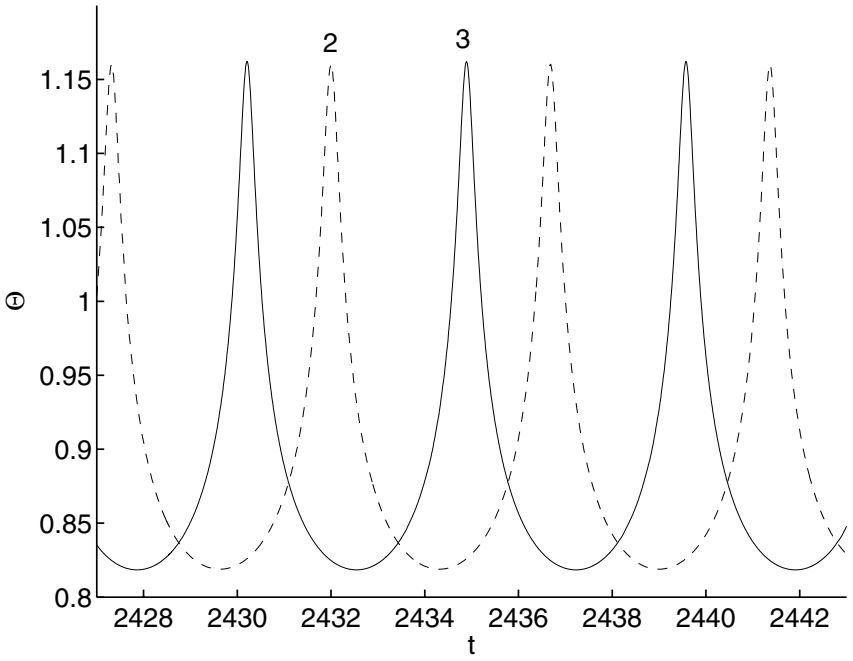


FIG. 21. $\Theta_{F2}(t)$ *and* $\Theta_{F3}(t)$ *plotted for small time interval near* $t = 2440$ *for* $Z = 4.3$ *and* $R = 35$.

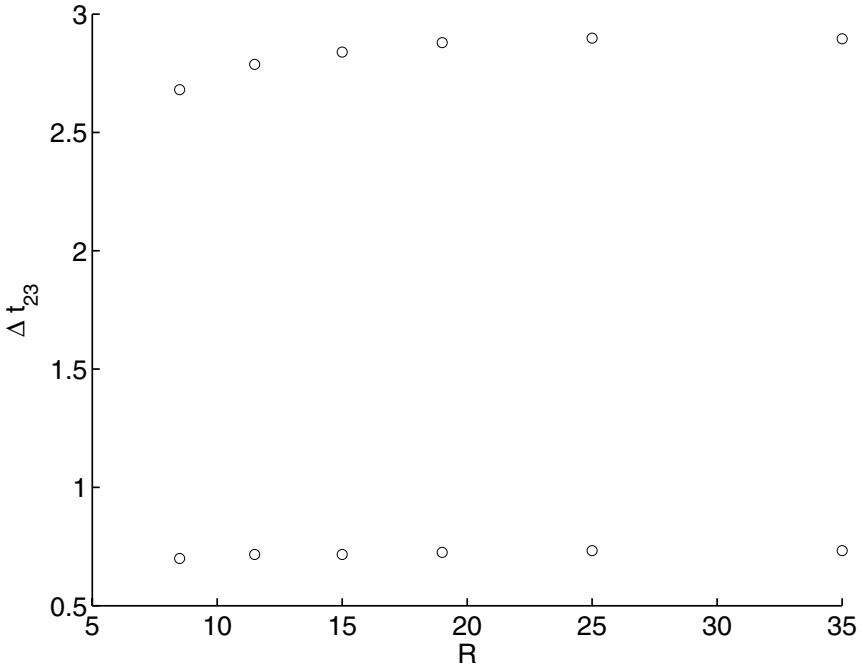


FIG. 22. *Maximum and minimum time difference between successive firings of rods* 2 *and* 3 *for QP solutions as a function of* R *for* $Z = 4.3$.

rods. In the direction reversing mode, the firing direction reverses after one complete circuit around the cylinder. Thus, the firing order would be 324, 423, 324, 423, ... Rods 3 and 4 fire *twice* in succession, signaling a reversal of direction. Furthermore, the two successive firings are at different amplitudes, whereas rod 2 always fires at the same amplitude. An overall period consists of two groups of three, or equivalently two circuits around the cylinder, with a direction reversal. If we denote the period between successive firings of rod 2 as T, then rods 3 and 4 have periods $2T$. The axial rod (rod 1), which has period T, fires between successive firings of rods 3 and 4, so that a firing of the axial rod can be thought of as triggering the reversal. We illustrate this mode for $R = 7$ in Figure 23, where we plot $\Theta_{Fi}(t)$ for rods 3, 2, 4, and 1 over a time interval containing a complete cycle. The figure illustrates the firing sequence 324, 423. The first firing shown is for rod 3 (denoted by A in the figure). Then, rods 2 and 4 fire in order (B and C, respectively). Prior to the direction reversal, the axial rod (rod 1) fires (D). The firing order reverses as rod 4 fires again, but at a smaller amplitude (E). Then, rod 2 fires at the same amplitude at which it previously fired, indicating its period T (F). Finally, rod 3 fires at a larger amplitude than it previously did (G). The figure includes the start of the next reversal, with the subsequent axial firing (H) followed by another firing of rod 3 (I) at the same level as its initial firing. The time between firings A and I is $2T$.

Thus, there are two frequencies associated with this solution, the period between successive firings and the period of the reversal. For this value of R, the two frequencies are commensurate and the solution is periodic. For larger values of R, the two frequencies no longer appear to be commensurate, and we find QP direction reversing

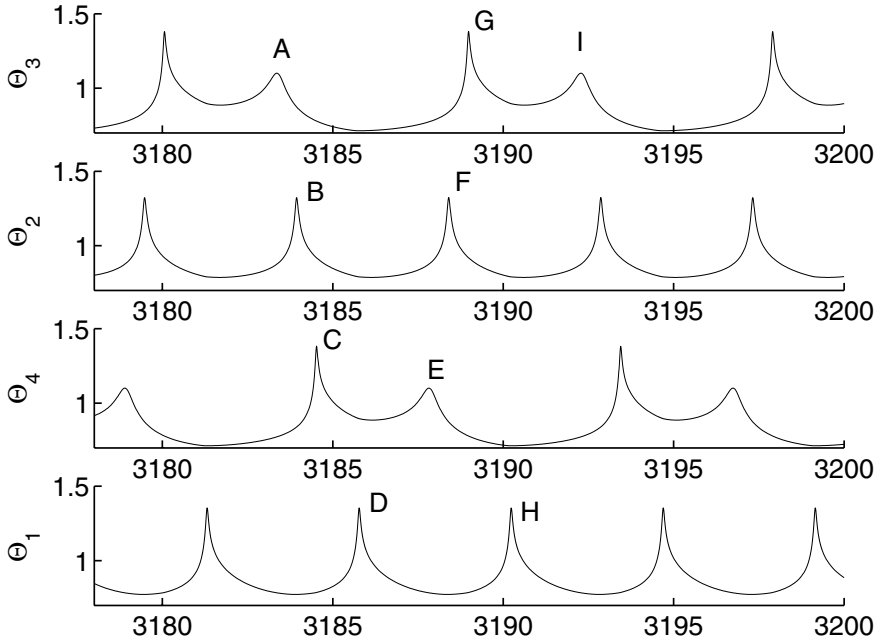


FIG. 23. $\Theta_{Fi}(t)$ $(i = 1, 2, 3, 4)$ *for periodic direction reversing mode for* $Z = 4.6$ *and* $R = 7$.

modes which exhibit similar dynamics to those described above but in a QP rather than periodic fashion. Upon increasing R yet further, we find a window of apparently chaotic solutions. Thus, the direction reversing mode represents an isolated window (in R) of laminar behavior, surrounded on both sides by intervals of apparently chaotic behavior.

Finally, for $R = 19$, we find a QP solution whose oscillations are periodically modulated. The envelopes, corresponding to the modulation frequency, exhibit a symmetry like that observed in [6] for a periodic solution. Specifically, the envelopes of $\Theta_{F3}(t)$ and $\Theta_{F4}(t)$ oscillate in synchrony with period T; thus, they have a common waveform. The envelopes of $\Theta_{F1}(t)$ and $\Theta_{F2}(t)$ also oscillate with period T, but they are not synchronous with each other nor with the envelopes of $\Theta_{F3}(t)$ or $\Theta_{F4}(t)$; they have different waveforms. This solution is shown in Figure 24. As R is increased, the amplitudes of the modulation oscillations decrease while their period increases.

5. Summary. We have numerically simulated 3D modes of solid flame waves employing limited computational resources, by employing a rod model in which a small number of appropriately located 1D rods interact with each other via heat transfer. The heat transfer coefficients correspond to a coarse-grained discretization of the transverse Laplacian. Both the rod model and the full 3D model allow uniformly propagating planar waves. By appropriately relating the heat transfer coefficients in the rod model to the radius R in the full 3D model, the dispersion relations for the two models can be shown to be identical.

We consider a simple rod model involving three outer rods and one axial rod, thus limiting the class of modes that can be described and for which analogies can be drawn with the modes of the 3D model. Clearly, additional modes, e.g., multiheaded spin modes, can be described by a rod model with a larger number of rods.

We have described the detailed behavior of the model as the sample size increases

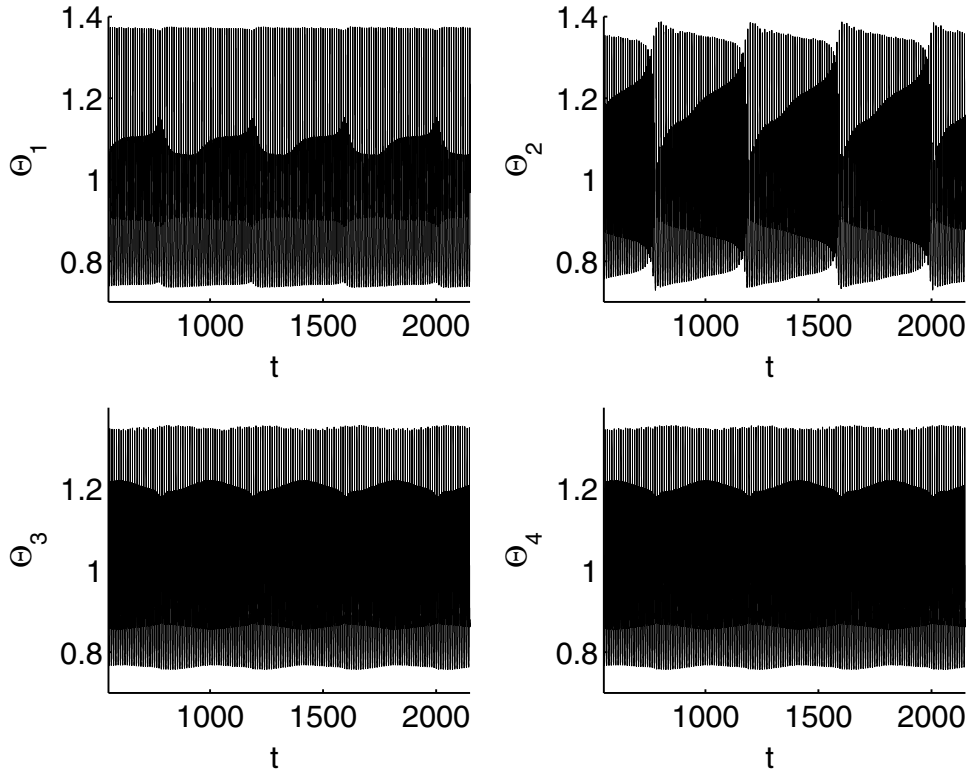


FIG. 24. $\Theta_{Fi}(t)$ $(i = 1, 2, 3, 4)$ *for QP solution with* $Z = 4.6$ *and* $R = 19$.

for three values of Z, one below and two above the pulsating stability boundary. When Z is below the stability boundary, there are transitions from the uniform mode to spin behavior as R increases. For larger values of R periodic radial modes are found. There is a branch of QP modes connecting the spin and radial branches, which involve an interaction between spin and radial behavior. Spin behavior dominates for smaller values of R (near the spin-QP transition), and radial behavior dominates for larger values of R (near the QP-radial transition). There is no indication of bistability for this value of Z.

For Z above the pulsating stability threshold, we find PP solutions for small R and spin modes for larger values of R, with a region of bistability between them. For larger values of R, only QP spin modes are found. Spectral analysis of the QP modes indicates that they represent the interaction of spin and radial modes. As R increases, the frequency spectrum of the QP modes approaches that of the PP solution, with a phase shift between adjacent outer rods. Thus, the pulsation within each rod is approximately that of the PP solution. We term this behavior SPP behavior. The behavior for large R can be described as a dynamical transformation of the PP mode, stable for small R, into a QP spin mode. This behavior persists over intermediate time scales which increase with R. Over long time scales there is persistent phase modulation between successive firings of the outer rods while the amplitude modulation decays to 0. Thus, the solutions remain QP with phase modulations manifested over progressively longer time scales as $R \to \infty$.

For yet larger values of Z, so that the PP solution is period doubled, we find spin and QP spin modes as before. We also find a window (in R) where direction reversing modes exist and are stable. These modes are characterized by a firing sequence which reverses direction, either periodically or quasi-periodically. Thus, outer rods fire in a sequence, e.g., 324, 423, 324, 423,..., so that the direction of firings is reversed. The middle rod in this sequence, in this case rod 2, is periodic with period T, say. The end rods, rods 3 and 4, fire consecutively as the sequence reverses. These firings are at different amplitude levels, so that these rods are periodic with period $2T$. The axial rod fires between two consecutive firings of rods 3 and 4. Thus, a firing on the axis can be thought of as triggering the reversal.

Our rod model is designed to provide insight into the qualitative behavior of the full 3D problem, employing vastly reduced computational resources. The computations presented here are for a primitive rod model, consisting of only four rods in total. The model is very computationally efficient, although it is incapable of reproducing multiheaded spins and standing wave–type solutions exhibiting, e.g., creation and annihilation, as is found in, e.g., surface gasless combustion [2]. Thus, the model can reproduce at most a limited subset of the solutions that can occur for the full 3D problem, although the results in [2] suggest that branches involving primarily one spot may extend stably for relatively large values of R and thus describe behavior for large diameter samples. The equivalence of the dispersion relation between the model and the full 3D problem suggests that 1-headed spin and radial-type behavior is reproduced at least qualitatively by the model. Whether the nonlinear behavior that we have observed is in fact also observed in the 3D problem has not yet been determined, though some of the computations described in [7, 8] may exhibit QP spin behavior. We should point out that the SPP mode (observed for large R) and the direction reversing mode (observed for large Z) occur in the regimes where numerical computations are most difficult and computationally intensive. Of course, if the number of rods is sufficiently large, the results for the rod model will surely closely approximate those for the full 3D model. However, our interest is in employing a finite, indeed a relatively small, number of rods. In addition to the question regarding the discretization, i.e., the use of a small number of rods, there is also the question of whether certain dynamical behaviors are a reflection of the symmetries of the rod arrangement. In view of this, it is important to ascertain whether such modes are observed for rod models involving a larger number of rods, arranged to exhibit different symmetries. We have obtained preliminary results with models involving more rods, which indicate that many of the modes reported here persist as the number of rods is increased, thus suggesting that they occur in the full 3D model as well.

REFERENCES

[1] A. P. Aldushin, A. Bayliss, and B. J. Matkowsky, *Dynamics of layer models of solid flame propagation*, Phys. D, 143 (2000), pp. 109–137.

[2] A. Bayliss, B. J. Matkowsky, and A. P. Aldushin, *Dynamics of hot spots in solid fuel combustion*, Phys. D, 166 (2002), pp. 104–130.

[3] A. Bayliss, D. Gottlieb, B. J. Matkowsky, and M. Minkoff, *An adaptive pseudo-spectral method for reaction diffusion problems*, J. Comput. Phys., 81 (1989), pp. 421–443.

[4] A. Bayliss and B. J. Matkowsky, *Fronts, relaxation oscillations, and period doubling in solid fuel combustion*, J. Comput. Phys., 71 (1987), pp. 147–168.

[5] A. Bayliss and B. J. Matkowsky, *Two routes to chaos in condensed phase combustion*, SIAM J. Appl. Math., 50 (1990), pp. 437–459.

[6] M. Golubitsky and I. Stewart, *Hopf bifurcation with dihedral group symmetry: Coupled nonlinear oscillators*, Contemp. Math., 56 (1986), pp. 131–173.

[7] T. P. Ivleva and A. G. Merzhanov, *Mathematical simulation of three dimensional spinning modes of gasless combustion waves,* Dokl. Phys., 44 (1999), pp. 739–744.

[8] T. P. Ivleva and A. G. Merzhanov, *Three dimensional spinning waves in the case of gas free combustion,* Dokl. Phys., 45 (2000), pp. 136–141.

[9] S. B. Margolis, *An asymptotic theory of condensed two-phase flame propagation,* SIAM J. Appl. Math., 43 (1983), pp. 351–369.

[10] S. B. Margolis, H. G. Kaper, G. K. Leaf, and B. J. Matkowsky, *Bifurcation of pulsating and spinning reaction fronts in condensed two-phase combustion,* Comb. Sci. and Tech., 43 (1985), pp. 127–165.

[11] B. J. Matkowsky and G. I. Sivashinsky, *Propagation of a pulsating reaction front in solid fuel combustion*, SIAM J. Appl. Math., 35 (1978), pp. 465–478.

[12] A. G. Merzhanov, *Solid flames: Discovery, concepts, and horizons of cognition*, Comb. Sci. and Tech., 98 (1994), pp. 307-336

[13] J. H. Park, A. Bayliss, and B. J. Matkowsky, *On the transition from spinning to radial solid flame waves*, Appl. Math. Lett., 17 (2004), pp. 123–131.

SIAM J. APPL. MATH.
Vol. 65, No. 2, pp. 550–566

A DELAY REACTION-DIFFUSION MODEL OF THE SPREAD OF BACTERIOPHAGE INFECTION*

STEPHEN A. GOURLEY† AND YANG KUANG‡

Abstract. This paper is a continuation of recent attempts to understand, via mathematical modeling, the dynamics of marine bacteriophage infections. Previous authors have proposed systems of ordinary differential delay equations with delay dependent coefficients. In this paper we continue these studies in two respects. First, we show that the dynamics is sensitive to the phage mortality function, and in particular to the parameter we use to measure the density dependent phage mortality rate. Second, we incorporate spatial effects by deriving, in one spatial dimension, a delay reaction-diffusion model in which the delay term is rigorously derived by solving a von Foerster equation. Using this model, we formally compute the speed at which the viral infection spreads through the domain and investigate how this speed depends on the system parameters. Numerical simulations suggest that the minimum speed according to linear theory is the asymptotic speed of propagation.

Key words. reaction-diffusion, delay, stage structure, through-stage death rate, traveling wave

AMS subject classifications. 92D25, 35K57, 35R10

DOI. 10.1137/S0036139903436613

1. Introduction.

It is known that bacteriophage infection can be a significant mechanism of mortality in marine prokaryotes (Bergh et al. [6], Proctor and Fuhrman [16]). These mortality mechanisms are critical in understanding the marine production processes. The constituents released by cell lysis can be an important pathway of nutrient recycling. This has direct bearing on issues such as global warming and topics of geochemical cycles. Viral infection also has direct implications for genetic exchange in the sea (Lenski and Levin [14], Bohannan and Lenski [7]).

Although we do not yet have a good understanding of the temporal or spatial scales at which host-virus encounters occur, it is clear that viral mortality must be explicitly considered in most models of the marine system. A case in point, recent experimental work suggests that the contamination of algal cells by viruses can serve as a regulatory mechanism in its bloom dynamics. Beltrami and Carroll [1] formulated a simple trophic model including virus-induced mortality. Their model succeeded in mimicking the actual algal bloom patterns of several species.

Our main interest in this paper is to explore how viral mortality affects both the temporal and spatial dynamics of marine bacteria and cyanobacteria. Recently, Beretta and Kuang [4] formulated and carried out a detailed study of the temporal viral-bacteria model

$$\begin{aligned}
\frac{dS}{dt} &= \alpha S(t)\left(1-\frac{S(t)+I(t)}{C}\right) - KS(t)P(t),\\
\frac{dI}{dt} &= -\mu_i I(t) + KS(t)P(t) - e^{-\mu_i T}KS(t-T)P(t-T),\\
\frac{dP}{dt} &= \beta - \mu_p P(t) - KS(t)P(t) + be^{-\mu_i T}KS(t-T)P(t-T).
\end{aligned} \tag{1.1}$$

*Received by the editors October 24, 2003; accepted for publication (in revised form) May 17, 2004; published electronically January 5, 2005.
http://www.siam.org/journals/siap/65-2/43661.html

†Corresponding author. Department of Mathematics and Statistics, University of Surrey, Guildford, Surrey GU2 7XH, UK (s.gourley@surrey.ac.uk). Correspondence should be directed to this author.

‡Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287 (kuang@asu.edu). The research of this author was partially supported by NSF grant DMS-0077790.

This system of delay differential equations models a population of marine bacteria in which the individuals are subject to infection by viruses, also known as bacteriophages. Prior to that, these authors (Beretta and Kuang [2]) modeled and studied the same process by a set of nonlinear ordinary differential equations, and Carletti [10] has studied the stochastic extension of that model. In system (1.1), S is the density (i.e., number of bacteria per liter) of susceptible bacteria, I is the density of infected bacteria, and P is the density (number of viruses per liter) of viruses (phages). Viruses P attack the susceptible bacteria S, and a bacterium becomes infected I when a virus successfully injects itself through the bacterial membrane. The virus then starts replicating inside the bacterium, and then all the bacterium's resources are directed to replication of the virus. The infected bacterium does not replicate itself by division; only susceptible bacteria are capable of doing so. After a latency time T, an infected bacterium will die by lysis; i.e., the bacterium explodes releasing b copies ($b > 1$) of the virus into the solution, which are then free to attack other susceptible bacteria. An infected bacterium may die other than by viral lysis; we allow for this by the term $-\mu_i I(t)$. The differential equation for $I(t)$ is derived from the fact that $I(t)$ is given by

$$I(t) = \int_0^T e^{-\mu_i \tau} KS(t-\tau)P(t-\tau)\, d\tau, \tag{1.2}$$

which expresses the fact that the number of recruits into the infected class between times $t-(\tau+d\tau)$ and $t-\tau$ is $KS(t-\tau)P(t-\tau)\,d\tau$, the number of these still alive at time t is obtained by multiplying by $e^{-\mu_i\tau}$, and then the integral totals up the contributions from all relevant previous times, i.e., up to T time units ago.

In the virus equation, the third equation of (1.1), all mortalities of viruses are accounted for by the term $-\mu_p P(t)$. The β term, where $\beta > 0$, models a constant inflow of phages from outside the system. In the absence of viruses the bacteria grow logistically. The rate of infection is given by the law of mass action to be $KS(t)P(t)$.

Beretta and Kuang [4] assumed that infected bacteria still compete with susceptible bacteria for common resources. This is represented by the $-(S+I)/C$ term in the first equation of (1.1). This is clearly a disputed subject. For example, a model by Campbell [8] consists of the following equations:

$$\begin{aligned} \frac{dS(t)}{dt} &= \alpha S(t)\left(1-\frac{S(t)}{C}\right) - KS(t)P(t), \\ \frac{dP(t)}{dt} &= bKS(t-T)P(t-T) - \mu_p P(t) - KS(t)P(t), \end{aligned} \tag{1.3}$$

where

$$I(t) = \int_{t-T}^{t} KS(\theta)P(\theta)d\theta. \tag{1.4}$$

Clearly, in (1.3) the competition for common resources and additional mortality rate endured by infected bacteria is neglected. The equations (1.3), (1.4) can be obtained from (1.1) by setting $\beta = 0$, $\mu_i = 0$. Extensions of the above Campbell model can be found in Beretta, Carletti, and Solimano [3] (taking into account environmental fluctuations) and Carletti [9] (replacing b by $be^{-\mu_i T}$).

In the present paper, like the model of Campbell [8], we assume that once a bacterium becomes infected by a virus, it no longer competes with susceptibles for

resources. We will allow the possibility of a density dependent mortality term in the phage equation. In Beretta and Kuang [4], and also in the present paper, it is assumed that T and b are constant and the same for the whole population. Modifications of this assumption (e.g., replacing the constant incubation time T by a distribution of incubation times modeled using a probability density function) are the subject of further work presently in progress.

In the next section, we will present our delay model of bacteriophage infection and a simple preliminary result on the positivity of its solutions. This is followed by a short section on the global stability of the disease-free equilibrium. The analysis of endemic equilibrium is highly nontrivial and we provide only generic conditions for its stability switch. To complement this analytic work, we present some carefully designed and data-based simulation results. We then proceed to formulate and study a delay reaction-diffusion model of the spread of bacteriophage infection. The paper ends with a discussion.

2. Preliminaries. Most of our effort will be devoted to understanding the system

$$\begin{aligned} S'(t) &= \alpha S(t)\left(1-\tfrac{S(t)}{\gamma}\right) - KS(t)P(t), \\ P'(t) &= -\mu_p P(t) - mP^2(t) - KS(t)P(t) + bKe^{-\mu_i T}S(t-T)P(t-T), \end{aligned} \tag{2.1}$$

and with a reaction-diffusion version of (2.1). The initial conditions for (2.1) are

$$\begin{aligned} S(s) &= S^0(s) \geq 0,\ s \in [-T,0], \quad \text{with } S^0(0) > 0, \\ P(s) &= P^0(s) \geq 0,\ s \in [-T,0], \quad \text{with } P^0(0) > 0, \end{aligned} \tag{2.2}$$

where S^0 and P^0 are prescribed continuous functions. Our system (2.1) differs from that studied in [4] in three respects: (i) we do not have an inflow of phages from outside the system, (ii) we allow the possibility of a density dependent mortality term (the term $-mP^2$ in (2.1)), and (iii) we assume that an infected bacterium no longer competes with the susceptibles for resources. The latter assumption means that we do not need the differential equation for $I(t)$ for the analysis (though $I(t)$ is still given by (1.2)). An additional difference is that in the present paper we shall consider the effects of including diffusion to model the motion of the phages and bacteria.

If we had $P^0(s) \equiv 0$ on $[-T,0]$, the method of steps would immediately yield $P(t) = 0$ for all $t > 0$. The dynamics of $S(t)$ would then be governed by the logistic equation. Similarly, if $S^0(s) \equiv 0$, then clearly $S(t)$ remains zero for all $t > 0$ and thus $P(t) \to 0$ as $t \to \infty$. These trivial cases are removed from consideration by the assumptions in (2.2).

PROPOSITION 1. *Solutions of* (2.1), (2.2) *satisfy* $S(t) > 0$, $P(t) > 0$ *for all* $t > 0$.

Proof. The equation for $S(t)$ in (2.1) contains a factor of $S(t)$ and therefore positivity for $S(t)$ follows by the standard argument. For $P(t)$, note that on $t \in [0,T]$ we have $P'(t) \geq -\mu_p P(t) - mP^2(t) - KS(t)P(t)$ so that $P(t) \geq \tilde{P}(t)$, where $\tilde{P}$ is the solution of $\tilde{P}'(t) = -\mu_p \tilde{P}(t) - m\tilde{P}^2(t) - KS(t)\tilde{P}(t)$ satisfying $\tilde{P}(0) = P(0) > 0$. Clearly $\tilde{P}(t) > 0$ for all $t > 0$, and so we conclude that $P(t) > 0$ for all $t > 0$. The proof is complete. ∎

3. Equilibria and their stability. The equilibria of (2.1) are $(S,P) = (0,0)$, the disease-free equilibrium $(\gamma, 0)$, and possibly an endemic equilibrium

$$(S^*, P^*) := \left(\frac{m\gamma\alpha + K\gamma\mu_p}{m\alpha + K^2\gamma(be^{-\mu_i T} - 1)},\ \frac{\alpha\gamma K(be^{-\mu_i T} - 1) - \alpha\mu_p}{m\alpha + K^2\gamma(be^{-\mu_i T} - 1)} \right). \tag{3.1}$$

The latter is ecologically relevant if and only if

$$be^{-\mu_i T} > 1 + \frac{\mu_p}{\gamma K}, \tag{3.2}$$

which, of course, can only possibly hold for T up to a finite value. As long as (3.2) holds, there is an endemic equilibrium. Note that as $m \to \infty$ the endemic equilibrium approaches the disease-free equilibrium $(\gamma, 0)$.

We shall first prove that, if condition (3.2) does not hold, then any positive solution approaches the disease-free equilibrium $(\gamma, 0)$.

THEOREM 1. *Assume that*

$$be^{-\mu_i T} \leq 1 + \frac{\mu_p}{\gamma K}.$$

Then any solution of (2.1), (2.2) *satisfies*

$$\lim_{t \to \infty} (S(t), P(t)) = (\gamma, 0).$$

Proof. Consider the positive definite functional

$$V = S - \gamma - \gamma \ln \frac{S}{\gamma} + \frac{\gamma K}{\mu_p} P + \frac{b\gamma K^2}{\mu_p} e^{-\mu_i T} \int_{t-T}^{t} S(s)P(s)\, ds.$$

Differentiating along solutions of (2.1) yields

$$\begin{aligned} V' &= -\frac{\alpha}{\gamma}(S-\gamma)^2 - \frac{\gamma m K}{\mu_p} P^2 + K\left(\frac{b\gamma K}{\mu_p} e^{-\mu_i T} - \frac{\gamma K}{\mu_p} - 1\right) SP \\ &\leq -\frac{\alpha}{\gamma}(S-\gamma)^2. \end{aligned}$$

Thus

$$V(t) + \frac{\alpha}{\gamma} \int_0^t (S(s) - \gamma)^2\, ds \leq V(0),$$

and, letting $t \to \infty$, we conclude that $|S(t) - \gamma| \in L^2(0, \infty)$ so that $S(t) \to \gamma$ as $t \to \infty$. The differential equations (2.1) then yield $P(t) \to 0$. The proof is complete. ∎

3.1. The endemic equilibrium: Linearized analysis. Let us investigate the endemic equilibrium (S^*, P^*) given by (3.1). In this subsection we shall assume, of course, that (3.2) holds, so that the equilibrium is feasible. The linearized analysis about the endemic equilibrium is algebraically quite complicated. The main reason for this is that the delay T appears not only in the $S(t-T)P(t-T)$ term in the second equation of (2.1), but also in the factor $e^{-\mu_i T}$ in front of that term. The paper by Wolkowicz, Xia, and Wu [20] shows how such additional factors involving time delay can appear in distributed delay equations. Surprisingly, this represents a significant complication and prevents us from analytically computing the precise parameter regimes in which the endemic equilibrium can change stability as the delay T is increased, or the actual values of T when stability switches occur. Note further that the equilibrium itself depends on T and exists only for T up to a finite value. This renders many of the existing stability switch methods (see Kuang [12]) powerless. However, a method has recently been developed by Beretta and Kuang [5] to address the problem of computing stability switches for delay equations which do not lend

themselves to classical methods because of these complications. We shall use this method in this section.

To linearize about (S^*, P^*) we set $S = S^* + \tilde{S}$ and $P = P^* + \tilde{P}$. Ignoring higher order terms in $\tilde{S}, \tilde{P}$ gives us the linearized system

$$\begin{aligned} \tilde{S}'(t) &= -\tfrac{\alpha}{\gamma} S^* \tilde{S}(t) - K S^* \tilde{P}(t), \\ \tilde{P}'(t) &= -K P^* \tilde{S}(t) - (\mu_p + 2mP^* + KS^*)\tilde{P}(t) \\ &\quad + bKe^{-\mu_i T}(P^* \tilde{S}(t-T) + S^* \tilde{P}(t-T)). \end{aligned} \tag{3.3}$$

We shall find it convenient to introduce the parameter

$$\rho_T = \frac{\gamma K}{\mu_p}(be^{-\mu_i T} - 1). \tag{3.4}$$

Then the endemic equilibrium (S^*, P^*) exists if and only if

$$\rho_T > 1.$$

In terms of ρ_T,

$$(S^*, P^*) = \left(\frac{\gamma(m\alpha + K\mu_p)}{m\alpha + K\mu_p \rho_T}, \ \frac{\alpha \mu_p (\rho_T - 1)}{m\alpha + K\mu_p \rho_T} \right).$$

Nontrivial solutions of the linearized system of the form $(\tilde{S}(t), \tilde{P}(t)) = e^{\lambda t}(c_1, c_2)$ exist if and only if

$$D(\lambda; T) = 0,$$

where

$$D(\lambda; T) = \lambda^2 + a(T)\lambda + b(T)\lambda e^{-\lambda T} + c(T) + d(T)e^{-\lambda T} \tag{3.5}$$

and

$$a(T) = \frac{\alpha(m\alpha + K\mu_p) + m\alpha(K\gamma + 2\mu_p\rho_T - \mu_p) + K\mu_p(K\gamma + \mu_p\rho_T)}{m\alpha + K\mu_p\rho_T}, \tag{3.6}$$

$$b(T) = -\frac{b\gamma K e^{-\mu_i T}(m\alpha + K\mu_p)}{m\alpha + K\mu_p\rho_T}, \tag{3.7}$$

$$c(T) = \frac{\alpha(m\alpha + K\mu_p)\left\{m\alpha(K\gamma + (2\rho_T - 1)\mu_p) + K\mu_p(\mu_p\rho_T + (2 - \rho_T)K\gamma)\right\}}{(m\alpha + K\mu_p\rho_T)^2}, \tag{3.8}$$

$$d(T) = \frac{bK\gamma\alpha e^{-\mu_i T}(m\alpha + K\mu_p)\left\{K\mu_p(\rho_T - 2) - m\alpha\right\}}{(m\alpha + K\mu_p\rho_T)^2}. \tag{3.9}$$

Keeping in mind that $b > 1$, it is straightforward to see that when $T = 0$ the equilibrium (S^*, P^*), if feasible, is linearly stable. This is because when $T = 0$, (3.5) becomes a quadratic in λ, and it is easy to see that $a(0) + b(0) > 0$ and $c(0) + d(0) > 0$. The question is whether the equilibrium can undergo any stability switch as T is increased, remembering that the equilibrium is only feasible up to a finite value of T. To identify a stability switch we seek solutions of the characteristic equation $D(\lambda; T) = 0$ of the

form $\lambda = \pm i\omega$, with ω a real positive number. We find that it is necessary for ω to satisfy

$$\omega^4 + (a^2(T) - 2c(T) - b^2(T))\omega^2 + c^2(T) - d^2(T) = 0. \tag{3.10}$$

However, the existence for a particular T of a real root $\omega(T)$ of (3.10) does not in itself imply that a stability switch occurs at that value of T, since T also has to satisfy (3.11) and (3.12) below. Nonetheless, certain general analytical conclusions can be drawn in spite of the algebra. Straightforward but tedious computations show that, for any parameter values consistent with $\rho_T > 1$ (i.e., with existence of the endemic equilibrium (S^*, P^*)), we have

$$a^2(T) - 2c(T) - b^2(T) > 0.$$

In light of this fact, and assuming that (S^*, P^*) is feasible when $T = 0$, certain conclusions follow.

(i) A stability switch cannot occur in an interval of T throughout which $c^2(T) > d^2(T)$.

(ii) If there are values of T with $c^2(T) < d^2(T)$, then a stability switch may occur as T is varied. Pairs of eigenvalues cross the imaginary axis as T passes through certain critical values. The critical values of T and the corresponding purely imaginary eigenvalues $\pm i\omega(T)$, $\omega(T) > 0$, are given implicitly by

$$\sin(\omega(T)\,T) = \frac{b(T)\omega(T)(\omega^2(T) - c(T)) + \omega(T)a(T)d(T)}{\omega^2(T)b^2(T) + d^2(T)}, \tag{3.11}$$

$$\cos(\omega(T)\,T) = \frac{d(T)(\omega^2(T) - c(T)) - \omega^2(T)a(T)b(T)}{\omega^2(T)b^2(T) + d^2(T)}, \tag{3.12}$$

$$\begin{aligned}\omega^2(T) = \frac{1}{2}\Big(&- a^2(T) + 2c(T) + b^2(T) \\ &+ \sqrt{a^4(T) - 4a^2(T)c(T) - 2a^2(T)b^2(T) + 4c(T)b^2(T) + b^4(T) + 4d^2(T)}\Big),\end{aligned} \tag{3.13}$$

where $a(T)$, $b(T)$, $c(T)$, and $d(T)$ are given by (3.6), (3.7), (3.8), and (3.9) above. It is impossible to solve these equations for T explicitly, so we shall use the procedure described in Beretta and Kuang [5]. According to this procedure, we define $\theta(T) \in [0, 2\pi)$ such that $\sin\theta(T)$ and $\cos\theta(T)$ are given by the right-hand sides of (3.11) and (3.12), respectively, with $\omega(T)$ given by (3.13). This defines $\theta(T)$ in a form suitable for numerical evaluation using standard software. Then T is given (still implicitly) by

$$T = \frac{\theta(T) + 2n\pi}{\omega(T)}, \qquad n = 0, 1, 2, \ldots,$$

and the idea is to identify the roots of this equation for various n, i.e., to solve numerically the equation $S_n(T) = 0$ for $n = 0, 1, 2$, where

$$S_n(T) = T - \left(\frac{\theta(T) + 2n\pi}{\omega(T)}\right), \qquad n = 0, 1, 2, \ldots. \tag{3.14}$$

Accurate plots of these functions $S_n(T)$ quickly reveal whether stability switches can occur or not, but one must remember to keep track of the feasibility of the equilibrium (S^*, P^*) since it disappears completely (by coalescing with the disease-free equilibrium $(\gamma, 0)$) at a finite value of the delay T.

By reference to (i) above, it is possible to obtain sufficient and easily verifiable conditions for the equilibrium (S^*, P^*) to remain locally stable. Indeed, the condition $c^2(T) > d^2(T)$ amounts to

(3.15)
$$\begin{aligned} &m^2\left\{\alpha^2(K\gamma+(2\rho_T-1)\mu_p)^2-\alpha^2b^2K^2\gamma^2e^{-2\mu_iT}\right\}\\ &+2\alpha K\mu_pm\left\{(K\gamma+(2\rho_T-1)\mu_p)(\mu_p\rho_T+(2-\rho_T)K\gamma)+(\rho_T-2)b^2K^2\gamma^2e^{-2\mu_iT}\right\}\\ &+K^2\mu_p^2(\mu_p\rho_T+(2-\rho_T)K\gamma)^2-K^2\mu_p^2(\rho_T-2)^2b^2K^2\gamma^2e^{-2\mu_iT}>0. \end{aligned}$$

Thus, if (3.15) holds, then (S^*, P^*), if feasible, is locally stable. The coefficient of m^2 in (3.15) is automatically positive if $\rho_T > 1$ (the condition for feasibility of (S^*, P^*)), and therefore one parameter regime in which (3.15) is satisfied is that the parameter m be large.

For the convenience of comparison and computation, we perform the same dimensionless analysis as was carried out in Beretta and Kuang [4]. We choose the dimensionless time as $\tau = K\gamma t$. Note that one unit of the dimensionless time scale, i.e., $\tau = 1$, corresponds to $t_\tau = (1/K\gamma)$ in the original time unit. We also need the dimensionless variables

$$s = \frac{S}{\gamma}, \quad p = \frac{P}{\gamma}.$$

Below are the dimensionless parameters:

$$a = \frac{\alpha}{K\gamma}, \quad m_p = \frac{\mu_p}{K\gamma}, \quad m_i = \frac{\mu_i}{K\gamma}, \quad m_q = \frac{m}{K}.$$

Equations (2.1) have the dimensionless form

(3.16)
$$\begin{cases} \dfrac{ds(\tau)}{d\tau} = as(\tau) - as^2(\tau) - s(\tau)p(\tau), \\ \dfrac{dp(\tau)}{d\tau} = -m_pp(\tau) - m_qp^2(\tau) - s(\tau)p(\tau) + be^{-m_iT_\tau}s(\tau - T_\tau)p(\tau - T_\tau). \end{cases}$$

The values for the dimensionless parameters and the dimensionless time scale are taken from the model of Beretta and Kuang [4] (the original parameter estimates are due to Okubo). They are

(3.17)
$$a = 10, \quad m_p = 14.925,$$

with $t_\tau = (1/K\gamma) = 7.4627$ days and an average latency time $T \simeq 0.303$ days. We have no estimates for $m_i = (\mu_i/K\gamma)$, but it seems reasonable to assume it is smaller than m (since the main cause of mortality is the lysis of infected cells). We assume $m_i \simeq 0.1m_p$. In addition, we do not have an estimate on m_q. In the following computational work, we assume that $m_q \simeq 0.1$, a value close to zero. Figure 1 is the result of an application of the stability switch theory of Beretta and Kuang [5] for this set of parameters (except that we vary the latency period).

Figure 2 provides simulation results for the above set of parameters with four representative values of latency periods. Clearly Figure 2 confirms the findings embodied in Figure 1.

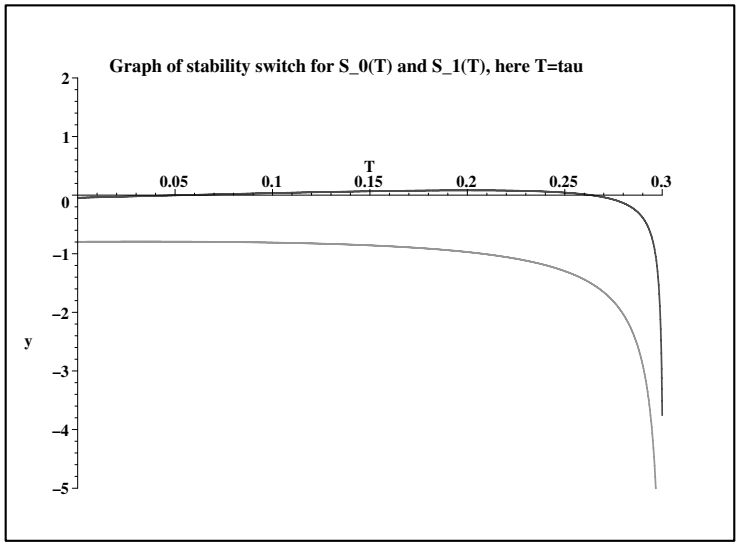


FIG. 1. *Plots of the functions* $S_0(\tau)$ *(upper curve) and* $S_1(\tau)$ *(lower curve). Parameter values used are* $\mu_p = 14.925$, $b = 75$, $\mu_i = 1.5$, $\alpha = 10$, *and* $m = 0.1$. *The equilibrium is feasible for* $0 \le \tau < \ln(b/(1+\mu_p))/\mu_i \equiv \tau_e \approx 1.033$.

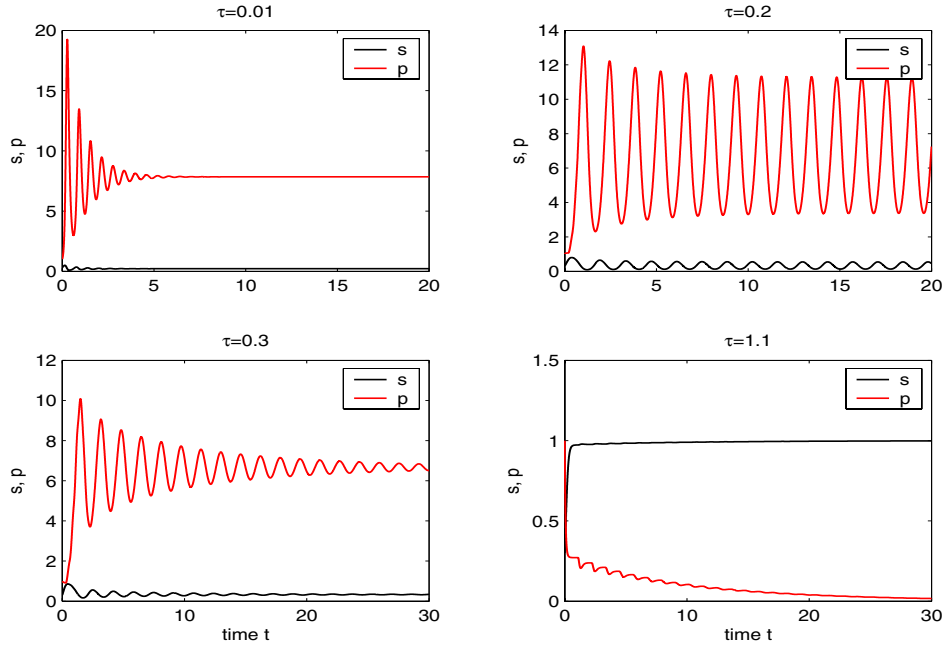


FIG. 2. *A solution of model* (3.16) *with* $s(\theta) = 0.3$, $p(\theta) = 1$, $\theta \in [-\tau, 0]$, *where* $\mu_p = 14.925$, $b = 75$, $\mu_i = 1.5$, $\alpha = 10$, $m = 0.1$, *and* τ *varies from* 0.01 *to* 1.1.

4. Diffusive models. In this section we propose some reaction-diffusion extensions of system (2.1). The main issues here are (i) what types of diffusion are appropriate, and (ii) derivation of the time-delay terms for the case when there is diffusion. The latter point is important because infectives can move during the period between infection and lysis, so that when an infective dies by lysis it will release the b copies of the virus into the water at a different location from where it originally became infected. We shall show how this can be accounted for in the modeling by including time and age as independent variables and using an age-structured model approach. The approach described here has also been used by many other investigators (see, e.g., Smith [17], So, Wu, and Zou [18], and Gourley and So [11]).

For the simplest case of Fickian diffusion, and working on an infinite one-dimensional domain $-\infty < x < \infty$, system (2.1) becomes

$$\begin{aligned} \frac{\partial S(x,t)}{\partial t} &= D_s \frac{\partial^2 S(x,t)}{\partial x^2} + \alpha S(x,t)\left(1 - \frac{S(x,t)}{\gamma}\right) - KS(x,t)P(x,t), \\ \frac{\partial P(x,t)}{\partial t} &= D_p \frac{\partial^2 P(x,t)}{\partial x^2} - \mu_p P(x,t) - mP^2(x,t) - KS(x,t)P(x,t) \\ &\quad + b \times \{\text{rate of death of infectives by lysis}\}, \end{aligned} \tag{4.1}$$

where D_s and D_p are the diffusivities of the susceptibles and the phages. The last term in the P equation reflects the fact that each time an infective dies by lysis, it releases b copies of the virus, and we must now compute an expression for the term in curly brackets. As a first step in doing so, we shall indicate how to compute the density $I(x,t)$ of infectives at (x,t). This will be achieved by using a standard age-structured model approach. Let $i(x,t,a)$ be the density of infectives at (x,t) of age a. We assume that i satisfies the von Foerster-type equation

$$\frac{\partial i}{\partial t} + \frac{\partial i}{\partial a} = D_i \frac{\partial^2 i}{\partial x^2} - \mu_i i, \tag{4.2}$$

where D_i is the diffusivity of the infectives. The age of an infective will be measured from its time of infection so that, by the law of mass action,

$$i(x,t,0) = KS(x,t)P(x,t). \tag{4.3}$$

We want to solve (4.2) subject to (4.3) to obtain $i(x,t,a)$. The total density of infectives at (x,t) will then be obtained by totaling all those of "age" less than T (since older ones will have died by lysis); thus

$$I(x,t) = \int_0^T i(x,t,a)\, da. \tag{4.4}$$

Expression (4.4) can then be used to find the rate of death of infectives by lysis which is required for model (4.1).

Let

$$i^r(x,a) = i(x,a+r,a).$$

Then

$$\frac{\partial i^r}{\partial a} = \left[\frac{\partial i}{\partial t} + \frac{\partial i}{\partial a}\right]_{t=a+r} = \left[D_i \frac{\partial^2 i}{\partial x^2} - \mu_i i\right]_{t=a+r}$$

so that

$$\frac{\partial i^r}{\partial a} = D_i \frac{\partial^2 i^r}{\partial x^2} - \mu_i i^r. \tag{4.5}$$

Applying the Fourier transform

$$\widehat{i^r}(s,a) = \mathcal{F}\{i^r(x,a);\, x \to s\} = \int_{-\infty}^{\infty} i^r(x,a) e^{-isx}\, dx$$

to (4.5) gives

$$\frac{\partial \widehat{i^r}(s,a)}{\partial a} = -(D_i s^2 + \mu_i)\widehat{i^r}(s,a),$$

the solution of which is

$$\begin{aligned}
\widehat{i^r}(s,a) &= \widehat{i^r}(s,0) e^{-(D_i s^2 + \mu_i)\, a} \\
&= \mathcal{F}\left\{KS(x,r)P(x,r);\, x \to s\right\} e^{-(D_i s^2 + \mu_i)\, a} \\
&= \mathcal{F}\left\{KS(x,r)P(x,r);\, x \to s\right\} \mathcal{F}\left\{\frac{e^{-\mu_i a}}{2\sqrt{\pi D_i a}} e^{-x^2/4D_i a};\, x \to s\right\}
\end{aligned}$$

since

$$\widehat{i^r}(s,0) = \mathcal{F}\left\{i(x,r,0);\, x \to s\right\} = \mathcal{F}\left\{KS(x,r)P(x,r);\, x \to s\right\}$$

and

$$e^{-(D_i s^2 + \mu_i)\, a} = \mathcal{F}\left\{\frac{e^{-\mu_i a}}{2\sqrt{\pi D_i a}} e^{-x^2/4D_i a};\, x \to s\right\}.$$

By the convolution theorem for Fourier transforms,

$$i(x, a+r, a) = i^r(x,a) = \int_{-\infty}^{\infty} \frac{e^{-\mu_i a}}{2\sqrt{\pi D_i a}} e^{-(x-y)^2/4D_i a} KS(y,r)P(y,r)\, dy.$$

Hence

$$i(x,t,a) = \int_{-\infty}^{\infty} \frac{e^{-\mu_i a}}{2\sqrt{\pi D_i a}} e^{-(x-y)^2/4D_i a} KS(y, t-a)P(y, t-a)\, dy,$$

and so

$$I(x,t) = \int_0^T \int_{-\infty}^{\infty} \frac{e^{-\mu_i a}}{2\sqrt{\pi D_i a}} e^{-(x-y)^2/4D_i a} KS(y, t-a)P(y, t-a)\, dy\, da$$

or, after the substitution $a = t - \tau$,

$$I(x,t) = \int_{t-T}^{t} \int_{-\infty}^{\infty} \frac{e^{-\mu_i (t-\tau)}}{2\sqrt{\pi D_i (t-\tau)}} e^{-(x-y)^2/4D_i (t-\tau)} KS(y,\tau)P(y,\tau)\, dy\, d\tau.$$

From this, we see that $I(x,t)$ obeys

$$\begin{aligned}
\frac{\partial I(x,t)}{\partial t} &= D_i \frac{\partial^2 I(x,t)}{\partial x^2} - \mu_i I(x,t) + KS(x,t)P(x,t) \\
&\quad - Ke^{-\mu_i T} \int_{-\infty}^{\infty} \frac{e^{-(x-y)^2/4D_i T}}{2\sqrt{\pi D_i T}} S(y, t-T)P(y, t-T)\, dy,
\end{aligned}$$

and it is clear that the last term of this is the rate of death of infectives by lysis. Thus, system (4.1) becomes

$$
\begin{aligned}
\frac{\partial S(x,t)}{\partial t} &= D_s \frac{\partial^2 S(x,t)}{\partial x^2} + \alpha S(x,t)\left(1 - \frac{S(x,t)}{\gamma}\right) - KS(x,t)P(x,t), \\
\frac{\partial P(x,t)}{\partial t} &= D_p \frac{\partial^2 P(x,t)}{\partial x^2} - \mu_p P(x,t) - mP^2(x,t) - KS(x,t)P(x,t) \\
&\quad + bKe^{-\mu_i T}\int_{-\infty}^{\infty} \frac{e^{-(x-y)^2/4D_i T}}{2\sqrt{\pi D_i T}} S(y,t-T)P(y,t-T)\,dy.
\end{aligned} \tag{4.6}
$$

The formulation of a simple reaction-diffusion extension of (2.1) is complete. Like (2.1), system (4.6) does not involve the infectives $I(x,t)$ directly, but it does involve the parameter D_i which measures their diffusivity.

System (4.6) is to be solved on the domain $-\infty < x < \infty$. Reaction-diffusion systems with delay are quite difficult to study, and in this paper we will not attempt a systematic study of all the dynamics of (4.6). It is of interest to investigate what (4.6) tells us about the spatial spread of a virus infection in a population of bacteria. Mathematically, it is therefore reasonable to look for traveling wave solutions of (4.6) connecting the disease-free equilibrium $(\gamma, 0)$ with the endemic equilibrium (S^*, P^*) given by (3.1), assuming (3.2) holds so that an endemic equilibrium exists. A traveling front solution connecting these equilibria can model an invasion by the virus into the domain.

A traveling wave solution is one that travels at a constant speed c without changing shape. Mathematically, it is a solution that depends on x and t through the single variable $z = x + ct$, with $c \geq 0$ without loss of generality (this gives a leftward moving wave). In terms of the variable z, system (4.6) becomes

$$
\begin{aligned}
c\,S'(z) &= D_s S''(z) + \alpha S(z)\left(1 - \frac{S(z)}{\gamma}\right) - KS(z)P(z), \\
c\,P'(z) &= D_p P''(z) - \mu_p P(z) - mP^2(z) - KS(z)P(z) \\
&\quad + bKe^{-\mu_i T}\int_{-\infty}^{\infty} \frac{e^{-y^2/4D_i T}}{2\sqrt{\pi D_i T}} S(z - cT - y)P(z - cT - y)\,dy,
\end{aligned} \tag{4.7}
$$

where prime denotes differentiation with respect to z, and we need to solve (4.7) for $S(z)$ and $P(z)$ subject to

$$
(S,P)(-\infty) = (\gamma, 0) \quad \text{and} \quad (S,P)(+\infty) = (S^*, P^*). \tag{4.8}
$$

System (4.7), (4.8) remains a difficult mathematical problem, and we have not been able to establish the existence of a solution, even with the most recently developed methods for proving existence of traveling front solutions of delay reaction-diffusion systems such as those of Wu and Zou [21]. We shall therefore assume that such a solution exists and concentrate on finding out as much as possible about the speed c at which the virus infection spreads through the spatial domain. On the further assumption that the infection spreads at the minimum speed consistent with having an ecologically realistic solution satisfying $S(z), P(z) \geq 0$ for all $z \in (-\infty, \infty)$, we shall formally calculate this minimum speed by examining the situation as $z \to -\infty$, where $P(z) \to 0$, and obtaining conditions on c which are necessary for the convergence of $P(z)$ to 0 to be nonoscillatory. Linearizing as $z \to -\infty$, when $P \to 0$ and $S \to \gamma$, the

second equation of (4.7) becomes, approximately,

$$c\,P'(z) = D_p P''(z) - \mu_p P(z) - \gamma K P(z) + b\gamma K e^{-\mu_i T} \int_{-\infty}^{\infty} \frac{e^{-y^2/4D_i T}}{2\sqrt{\pi D_i T}} P(z - cT - y)\,dy$$

and has solutions of the form $P(z) = \exp(\lambda z)$ whenever λ satisfies

$$c\lambda - D_p\lambda^2 + \mu_p + \gamma K = b\gamma K e^{-\mu_i T} e^{-\lambda c T} e^{\lambda^2 D_i T}. \tag{4.9}$$

Since this analysis is for $z \to -\infty$, it is necessary that (4.9) have at least one real positive root if $P(z)$ is to approach 0 in a nonoscillatory manner. Whether (4.9) has real positive roots or not depends on the value of c, as can be easily seen by plotting the left- and right-hand sides of (4.9) against λ and remembering that $b\gamma K e^{-\mu_i T} > \mu_p + \gamma K$, since this is the condition for the existence of (S^*, P^*). If c is very small, then (4.9) has no real positive roots, but if c is gradually increased, there is a critical value of c which we shall call $c_{\min}$ (depending on T) such that when $c = c_{\min}$ (4.9) has one positive root (a double root), and when $c > c_{\min}$ the equation has precisely two real distinct positive roots. Only traveling fronts for which $c \geq c_{\min}$ are ecologically realistic, and we assume that the virus infection travels with speed $c_{\min}$ since it is usually the case in reaction-diffusion equations that the front one actually sees is the one with minimum speed (those with $c > c_{\min}$ usually have very small basins of attraction that rule out all but special initial conditions having very specific exponential decay rates).

Our aim now is to find out more about $c_{\min}$ and its dependence on the parameters. It is not possible to find an explicit expression for $c_{\min}$, but we can find some information about it. Indeed, $c_{\min}$ is the value of c for which (4.9) has a double root λ_*. Therefore, $c_{\min}$ and the double root λ_* must satisfy the simultaneous equations

$$\begin{aligned} c_{\min}\lambda_* - D_p\lambda_*^2 + \mu_p + \gamma K &= b\gamma K \exp(-\mu_i T - \lambda_* c_{\min} T + \lambda_*^2 D_i T), \\ c_{\min} - 2D_p\lambda_* &= b\gamma K(2\lambda_* D_i T - c_{\min} T)\exp(-\mu_i T - \lambda_* c_{\min} T + \lambda_*^2 D_i T). \end{aligned} \tag{4.10}$$

From these equations, we see that λ_* must satisfy $f(\lambda) = 0$, where

$$\begin{aligned} f(\lambda) := 2D_i D_p T\lambda^3 - (2c_{\min} D_i T + c_{\min} D_p T)\lambda^2 - (2D_i T(\mu_p + \gamma K) - c_{\min}^2 T + 2D_p)\lambda \\ + c_{\min} + c_{\min} T(\mu_p + \gamma K). \end{aligned}$$

Now f is a cubic and is such that $f(0) > 0$ and

$$f\left(\frac{c_{\min} + \sqrt{c_{\min}^2 + 4D_p(\mu_p + \gamma K)}}{2D_p}\right) = -\sqrt{c_{\min}^2 + 4D_p(\mu_p + \gamma K)} < 0.$$

These facts imply that the equation $f(\lambda) = 0$ has one real negative root and two real distinct positive roots. The larger of the two positive roots cannot satisfy the first equation of (4.10). Therefore, λ_* is the smaller of the two real positive roots of $f(\lambda) = 0$. Furthermore,

$$0 < \lambda_* < \frac{c_{\min} + \sqrt{c_{\min}^2 + 4D_p(\mu_p + \gamma K)}}{2D_p}. \tag{4.11}$$

The roots of a cubic equation are difficult to write down in general terms because there are numerous cases depending on the signs of various quantities defined in terms of the coefficients in the equation. An appendix to the book by Murray [15] gives all the

details. Although the coefficients of our particular cubic equation are complicated, we know a priori that our cubic equation has only real roots, and this narrows down the possibilities considerably. In fact, if we let

$$a_* = -\frac{c_{\min}\,(2D_i + D_p)}{6D_iD_p},$$

$$\alpha_* = \frac{4c_{\min}^2 TD_i^2 - 2c_{\min}^2 TD_iD_p + c_{\min}^2 TD_p{}^2 + 12D_i^2D_pT\mu_p + 12D_i^2D_pT\gamma K + 12D_iD_p^2}{36D_i^2D_p^2T}$$

(it is easily shown that $\alpha_* > 0$),

$$N = 8c_{\min}^2 TD_i^2 + 36D_i^2D_pT\mu_p + 36D_i^2D_pT\gamma\,K + 2c_{\min}^2 TD_i\,D_p - 18D_iD_p^2 - c_{\min}^2 TD_p^2,$$

$$\beta_* = \frac{c_{\min}\,(D_p - D_i)\,N}{108D_i{}^3D_p{}^3T},$$

and

$$\phi = (1/3)\sin^{-1}\left(\frac{\beta_*}{2\alpha_*^{3/2}}\right), \qquad \phi \in [-\pi/6, \pi/6],$$

then the only root of $f(\lambda) = 0$ satisfying (4.11) can be shown to be

$$\lambda_* = 2\alpha_*^{1/2}\sin\phi - a_*. \tag{4.12}$$

Substituting λ_* into either equation of (4.10) then gives a single, but very complicated, equation determining the speed $c_{\min}$.

We define the function $g(c)$ to be the left-hand side minus the right-hand side of the second equation of (4.10), with λ_* given by (4.12) and $c_{\min}$ replaced by c. The resulting function is too complicated to write out explicitly but is easily handled in MAPLE. Of course, $c_{\min}$ solves $g(c_{\min}) = 0$ and can easily be found either by reading off the root from an accurate plot of $g(c)$ or by using MAPLE commands for finding roots numerically. Figure 3 shows a plot of $g(c)$ for typical parameter values (see caption). We investigated how $c_{\min}$ depends on the values of all the parameters, and our main observations were as follows:

- If μ_i, μ_p, or T is increased, the result is a decrease in $c_{\min}$.
- If K, γ, b, D_i, or D_p is increased, the result is an increase in $c_{\min}$.
- If the delay T is large, then the value of $c_{\min}$ is much more sensitive to D_i than to D_p. Presumably this is because virus particles with a host are transported at the diffusivity of the infectives. To illustrate this, let $T = 7$ and other parameters retain their Figure 3 values. Then $c_{\min} = 1.265$. Keeping $T = 7$, if D_i is then raised to 100, $c_{\min}$ rises to 5.623. But if instead $D_i = 5$ and D_p is raised to 100, then $c_{\min}$ rises only to 1.976.

Analytical estimates for $c_{\min}$ can be obtained from other arguments, involving consideration of the graphs of the left- and right-hand sides of (4.9) as functions of λ. When $c = c_{\min}$ these two graphs just touch, at the value λ_* just discussed. Consider first the case when $D_i < D_p$ (so the minimum of the right-hand side is to the right of the maximum of the left-hand side). In this situation the maximum of the left-hand

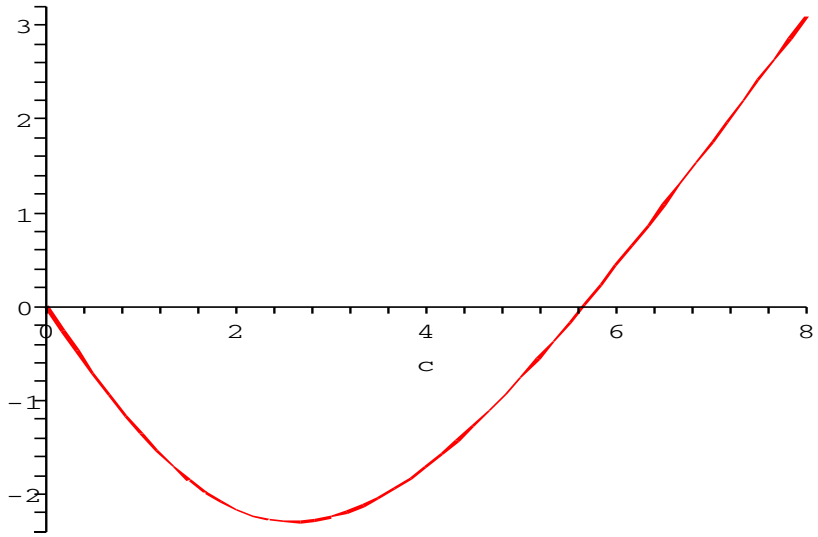


FIG. 3. *Plot of the function $g(c)$ defined in the text. The virus is predicted to spread at the speed $c_{\min} > 0$ such that $g(c_{\min}) = 0$. Parameter values used for this graph were $K = 0.134$, $\mu_i = 0.1$, $\mu_p = 2$, $T = 0.2$, $\gamma = 1$, $D_i = 5$, $D_p = 1$, and $b = 60$. For these values, $c_{\min} = 5.646$.*

side as a function of λ is $c_{\min}^2/(4D_p) + \mu_p + \gamma K$, and this must be less than the value of the right-hand side when $\lambda = 0$, which is $b\gamma K e^{-\mu_i T}$. This leads to the estimate

$$c_{\min} < 2\sqrt{D_p\{\gamma K(be^{-\mu_i T} - 1) - \mu_p\}} \quad \text{if } D_i < D_p.$$

If D_i is larger than D_p, but not too much larger, the above estimate on $c_{\min}$ will still hold.

We also carried out some numerical simulations of system (4.6) with a view to finding out whether the minimum speed $c_{\min}$ found from the linearized analysis is the speed which would be observed in practice. The question is whether the minimum speed wave is in some sense robust, attracting large classes of initial data. These questions are difficult to resolve analytically. In a recent paper, Thieme and Zhao [19] proved results on asymptotic speeds of spread for a class of nonlinear integral equations which include many reaction-diffusion models with delay, but their results do not include system (4.6). Figure 4 shows the results of a numerical simulation of system (4.6). For initial data, susceptibles S were set equal to γ throughout the domain, and some phages were introduced at $x = 0$ into an otherwise phage-free domain. Figure 4 shows how the phages spread out into the domain and the effect on the density of susceptible bacteria. Note that the traveling wave profiles are nonmonotone. Careful examination of the profiles suggests that the traveling fronts advance at the minimum speed $c_{\min}$ computed from the linearized analysis. The numerically computed front actually appears to travel at a slightly higher speed, but we are confident that this is purely a consequence of the discretization procedure. The speed varied slightly with the number of spatial grid points but seemed to approach $c_{\min}$ from above as

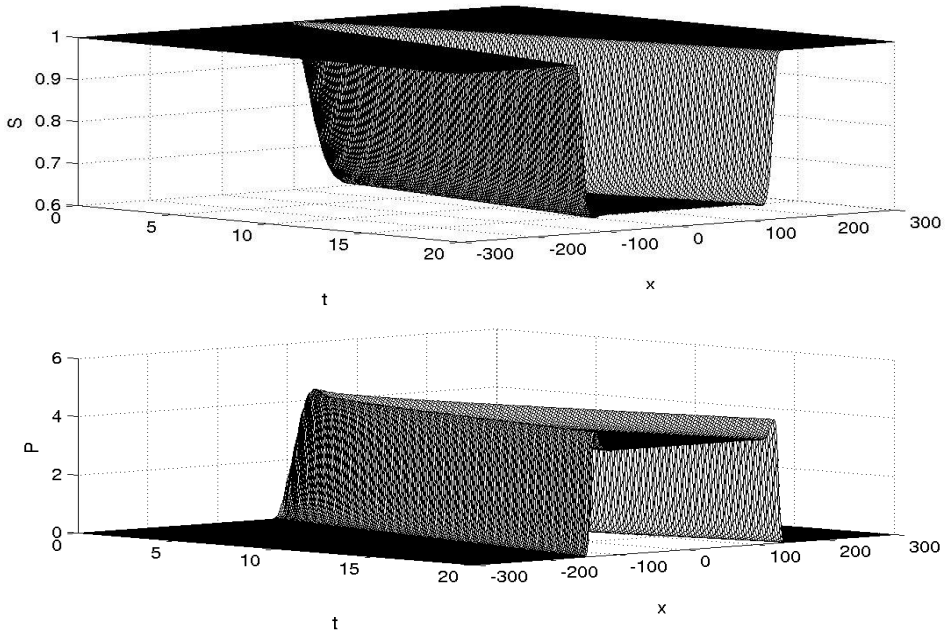


FIG. 4. *Numerical simulation of system* (4.6). *Parameter values were* $D_s = 5$, $m = 1$, $\alpha = 1.34$, *and the remaining parameters were as in Figure* 3. *The simulation suggests that the asymptotic speed of spread is* $c_{\min}$, *the minimum speed according to the linearized analysis.*

the number of grid points was increased. As a result, we suggest that the asymptotic speed of spread is indeed the speed $c_{\min}$ found from the linearized analysis.

5. Discussion. A key observation of Beretta and Kuang [2, 4] is the sensitivity of the dynamics on the phage reproduction rate b. This remains so for model (2.1). The novel observation of this work is the ultrasensitivity of the dynamics on the phage density dependent mortality rate m. This suggests that the density dependent mortality rate must be carefully measured to gain a better understanding of the bacteriophage infection dynamics in marine bacteria. Indeed, the recent work of Kuang, Fagan, and Loladze [13] contends that the predator death rate almost always positively correlates with the predator density in nature. To see this for model (2.1), we present Figures 5 and 6. Both figures use initial data and parameter values identical to those in Figure 2, except that in Figure 5, $m = 0$, while in Figure 6, $m = 0.2$.

The second novel aspect of our work is the rigorous derivation of a delay reaction-diffusion system to model the spatial spread of the virus infection and the use of this system to formally calculate the speed at which the infection spreads through a one-dimensional environment. The speed does not depend on the density dependent mortality parameter m just discussed. Unfortunately, it is not possible to find a simple expression for the speed, but it can be found from numerical computation. As we would expect, the speed depends on the diffusivity of both the infectives and the phages but is much more sensitive to the value of the former than the latter. This would be because virus replication takes place only inside a host, and therefore during replication the diffusivity of the viruses is effectively the host diffusivity D_i rather than D_p.

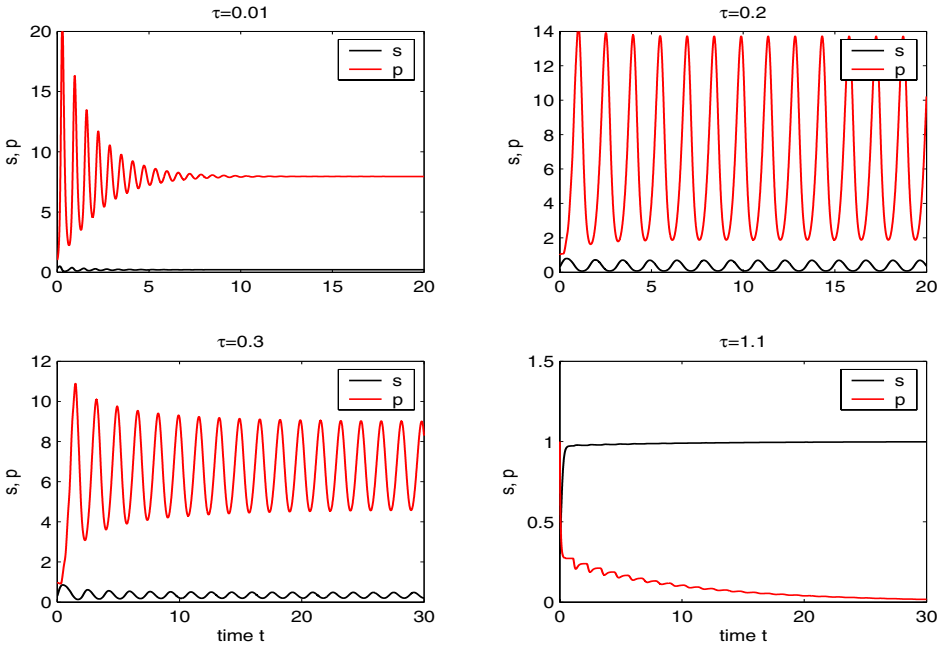


FIG. 5. *A solution of model* (3.16) *with* $s(\theta) = 0.3$, $p(\theta) = 1$, $\theta \in [-\tau, 0]$, *where* $\mu_p = 14.925$, $b = 75$, $\mu_i = 1.5$, $\alpha = 10$, $m = 0$, *and* τ *varies from* 0.01 *to* 1.1.

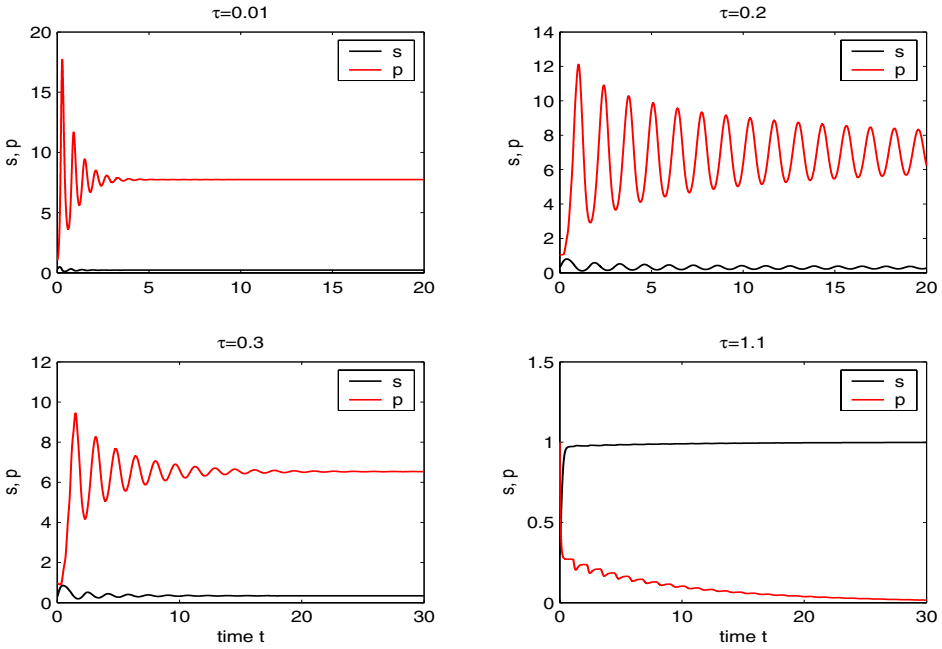


FIG. 6. *A solution of model* (3.16) *with* $s(\theta) = 0.3$, $p(\theta) = 1$, $\theta \in [-\tau, 0]$, *where* $\mu_p = 14.925$, $b = 75$, $\mu_i = 1.5$, $\alpha = 10$, $m = 0.2$, *and* τ *varies from* 0.01 *to* 1.1.

Acknowledgment. We would like to thank Zdzislaw Jackiewicz of Arizona State University for help with the numerical simulation of the reaction-diffusion model.

REFERENCES

[1] E. Beltrami and T. O. Carroll, *Modeling the role of viral disease in recurrent phytoplankton blooms*, J. Math. Biol., 32 (1994), pp. 857–863.
[2] E. Beretta and Y. Kuang, *Modeling and analysis of a marine bacteriophage infection*, Math. Biosci., 149 (1998), pp. 57–76.
[3] E. Beretta, M. Carletti, and F. Solimano, *On the effects of environmental fluctuations in a simple model of bacteria-bacteriophage infection*, Canad. Appl. Math. Quart., 8 (2000), pp. 321–366.
[4] E. Beretta and Y. Kuang, *Modeling and analysis of a marine bacteriophage infection with latency period*, Nonlinear Anal. Real World Appl., 2 (2001), pp. 35–74.
[5] E. Beretta and Y. Kuang, *Geometric stability switch criteria in delay differential systems with delay dependent parameters*, SIAM. J. Math. Anal., 33 (2002), pp. 1144–1165.
[6] O. Bergh, K.Y. Borsheim, G. Bratbak, and M. Heldal, *High abundance of viruses found in aquatic environments*, Nature, 340 (1989), pp. 467–468.
[7] B. J. M. Bohannan and R. E. Lenski, *Linking genetic change to community evolution: Insights from studies of bacteria and bacteriophage*, Ecology Letters, 3 (2000), pp. 362–377.
[8] A. Campbell, *Conditions for the existence of bacteriophage*, Evolution, 15 (1961), pp. 153–165.
[9] M. Carletti, *Numerical determination of the instability region for a delay model of phage-bacteria interaction*, Numer. Algorithms, 28 (2001), pp. 27–44.
[10] M. Carletti, *On the stability properties of a stochastic model for phage-bacteria interaction in open marine environment*, Math. Biosci., 175 (2002), pp. 117–131.
[11] S. A. Gourley, and J. W.-H. So, *Extinction and wavefront propagation in a reaction-diffusion model of a structured population with distributed maturation delay*, Proc. Roy. Soc. Edinburgh Sect. A, 133 (2003), pp. 527–548.
[12] Y. Kuang, *Delay Differential Equations, with Applications in Population Dynamics*, Academic Press, Boston, 1993.
[13] Y. Kuang, W. F. Fagan, and I. Loladze, *Biodiversity, habitat area, resource growth rate and interference competition*, Bull. Math. Biol., 65 (2003), pp. 497–518.
[14] R. E. Lenski and B. R. Levin, *Constraints on the coevolution of bacteria and virulent phage: A model, some experiments, and predictions for natural communities*, Amer. Naturalist, 125 (1985), pp. 585–602.
[15] J. D. Murray, *Mathematical Biology*, Springer-Verlag, Berlin, 1993.
[16] L. M. Proctor and J. A. Fuhrman, *Viral mortality of marine bacteria and cyanobacteria*, Nature, 343 (1990), pp. 60–62.
[17] H. L. Smith, *A structured population model and a related functional-differential equation: Global attractors and uniform persistence*, J. Dynam. Differential Equations, 6 (1994), pp. 71–99.
[18] J. W.-H. So, J. Wu, and X. Zou, *A reaction-diffusion model for a single species with age structure.* I. *Travelling wavefronts on unbounded domains*, Proc. Roy. Soc. London Ser. A, 457 (2001), pp. 1841–1853.
[19] H. R. Thieme and X.-Q. Zhao, *Asymptotic speeds of spread and traveling waves for integral equations and delayed reaction-diffusion models*, J. Differential Equations, 195 (2003), pp. 430–470.
[20] G. S. K. Wolkowicz, H. Xia, and J. Wu, *Global dynamics of a chemostat competition model with distributed delay*, J. Math. Biol., 38 (1999), pp. 285–316.
[21] J. Wu and X. Zou, *Travelling wave fronts of reaction-diffusion systems with delay*, J. Dynam. Differential Equations, 13 (2001), pp. 651–687.

SIAM J. APPL. MATH.
Vol. 65, No. 2, pp. 567–586

TWO-TIME-SCALE MARKOV CHAINS AND APPLICATIONS TO QUASI-BIRTH-DEATH QUEUES*

G. YIN† AND HANQIN ZHANG‡

Abstract. Aiming at reduction of complexity, this work is concerned with two-time-scale Markov chains and applications to quasi-birth-death queues. Asymptotic expansions of probability vectors are constructed and justified. Lumping all states of the Markov chain in each subspace into a single state, an aggregated process is shown to converge to a continuous-time Markov chain whose generator is an average with respect to the stationary measures. Then a suitably scaled sequence is shown to converge to a switching diffusion process. Extensions of the results are presented together with examples of quasi-birth-death queues.

Key words. Markov chain, singular perturbation, countable state space, asymptotic expansion, occupation measure, aggregation, switching diffusion, quasi-birth-death queue

AMS subject classifications. 34E05, 60J27, 60F05

DOI. 10.1137/S003613990139756X

1. Introduction.

Much effort has gone into evaluating performance measures of queueing systems in the past two decades; see, for example, [3, 22, 24, 26] and the references therein. Since exact solutions are difficult to obtain in many queueing problems, we are content with approximate solutions. To treat large-scale Markovian queueing systems, one of the most popular methods is decomposition, which consists of breaking the underlying network into smaller pieces (e.g., one station in each piece); see [3, 24, 26] among others. Although one often uses time-homogeneous Markovian models for approximating the actual systems, many queueing systems in real life are nonstationary (time-dependent); for example, the arrival and service rates in the systems are time-varying. Recently, time-inhomogeneous Markovian queueing networks have been widely used to model telecommunication systems; see [5]. Developing computational methods and approximation techniques for these quantities involved in *time-inhomogeneous* queueing problems has long been regarded as a challenging task (see [14] and [20]); see also [12, 18] and the references therein for earlier effort in this direction.

The motivation of our study stems from the recent advances in understanding asymptotic properties of singularly perturbed Markov chains aimed at reducing of complexity. We began our investigation in [13]. By combining matched asymptotic expansions and stochastic analysis, our effort was subsequently extended to treat more complex models and applications in control and optimization; see [29, 30]. These results are about Markov chains with finite-state spaces, and they are applicable to queues with a finite number of waiting rooms or finite capacity; see also [20] for a related work. Continuing our effort initiated in [28], using a model similar to that

*Received by the editors November 6, 2001; accepted for publication (in revised form) May 21, 2004; published electronically January 5, 2005.
http://www.siam.org/journals/siap/65-2/39756.html

†Department of Mathematics, Wayne State University, Detroit, MI 48202 (gyin@math.wayne.edu). The research of this author was supported in part by National Science Foundation grant DMS-0304928 and in part by the Wayne State University Research Enhancement Program.

‡Institute of Applied Mathematics, Academy of Mathematics and Systems Sciences, Academia Sinica, Beijing, 100080, China (hanqin@mail.amt.ac.cn). The research of this author was supported in part by a Distinguished Young Investigator grant from the National Natural Sciences Foundation of China and a grant from the Hundred Talents Program of the Chinese Academy of Sciences.

presented in [26] as our starting point, we consider queueing systems in which the generator of the queue length process includes both a fast varying part and a slowly changing part reflecting both strong and weak interactions of states belonging to different irreducible classes. Compared with the methods used in [3, 24, 26], we adopt a singular perturbation approach via time-scale separation to model the different transition intensities. In [28], treating countable-state-space Markov chains, one of the main assumptions used is that the underlying Markov chain is irreducible. In the context of queueing theory, it basically corresponds to the consideration of a single queue. The goal of this paper is to treat multistation queueing networks and to develop decomposition and aggregation methods for queueing network problems. Motivated by queueing networks involving quasi-birth-death processes, we start with decomposition by splitting the entire state space into a number of subspaces. Usually, the transitions within each subspace are much more intensive and frequent than those among different subspaces. We introduce a small parameter $\varepsilon > 0$ to highlight the different rates of transition intensities. Then we proceed to derive the asymptotic expansions of the probability distribution of the queue length. By lumping all the states in each subspace into a single state, we obtain a sequence of aggregated processes. We demonstrate that this sequence converges to a Markov chain. Furthermore, we obtain limit results for suitably scaled sequences.

The rest of the paper is arranged as follows. Section 2 begins with the precise formulation of the problem. Section 3 provides asymptotic expansions. A sequence of occupation measures is defined in section 4; its probabilistic properties, including mean square estimate, aggregation, and switching diffusion limit of a suitably scaled sequence, are examined. Section 5 presents extensions of results and queueing examples. Finally, an appendix is provided to include some technical complements.

2. Formulation. Working with a finite time horizon $t \in [0, T]$ for some $T > 0$, our focus is on time-inhomogeneous Markov chains. Suppose that $\beta(t)$ is a continuous-time Markov chain with countable state space $\mathbb{N} = \{1, 2, \ldots\}$. An infinite dimensional matrix-valued function $Q(t) = (q^{ij}(t))$ defined on $[0, T]$ is a generator of the Markov chain $\beta(t)$ if $q^{ij}(\cdot)$ is Borel measurable and bounded for each $i, j \in \mathbb{N}$, $q^{ij}(t) \geq 0$ for all $i \neq j$, $\sum_{j=1}^{\infty} q^{ij}(t) = 0$ for all $i \in \mathbb{N}$, and for any bounded and Borel-measurable function $\widetilde{g}(\cdot)$ defined on $\mathbb{N}$, $\widetilde{g}(\beta(t)) - \int_0^t Q(s)\widetilde{g}(\cdot)(\beta(s))ds$ is a martingale, where

$$Q(t)\widetilde{g}(\cdot)(i) = \sum_{j=1}^{\infty} q^{ij}(t)\widetilde{g}(j) \quad \text{for each} \quad i \in \mathbb{N}. \tag{2.1}$$

Throughout the paper, we use K to denote a generic positive constant. The conventions $K + K = K$ and $KK = K$ are used for simplicity. We use $\mathbb{1}_{m_0}$ and $\mathbb{1}$ to denote an m_0-dimensional and infinite dimensional column vector with all components being 1. For a vector z and a matrix H, we use z' and H' to denote their transposes, and we use z^i and h^{ij} to denote the ith component of z and the ijth entry of $H = (h^{ij})$, respectively. For a given matrix $H = (h^{ij})_{\infty \times \infty}$ with infinite columns and infinite rows, H_a is an augmented matrix given by $H_a = (H \vdots \mathbb{1})$. In addition, we use a subscript to index a sequence.

A Markov chain $\beta(t)$ or its generator $Q(t)$ is weakly irreducible, if the system of equations $g(t)Q_a(t) = (0 \vdots 1)$ has a unique solution $g(t) = (g^i(t))$ with $g^i(t) \geq 0$ for each $i \in \mathbb{N}$, where $0 = (0, 0, \ldots)$ is an infinite dimensional 0 vector. The unique nonnegative solution is termed a quasi-stationary distribution. (An equivalent way to write the system of equations in the above definition is $g(t)Q(t) = 0$, $g(t)\mathbb{1} = 1$.) The

definition is an extension of the usual notion of irreducibility, and the weak irreducibility given in [13] for finite-state Markov chains. Compared with the usual definition of irreducibility, it deals with time-varying generators and allows some components of the quasi-stationary distribution to be 0.

Suppose that the Markovian queueing network has n_0 ($n_0 < \infty$) interconnected stations. Consider a vector-valued queue-length process taking values in $\mathbb{N}^{n_0} = \mathbb{N} \times \mathbb{N} \times \cdots \times \mathbb{N}_{n_0}$ (an n_0-fold product). We use $\alpha(\cdot)$, a continuous-time Markov chain, to model the queue length of the queueing network. Suppose that $\mathbb{N}^{n_0}$ can be divided into l subsets. Within each subset, the transitions (such as arrivals and departures of customers, etc.) take place an order of magnitude more frequent than that of among different subsets. To highlight this contrast, we introduce a small parameter $\varepsilon > 0$. Thus the continuous-time Markov chain is ε-dependent, i.e., $\alpha(\cdot) = \alpha_\varepsilon(\cdot)$. For $t \in [0, T]$, assume that the generator of the Markov chain $\alpha_\varepsilon(t)$ is given by

$$Q_\varepsilon(t) = \frac{A(t)}{\varepsilon} + B(t), \quad \text{with } A(t) = \operatorname{diag}(A^1(t), \ldots, A^l(t)), \tag{2.2}$$

where $A(t)$, $B(t)$, and $A^i(t)$ (for $i = 1, \ldots, l$) are all generators of certain countable-state space Markov chains, and $\operatorname{diag}(A^1(t), \ldots, A^l(t))$ denotes a block diagonal matrix, each of the entries of which has appropriate dimension. Note that $B(t)$ is a generator and there is no need to assume it has the same diagonal matrix form as that of $A(t)$.

There is a certain hierarchy in the underlying network. Within each subset, one observes detailed variations of the networks such as arrivals and services of customers, etc., whereas at an upper system management level, instead of these variations, one observes the transitions among different subsets. Thus from an upper management point of view, by lumping all the states in each subspace into a single one, the system may be regarded as if it were a queue with finite waiting rooms (l rooms). Nevertheless, the aggregated process is generally non-Markovian. Fortunately, as will be shown, the aggregated process converges weakly to a limit process that is a finite-state Markovian queue. The significance of such a result is that in lieu of examining the detailed variations, one can study the aggregated process.

To reflect the decomposition in the network, write the state space of the queueing network as $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2 \cdots \cup \mathcal{M}_l$, where $\mathcal{M}_i = \{s_{i1}, s_{i2}, \ldots, \}$ for $i = 1, \ldots, l$. Following our approach for singularly perturbed Markov chains with finite-state spaces, we construct matched asymptotic expansions of the probability vector

$$p_\varepsilon(t) = (p_\varepsilon^{11}(t), \ldots, p_\varepsilon^{21}(t), \ldots, \ldots, p_\varepsilon^{l1}(t), \ldots), \quad p_\varepsilon^{ij}(t) = P(\alpha_\varepsilon(t) = s_{ij}). \tag{2.3}$$

For future use, partition an infinite dimensional vector v as $v = (v^1, \ldots, v^l)$ with $v^i = (v^{i1}, v^{i2}, \ldots)$. That is, v^i is an infinite dimensional vector corresponding to the subspace $\mathcal{M}_i$. Since we are dealing with countable state space Markov chains, we work with an infinite dimensional vector space. It is natural to consider the spaces $\ell_1 = \{(v^1, \ldots, v^l) : 1 \le i \le l,\ v^{ik} \in \mathbb{R} \text{ for each } k \in \mathbb{N}, \text{ and } \sum_{i=1}^{l} \sum_{k=1}^{\infty} |v^{ik}| < \infty\}$, $\ell_\infty = \{(v^1, \ldots, v^l)\colon 1 \le i \le l,\ v^{ik} \in \mathbb{R} \text{ for each } k \in \mathbb{N}, \text{ and } \sup_{1 \le i \le l} \sup_{1 \le k < \infty} |v^{ik}| < \infty\}$, equipped with $||v||_1 = \sum_{i=1}^{l} \sum_{k=1}^{\infty} |v^{ik}|$ and $||v||_\infty = \sup_{1 \le i \le l} \sup_{1 \le k < \infty} |v^{ik}|$, respectively; see [9, p. 11]. For a linear operator A defined on these spaces, we use its induced norm $||A|| = \sup_{||x||=1} ||Ax||$, where $||\cdot||$ is either norm $||\cdot||_1$ or $||\cdot||_\infty$. It is plain that $p_\varepsilon(t) \in \ell_1$ and that for each i ($1 \le i \le l$) and each k ($1 \le k < \infty$), $p_\varepsilon^{ik}(t) \ge 0$, and $\sum_{i=1}^{l} \sum_{k=1}^{\infty} p_\varepsilon^{ik}(t) = 1$. It is well known that $p_\varepsilon(t)$ satisfies the forward equation

$$\dot{p}_\varepsilon(t) = p_\varepsilon(t) Q_\varepsilon(t), \qquad p_\varepsilon(0) = p(0), \tag{2.4}$$

such that $p^{ik}(0) \geq 0$ and $\sum_{i=1}^{l} \sum_{k=1}^{\infty} p^{ik}(0) = 1$.

3. Asymptotic expansions. Following the approach of singular perturbation methods, we derive uniform (in the time variable t) asymptotic expansions of the probability vector as well as that of the transition probability matrices of the queue length processes. To proceed, we need the following conditions.

(A1) For each $t \in [0,T]$ and each $i = 1, \ldots, l$, $A^i(t)$ is weakly irreducible. There exists an integer n such that $(d^{n+2}/dt^{n+2})A^i(\cdot) \in \text{Lip}[0,T]$ and (d^{n+1}/dt^{n+1}) $B(\cdot) \in \text{Lip}[0,T]$, where $\text{Lip}[0,T]$ denotes the class of functions defined on $[0,T]$ that are Lipschitzian.

(A2) There is a $\kappa > 0$ such that for any real number $\tau > 0$,

$$\big|\big|\exp(A(0)\tau) - \text{diag}(\mathbb{1}\nu^1(0), \ldots, \mathbb{1}\nu^l(0))\big|\big|_\infty \leq K \exp(-\kappa\tau), \tag{3.1}$$

where $\nu^i(t) = (\nu^{i1}(t), \nu^{i2}(t), \ldots)$ is the quasi-stationary distribution corresponding to the generator $A^i(t)$.

Remark 3.1. From (A1) and the definition of weak irreducibility, $\nu^i(t)A^i_a(t) = (0 \vdots 1)$ has a unique solution. Furthermore, $\nu^i(t) = (0 \vdots 1)(A^i_a(t))'(A^i_a(t)(A^i_a(t))')^{-1}$. (A2) is a Doeblin-type condition. A condition in a slightly different form is given in [7, p. 192]. We will obtain the asymptotic expansions in two steps. The first step is a formal construction in which we (see [4, 10, 25] and [11, 13] among others) find the outer expansions and the initial layer corrections. The second step involves validating the formal expansions and deriving the desired error estimates.

3.1. Formal expansions. We seek matched asymptotic expansions of the form $\mathcal{O}_{\varepsilon,n}(t) + \mathcal{I}_{\varepsilon,n}(t) = \sum_{i=0}^{n} \varepsilon^i \phi_i(t) + \sum_{i=0}^{n} \varepsilon^i \psi_i(t/\varepsilon)$. For technical reasons, which will become clear in what follows, to justify the validity of the expansions, we also need to compute $\phi_{n+1}(t)$ and $\psi_{n+1}(t/\varepsilon)$. To make sure that the matching condition is satisfied, we use $\mathcal{O}_{\varepsilon,n+1}(0) + \mathcal{I}_{\varepsilon,n+1}(0) = p(0)$ or, more specifically, $\phi_0(0) + \psi_0(0) = p(0)$, $\phi_k(0) + \psi_k(0) = 0$ for $1 \leq k \leq n+1$. Our approach is based on constructions of the sequences $\{\phi_i(t)\}$ and $\{\psi_i(t/\varepsilon)\}$.

Substituting $\mathcal{O}_{\varepsilon,n+1}(t)$ into (2.4) and comparing coefficients of powers of ε^k, we obtain

$$\phi_0(t)A(t) = 0, \qquad \phi_k(t)A(t) = \dot{\phi}_{k-1}(t) - \phi_{k-1}(t)B(t), \quad 1 \leq k \leq n+1. \tag{3.2}$$

The outer expansions give us satisfactory approximation for $t > 0$ away from an initial layer of the order $O(\varepsilon)$, but it does not satisfy the initial condition and breaks down for sufficiently small t. To compensate, introduce a fast time variable $\tau = t/\varepsilon$. By the Lipschitz continuity given in (A1), taking Taylor expansions of $A(\varepsilon\tau)$ and $B(\varepsilon\tau)$ about 0 yields $A(\varepsilon\tau) = \sum_{i=0}^{n+1} \frac{(\varepsilon\tau)^i}{i!} \frac{d^iA(0)}{dt^i} + O((\varepsilon\tau)^{n+2})$, $\varepsilon B(\varepsilon\tau) = \sum_{i=0}^{n} \varepsilon \frac{(\varepsilon\tau)^i}{i!} \frac{d^iB(0)}{dt^i} + O((\varepsilon\tau)^{n+2})$. Substituting $\mathcal{I}_{\varepsilon,n+1}(t)$ into (2.4), using the above Taylor expansions of $A(t)$ and $B(t)$, and comparing powers of ε^i, we obtain the equations satisfied by the initial layer terms. We have $\psi_0(0) = p(0) - \phi_0(0)$, $\psi_k(0) = -\phi_k(0)$, and

$$\begin{aligned} &\frac{d\psi_0(\tau)}{d\tau} = \psi_0(\tau)A(0), \quad \frac{d\psi_k(\tau)}{d\tau} = \psi_k(\tau)A(0) + r_k(\tau), \quad 1 \leq k \leq n+1, \\ &r_k(\tau) = \sum_{i=0}^{k-1} \psi_{k-i-1}(\tau)\Big(\frac{\tau^{i+1}}{(i+1)!}\frac{d^{i+1}A(0)}{dt^{i+1}} + \frac{\tau^i}{i!}\frac{d^iB(0)}{dt^i}\Big), \ 1 \leq k \leq n+1. \end{aligned} \tag{3.3}$$

In [28], in which $A(t)$ consists of only one block, the outer expansions and initial layers can be obtained separately. Here, the constructions of $\{\phi_k(t)\}$ have to be done

in conjunction with the initial layer corrections. The equations in (3.3) together with the initial data are known as abstract Cauchy problems; see [17, p. 21]. It is well known that $\{\exp(A(0)\tau)\}$ is a uniformly continuous semigroup (see [17, 23]). With $\psi_k(0)$ to be determined, the representation of the solutions of (3.3) is given by

$$\begin{aligned}\psi_0(\tau) &= (p(0) - \phi_0(0))\exp(A(0)\tau),\\ \psi_k(\tau) &= \psi_k(0)\exp(A(0)\tau) + \int_0^\tau r_k(s)\exp(A(0)(\tau - s))ds,\ 1 \le k \le n+1.\end{aligned} \tag{3.4}$$

Step 1. Determine $\boldsymbol{\phi_0(t)}$ and $\boldsymbol{\psi_0(\tau)}$. We begin with the first equation in (3.2). Using the partitioned vector form introduced right after (2.3), we obtain $\phi_0^i(t)A^i(t) = 0$ for each $i = 1, \ldots, l$. These equations are not uniquely solvable since $A^i(t)$ have a 0 eigenvalue. However, by attaching $\sum_{j=1}^\infty \phi_0^{ij}(t) = \theta_0^i(t)$ to the equations, the resulting system of equations has a unique solution thanks to the weak irreducibility of $A^i(t)$. Thus, $\phi_0^i(t)$ must be proportional to $\nu^i(t)$, the quasi-stationary distribution corresponding to $A^i(t)$. That is, $\phi_0^i(t) = \theta_0^i(t)\nu^i(t)$, where $\theta_0^i(t) \in \mathbb{R}$ is to be determined. Define

$$\theta_0(t) = (\theta_0^1(t), \ldots, \theta_0^l(t)) \in \mathbb{R}^{1\times l},\ \widetilde{\mathbb{1}} = \operatorname{diag}(\mathbb{1}, \mathbb{1}, \ldots, \mathbb{1}),\ \nu(t) = \operatorname{diag}(\nu^1(t), \ldots, \nu^l(t)). \tag{3.5}$$

It is immediate that $A(t)$ is orthogonal to $\widetilde{\mathbb{1}}$ (i.e., $A(t)\widetilde{\mathbb{1}} = 0$) and $\phi_0(t) = \theta_0(t)\nu(t)$. For the equation with $k = 1$ in (3.2), multiplying from the right by $\widetilde{\mathbb{1}}$ leads to

$$\dot\theta_0(t) = \theta_0(t)\overline{B}(t), \quad \overline{B}(t) = \nu(t)B(t)\widetilde{\mathbb{1}}, \tag{3.6}$$

so $\overline{B}(t)$ is an average of $B(t)$ with respect to $\nu^1(t), \ldots, \nu^l(t)$. Note that $\overline{B}(t)$ is an $l \times l$ matrix-valued function. Note also that (3.6) is a linear system of differential equations and that it has a unique solution for each initial condition. Choose $\theta_0(0) = p(0)\widetilde{\mathbb{1}}$. Then the solution of (3.6) is uniquely determined. The $\phi_0^i(t)$ has the interpretation of total probability.

Concerning the initial layer correction $\psi_0(\tau)$, the solution is given by the first equation in (3.4) and is uniquely solved. We claim that $||\psi_0(\tau)||_\infty \le K\exp(-\kappa\tau)$ for some $\kappa > 0$ and $K > 0$. To prove this, note that $\widetilde{\mathbb{1}}\nu(0) = \operatorname{diag}(\mathbb{1}\nu^1(0), \ldots, \mathbb{1}\nu^l(0))$ and

$$\psi_0(0)\widetilde{\mathbb{1}}\nu(0) = [p(0) - \phi_0(0)]\widetilde{\mathbb{1}}\nu(0) = (0, 0, \ldots), \tag{3.7}$$

where $\phi_0(0) = \theta_0(0)\nu(0)$, $\phi_0(0)\widetilde{\mathbb{1}} = \theta_0(0)$, and $\theta_0(0) = p(0)\widetilde{\mathbb{1}}$ are used. By virtue of the orthogonality and (A2), $||\psi_0(\tau)||_\infty = ||\psi_0(0)(\exp(A(0)\tau) - \widetilde{\mathbb{1}}\nu(0))||_\infty \le K\exp(-\kappa\tau)$. To determine $\phi_i(t)$ and $\psi_i(t)$ for $i \ge 1$, we need the following lemma.

LEMMA 3.2. *Suppose that $Q(t)$ is a generator of a countable-state-space Markov chain such that $Q(t)$ is weakly irreducible for each $t \in [0, T]$ and $(d^{n+1}/dt^{n+1})Q(\cdot) \in Lip[0, T]$. Denote $\widetilde{Q}(t) = Q_a(t)Q_a'(t)$. Then for $k = 1, \ldots, n+1$, $(d^k/dt^k)\widetilde{Q}^{-1}(t)$ exists and belongs to C^{n+1-k} (the class of functions that are $(n+1-k)$-times continuously differentiable).*

Proof. Since $\widetilde{Q}(t)\widetilde{Q}^{-1}(t) = I$, differentiating both sides of the above equation leads to

$$\frac{d\widetilde{Q}^{-1}(t)}{dt} = -\widetilde{Q}^{-1}(t)\frac{d\widetilde{Q}(t)}{dt}\widetilde{Q}^{-1}(t), \tag{3.8}$$

so $\widetilde{Q}^{-1}(\cdot)$ is differentiable. Repeatedly differentiating equation (3.8) yields the desired result. ∎

Step 2. Determine $\phi_1(t)$ and $\psi_1(\tau)$. Lemma 3.2 and Remark 3.1 imply that $\phi_0(\cdot)$ is differentiable and is a function of class C^{n+1}. We proceed to determine $\phi_1(t)$ and $\psi_1(\tau)$. Note that the equation with $k=1$ in (3.2) is a nonhomogeneous equation whose right-hand side $\widetilde{\phi}_0(t) \stackrel{\text{def}}{=} \dot{\phi}_0(t) - \phi_0(t)B(t)$ is a known function since $\phi_0(t)$ has been found. Using (3.6), $[\dot{\phi}_0(t)-\phi_0(t)B(t)]\widetilde{\mathbb{1}} = \dot{\theta}_0(t)-\theta_0(t)\nu(t)B(t)\widetilde{\mathbb{1}} = (0,0,\ldots)$, and the Fredholm alternative yields that the equation with $k=1$ in (3.2) has a particular solution $\phi_{1,p}(t)$ being orthogonal to $\widetilde{\mathbb{1}}$. Assume the solution of $\phi_1(t)A(t) = \widetilde{\phi}_0(t)$ to be of the form $\phi_1(t) = \theta_1(t)\nu(t)+\phi_{1,p}(t)$. Since $\phi_{1,p}(t)$ is orthogonal to $\widetilde{\mathbb{1}}$, postmultiplying the equation with $k=2$ in (3.2) by $\widetilde{\mathbb{1}}$ leads to

$$\dot{\theta}_1(t) = \theta_1(t)\overline{B}(t) + \phi_{1,p}(t)B(t)\widetilde{\mathbb{1}}. \tag{3.9}$$

Once the initial condition is specified, (3.9) is uniquely solved. The initial condition $\theta_1(0)$ has to come from the initial layer correction term. With the selection of $\psi_1(0) = -\phi_1(0)$, by (3.4) with $k=1$, the unique solution is given by

$$\begin{aligned}\psi_1(\tau) = {} & \psi_1(0)\exp(A(0)\tau) + \int_0^\tau \psi_0(s)\exp(A(0)s)B(0)\exp(A(0)(\tau-s))ds \\ & + \int_0^\tau s\psi_0(s)\exp(A(0)s)\frac{dA(0)}{dt}\exp(A(0)(\tau-s))ds.\end{aligned} \tag{3.10}$$

By the exponential decay of $\psi_0(\tau)$, $\left|\left|\int_0^\tau \psi_0(s)\exp(A(0)s)ds\right|\right|_\infty = \int_0^\infty ||\psi_0(s)||_\infty\, ds < \infty$. Denote $\overline{\psi}_0 \stackrel{\text{def}}{=} \int_0^\infty \psi_0(s)\exp(A(0)s)dsB(0)$. Then from (A2),

$$\begin{aligned}\lim_{\tau\to\infty}\psi_1(0)\exp(A(0)\tau) &= \lim_{\tau\to\infty}\big(\psi_1(0)[\exp(A(0)\tau) - \widetilde{\mathbb{1}}\nu(0)] + \psi_1(0)\widetilde{\mathbb{1}}\nu(0)\big) \\ &= \psi_1(0)\widetilde{\mathbb{1}}\nu(0), \\ \lim_{\tau\to\infty}\int_0^\tau \psi_0(s)\exp(A(0)s)B(0)\exp(A(0)(\tau-s))ds &= \overline{\psi}_0\widetilde{\mathbb{1}}\nu(0).\end{aligned}$$

Note that

$$A(t)\widetilde{\mathbb{1}} = 0, \quad \frac{d^kA(t)}{dt^k}\widetilde{\mathbb{1}} = \frac{d^kA(t)\widetilde{\mathbb{1}}}{dt^k} = 0, \quad k = 1,\ldots,n+1. \tag{3.11}$$

Using the orthogonality (see (3.7) and (3.11)), we obtain

$$\left|\left|\int_0^\tau s\psi_0(s)\exp(A(0)s)\frac{dA(0)}{dt}\exp(A(0)(\tau-s))ds\right|\right|_\infty \le K\tau^2\exp(-\kappa\tau). \tag{3.12}$$

By demanding $\lim_{\tau\to\infty}\psi_1(\tau) = 0$, taking the limit as $\tau\to\infty$ in (3.10), and using the estimates above, we arrive at

$$\psi_1(0)\widetilde{\mathbb{1}}\nu(0) + \overline{\psi}_0\widetilde{\mathbb{1}}\nu(0) = 0. \tag{3.13}$$

An important observation indicates that there are only l unknowns in (3.13). Using the notation of partitioned vector given after (2.3), $\psi_1(0) = (\psi_1^1(0),\psi_1^2(0),\ldots,\psi_1^l(0))$ and $\overline{\psi}_0 = (\overline{\psi}_0^1,\overline{\psi}_0^2,\ldots,\overline{\psi}_0^l)$, the solution is given by $\psi_1^i(0)\mathbb{1} = -\overline{\psi}_0^i\mathbb{1}$, $i=1,\ldots,l$. Since $\psi_1(0)$ must be chosen so that $\phi_1(0)+\psi_1(0) = 0$, we obtain $\theta_1^i(0) = \overline{\psi}_0^i\mathbb{1}, i = 1,\ldots,l$. Thus both $\phi_1(t)$ and $\psi_1(\tau)$ have been determined. We claim that $\psi_1(\tau)$

decays exponentially fast. By adding and subtracting appropriate terms,

$$\begin{aligned}\psi_1(\tau) &= \psi_1(0)[\exp(A(0)\tau) - \widetilde{\mathbb{1}}\nu(0)] - \int_\tau^\infty \psi_0(s)\exp(A(0)s)dsB(0)\widetilde{\mathbb{1}}\nu(0)\\ &\quad + \psi_1(0)\widetilde{\mathbb{1}}\nu(0) + \int_0^\infty \psi_0(s)\exp(A(0)s)dsB(0)\widetilde{\mathbb{1}}\nu(0)\\ &\quad + \int_0^\tau \psi_0(s)\exp(A(0)s)B(0)[\exp(A(0)(\tau-s)) - \widetilde{\mathbb{1}}\nu(0)]ds\\ &\quad + \int_0^\tau s\psi_0(s)\exp(A(0)s)\frac{dA(0)}{dt}\exp(A(0)(\tau-s))ds.\end{aligned} \tag{3.14}$$

By (A2),

$$\begin{aligned}&\Big|\Big|\psi_1(0)[\exp(A(0)\tau) - \widetilde{\mathbb{1}}\nu(0)]\Big|\Big|_\infty \le K\exp(-\kappa\tau), \quad \text{and}\\ &\Big|\Big|\int_0^\tau \psi_0(s)\exp(A(0)s)B(0)[\exp(A(0)(\tau-s)) - \widetilde{\mathbb{1}}\nu(0)]ds\Big|\Big|_\infty\\ &\qquad\qquad \le K(1+\tau)\exp(-\kappa\tau).\end{aligned} \tag{3.15}$$

It then follows from the above estimates, $||\psi_1(\tau)||_\infty \le K(1+\tau+\tau^2)\exp(-\kappa\tau) \le K\exp(-\kappa_0\tau)$, for some $0 < \kappa_0 < \kappa$.

Step 3. Determine $\boldsymbol{\phi_k(t)}$ and $\boldsymbol{\psi_k(\tau)}$, for $\boldsymbol{1 < k \le n+1}$. We apply the same method for finding $\phi_1(t)$ and $\psi_1(\tau)$ to determine $\phi_k(t)$ and $\psi_k(\tau)$. For each $i = 1,\dots,l$, assume $\phi_k^i(t)$ is of the form $\phi_k^i(t) = \theta_k^i(t)\nu^i(t) + \phi_{k,p}^i(t)$, where $\phi_{k,p}(t) = (\phi_{k,p}^1(t),\dots,\phi_{k,p}^l(t))$ is a particular solution of the equation $\phi_k(t)A(t) = \dot\phi_{k-1}(t) - \phi_{k-1}(t)B(t)$, and $\phi_{k,p}(t)$ is orthogonal to $\widetilde{\mathbb{1}}$. We proceed to determine $\theta_k(t) = (\theta_k^i(t)) \in \mathbb{R}^{1\times l}$. By substitution, similar to (3.9), it is easily seen that $\theta_k(t)$ satisfies the differential equation

$$\dot\theta_k(t) = \theta_k(t)\overline{B}(t) + \phi_{k,p}(t)B(t)\widetilde{\mathbb{1}}. \tag{3.16}$$

To determine the initial condition $\theta_k(0)$, the definition of $r_k(\tau)$ in (3.4) gives us

$$\begin{aligned}\psi_k(\tau) &= \psi_k(0)\exp(A(0)\tau) + \int_0^\tau \sum_{j=0}^{k-1}\psi_{k-j-1}(s)\\ &\quad \times\Big(\frac{\tau^{j+1}}{(j+1)!}\frac{d^{j+1}A(0)}{dt^{j+1}} + \frac{\tau^j}{j!}\frac{d^jB(0)}{dt^j}\Big)\exp(A(0)(\tau-s))ds.\end{aligned} \tag{3.17}$$

Using the same techniques as that for $\theta_1(0)$, we obtain $\theta_k(0)$ and uniquely determine both $\phi_k(t)$ and $\psi_k(\tau)$. We then verify the exponential decay property of $\psi_k(\tau)$. The procedure is the same as for that of $\psi_1(\tau)$; we record the result as follows.

THEOREM 3.3. *Assume* (A1) *and* (A2). *Then the following assertions hold:*

(a) *The sequence $\{\phi_k(t)\}$ can be constructed by solving the system of equations*

$$\phi_k(t)A(t) = \dot\phi_{k-1}(t) - \phi_{k-1}(t)B(t) \stackrel{\text{def}}{=} \widetilde{\phi}_{k-1}(t) \tag{3.18}$$

for $k = 1,2,\dots,n$. The solution is $\phi_k(t) = \theta_k(t)\nu(t) + \phi_{k,p}(t)$, where $\phi_{k,p}(t)$ is a particular solution of (3.18)*, which is orthogonal to $\widetilde{\mathbb{1}}$, and $\theta_k(t)$ satisfies $\dot\theta_k(t) = \theta_k(t)\overline{B}(t) + \phi_{k,p}(t)B(t)\widetilde{\mathbb{1}}$.*

(b) *Find $\psi_k^i(0)\mathbb{1}$ from $\psi_k(0)\widetilde{\mathbb{1}}\nu(0) = -(\sum_{j=0}^{k-1}\int_0^\infty \frac{s^j}{j!}\psi_{k-j-1}(s)ds\frac{d^jA(0)}{dt^j})\widetilde{\mathbb{1}}\nu(0) \stackrel{\text{def}}{=} -\overline{\psi}_{k-1}\widetilde{\mathbb{1}}\nu(0)$. Choose $\theta_k^i(0) = -\psi_k^i(0)\mathbb{1} = \overline{\psi}_{k-1}^i\mathbb{1}$ for $i = 1,\dots,l$. Choose $\psi_k(0) = -\phi_k(0)$.*

(c) *For $k = 0, 1, \ldots, n+1$, $\theta_k(\cdot)$ and $\phi_k(\cdot)$ are $(n+1-k)$-times continuously differentiable on $[0, T]$; there exists a $\kappa_0 > 0$ such that $||\psi_k(\tau)||_\infty \le K \exp(-\kappa_0 \tau)$.*

3.2. Asymptotic justification. In this section, we obtain the desired error bounds and show that the asymptotic expansions hold uniformly in $t \in [0, T]$. Define an operator $\mathcal{L}_\varepsilon$ as

$$\mathcal{L}_\varepsilon f(t) = \varepsilon \frac{df(t)}{dt} - f(t)A(t) - \varepsilon f(t)B(t) \tag{3.19}$$

for a suitable smooth function $f(\cdot)$. First let us establish a lemma.

LEMMA 3.4. *Consider $\mathcal{L}_\varepsilon v_\varepsilon(t) = \Delta_{\varepsilon,k}(t)$, $v_\varepsilon(0) = 0$, with $\sup_{t\in[0,T]} ||\Delta_{\varepsilon,k}(t)||_\infty = O(\varepsilon^{k+1})$ for some k with $0 \le k \le n+1$. Then $\sup_{t\in[0,T]} ||v_\varepsilon(t)||_\infty = O(\varepsilon^k)$.*

Proof. Note that the initial value problem given above is a time-dependent abstract Cauchy problem or an evolution equation. The solution is given by $v_\varepsilon(t) = \frac{1}{\varepsilon}\int_0^t \Delta_{\varepsilon,k}(s) X_\varepsilon(t,s) ds$, where $X_\varepsilon(t,s)$ is a fundamental solution or an evolution operator. In fact (see Ladas and Lakshmikantham [17, p. 56]), $X_\varepsilon(t,s)$ is an operator-valued function with values in $L(\ell_1)$; the space of bounded linear operators, defined on ℓ_1 and strongly continuous in t, s for $0 \le s \le t \le T$ such that $(\partial/\partial t)X_\varepsilon(t,s)$ exists in strong topology of ℓ_1; $(\partial/\partial t)X_\varepsilon(t,s) \in L(\ell_1)$ for $0 \le s \le t \le T$ and $(\partial/\partial t)X_\varepsilon(t,s)$ is strongly continuous in t; the range of $X_\varepsilon(t,s)$ is in the domain of $Q_\varepsilon(t)$; and it satisfies the homogeneous problem $\varepsilon\frac{\partial X_\varepsilon(t,s)}{\partial t} - X_\varepsilon(t,s)A(t) - \varepsilon X_\varepsilon(t,s)B(t) = 0$, $X_\varepsilon(s,s) = I$, where the initial value I is the infinite dimensional identity matrix. Since it represents transition probabilities, $||X_\varepsilon(t,s)||_\infty$ is bounded uniformly in ε for all $t, s \in [0,T]$. Therefore, we have $\sup_{t\in[0,T]} ||v_\varepsilon(t)||_\infty \le \frac{K}{\varepsilon}\sup_{t\in[0,T]}\int_0^t ||\Delta_{\varepsilon,k}(s)||_\infty ds \le K\varepsilon^k$. The lemma is obtained. □

For each $k = 1, \ldots, n+1$, define a vector-valued error $e_{\varepsilon,k}(t)$

$$e_{\varepsilon,k}(t) = p_\varepsilon(t) - \sum_{i=0}^{k} \varepsilon^i \phi_i(t) - \sum_{i=0}^{k} \varepsilon^i \psi_i(t/\varepsilon), \tag{3.20}$$

where $p_\varepsilon(\cdot)$ is the solution of (2.4), $\phi_i(\cdot)$ and $\psi_i(\cdot)$ are the outer expansions and initial layer corrections, respectively. We must show that $e_{\varepsilon,n}(t) = O(\varepsilon^{n+1})$. For this purpose, we first derive a lemma whose proof is similar in spirit to the corresponding results for weakly irreducible generators [28] and is thus omitted.

LEMMA 3.5. *Under* (A1) *and* (A2), *for $k = 1, \ldots, n+1$, $\sup_{t\in[0,T]} ||\mathcal{L}_\varepsilon e_{\varepsilon,k}(t)||_\infty = O(\varepsilon^{k+1})$, and hence $\sup_{t\in[0,T]} ||e_{\varepsilon,k}(t)||_\infty = O(\varepsilon^k)$.*

Remark 3.6. Using Lemma 3.5, with $k = 1$, we obtain $\sup_{t\in[0,T]} ||e_{\varepsilon,1}(t)||_\infty = O(\varepsilon)$. However, $e_{\varepsilon,1}(t) = e_{\varepsilon,0}(t) - \varepsilon\phi_1(t) - \varepsilon\psi_1(t/\varepsilon)$. In view of the boundedness of $\phi_1(t)$ and $\psi_1(t/\varepsilon)$, $\varepsilon\phi_1(t) + \varepsilon\psi_1(t/\varepsilon) = O(\varepsilon)$. Thus, $e_{\varepsilon,0}(t) = O(\varepsilon)$ uniformly in $t \in [0,T]$. We can proceed inductively. By virtue of Lemma 3.5, with $k = n+1$, $\sup_{t\in[0,T]} ||\varepsilon_{\varepsilon,n+1}||_\infty = O(\varepsilon^{n+1})$. Going back one step and using $e_{\varepsilon,n+1}(t) = e_{\varepsilon,n}(t) + O(\varepsilon^{n+1})$, similar to the case of $k = 1$, we obtain $\sup_{t\in[0,T]} ||e_{\varepsilon,n}(t)||_\infty = O(\varepsilon^{n+1})$. We summarize the discussion thus far into the following theorem. This gives us the desired approximation results. Not only is the convergence of $P(\alpha_\varepsilon(t) = i)$ obtained, but also the rate of convergence is derived. Furthermore, a full asymptotic series is obtained.

THEOREM 3.7. *Under* (A1) *and* (A2), *we can construct two sequences $\{\phi_k(t)\}_{k=0}^n$ and $\{\psi_k(t/\varepsilon)\}_{k=0}^n$ by Theorem* 3.3 *such that $\phi_k(t) \in C^{n+1-k}$ and $\psi_k(t/\varepsilon)$ decay exponentially fast. Moreover with $e_{\varepsilon,n}(t)$ defined by* (3.20), $\sup_{t\in[0,T]} ||e_{\varepsilon,n}(t)||_\infty = O(\varepsilon^{n+1})$.

Using the same techniques, we can also obtain asymptotic expansions of transition probability matrices. Since the proofs are essentially the same, we will state only the results and omit the detailed argument. Let $P_\varepsilon(t_0,t)$ be the transition matrix $(p_\varepsilon^{in,jk}(t_0,t))$ with $p_\varepsilon^{in,jk}(t_0,t)=P(\alpha_\varepsilon(t)=s_{jk}|\alpha_\varepsilon(t_0)=s_{in})$ for all $s_{in},s_{jk}\in\mathcal{M}$.

THEOREM 3.8. *Assume* (A1) *and* (A2) *with* $n=1$. *Then*

$$P_\varepsilon(t_0,t)=\Phi_0(t_0,t)+\Psi_0(t_0,(t-t_0)/\varepsilon)+O(\varepsilon+\varepsilon\exp((t-t_0)/\varepsilon)) \tag{3.21}$$

uniformly in (t_0,t), *where* $0\le t_0\le t\le T$,

$$\Phi_0(t_0,t)=\widetilde{\mathbb{1}}\Theta(t_0,t)\nu(t),\qquad \frac{d\Psi_0(t_0,\tau)}{d\tau}=\Psi_0(t_0,\tau)A(t_0),\ \Psi_0(t_0,t_0)=I-\Phi_0(t_0,t_0), \tag{3.22}$$

where $\Theta(t_0,t)=(\theta^{ij}(t_0,t))\in\mathbb{R}^{l\times l}$ *is the solution of*

$$\frac{d}{dt}\Theta(t_0,t)=\Theta(t_0,t)\overline{B}(t),\quad \Theta(t_0,t_0)=I. \tag{3.23}$$

Remark 3.9. Owing to Theorems 3.7 and 3.8, for some $\kappa_0>0$, $p_\varepsilon(t)=\nu(t)+O(\varepsilon+\exp(-\kappa_0 t/\varepsilon))$, $P_\varepsilon(t_0,t)=\Phi_0(t_0,t)+O(\varepsilon+\exp(-\kappa_0(t-t_0)/\varepsilon))$.

4. Occupation measures, aggregation, and switching diffusion. This section presents further asymptotic results that are of probabilistic feature. The statements of the results are given, and the proofs are relegated to the appendix.

4.1. Occupations measures. For any $i=1,\dots,l$ and $j=1,2,\dots$, define sequences of occupation measures as

$$\mu_\varepsilon^{ij}(t)=\int_0^t z^{ij}(s,\alpha_\varepsilon(s))ds,\quad z^{ij}(t,\alpha_\varepsilon(t))=I_{\{\alpha_\varepsilon(t)=s_{ij}\}}-\nu^{ij}(t)I_{\{\alpha_\varepsilon(t)\in\mathcal{M}_i\}}, \tag{4.1}$$

where $\nu^{ij}(t)$ is the jth component of the quasi-stationary distribution $\nu^i(t)$ as defined in (A2). As a preparation, we first derive an order of magnitude estimate for $\mu_\varepsilon^{ij}(\cdot)$ defined in (4.1).

THEOREM 4.1. *Under the conditions of Theorem* 3.8, $\sup_{t\in[0,T]}E(\mu_\varepsilon^{ij}(t))^2=O(\varepsilon)$.

As alluded to in the introduction, we wish to reduce the complexity of the queueing network by aggregating states in each subspace as a single state. This leads to the definition of $\overline{\alpha}_\varepsilon(t)=i$ if $\alpha_\varepsilon(t)\in\mathcal{M}_i$.

THEOREM 4.2. *Under the conditions of Theorem* 3.8, $\overline{\alpha}_\varepsilon(\cdot)$ *converges weakly to* $\overline{\alpha}(\cdot)$, *a Markov chain generated by* $\overline{B}(t)$ *given in* (3.6).

4.2. Switching diffusion limit. Let $f(\cdot)$ be a real-valued function defined on $\mathcal{M}$ satisfying that $\{f(s_{ij}):1\le i\le l,\ 1\le j<\infty\}\in\ell_1$. For $t\in[0,T]$, define a sequence of real-valued functions by

$$x_\varepsilon(t)=\sum_{i=1}^{l}\sum_{j=1}^{\infty}\frac{1}{\sqrt{\varepsilon}}\int_0^t f(s_{ij})[I_{\{\alpha_\varepsilon(u)=s_{ij}\}}-\nu^{ij}(u)I_{\{\alpha_\varepsilon(u)\in\mathcal{M}_i\}}]du. \tag{4.2}$$

Our interest lies in the asymptotic properties of $x_\varepsilon(\cdot)$. To obtain the desired limit property, we consider a pair of processes $\{Y_\varepsilon(\cdot)\}=\{x_\varepsilon(\cdot),\overline{\alpha}_\varepsilon(\cdot)\}$ and aim to show that $Y_\varepsilon(\cdot)$ converges weakly to $Y(\cdot)$ with a suitable generator.

LEMMA 4.3. *Under the conditions of Theorem* 3.8, $\{Y_\varepsilon(\cdot)\}$ *is tight in* $D^2[0,T]$ (*the space of* $\mathbb{R}^2$*-valued functions that are right continuous and have left limits endowed with the Skorohod topology*).

Since $\{Y_\varepsilon(\cdot)\}$ is tight, by Prohorov's theorem, we can extract a weakly convergent subsequence. Select such a subsequence and still use ε as its index for notational simplicity. Denote the limit by $Y(\cdot) = (x(\cdot), \overline{\alpha}(\cdot))$. We proceed to characterize this limit process. Let C_L^2 be the collection of functions having bounded derivatives up to the second order with the second derivative being Lipschitz continuous. For each $i = 1, \ldots, l$, $g(\cdot, i) \in C_L^2$, define an operator $\mathcal{D}(t)$ by

$$\mathcal{D}(t)g(x,i) = \frac{1}{2}\sigma^2(t,i)\frac{\partial^2}{\partial x^2}g(x,i) + \overline{B}(t)g(x,\cdot)(i), \tag{4.3}$$

where $\sigma^2(t,i) > 0$ is a smooth function.

THEOREM 4.4. *Under the conditions of Theorem* 3.8, $Y_\varepsilon(\cdot)$ *converges weakly to* $Y(\cdot)$, *which is a solution of the martingale problem with operator* $\mathcal{D}(\cdot)$ *defined by* (4.3).

Using the above theorem, we can also get a weak convergence of a reflected process. Define $r_\varepsilon(t) = x_\varepsilon(t) - \inf_{0\le u\le t}\{x_\varepsilon(u)\}$. The weak convergence of $x_\varepsilon(t)$ and the continuous mapping theorem (see [2, p. 30, Theorem 5.1]) yield that $r_\varepsilon(\cdot)$ converges weakly to $r(\cdot)$, a reflected switching diffusion process.

5. Extensions and examples. In this section we give the extensions of the results and present examples of queueing problems.

5.1. Infinitely many blocks. We consider the case

$$Q_\varepsilon(t) = \frac{A(t)}{\varepsilon} + B(t) \quad \text{with } A(t) = \operatorname{diag}(A^1(t), A^2(t), \ldots). \tag{5.1}$$

That is, the matrix $A(t)$ given by (2.2) is a diagonal block one with infinitely many blocks. Instead of condition (A2), we assume (A2′).

(A2′) There is a $\kappa_l > 0$ such that for $\big|\big|\exp(A^l(0)\tau) - \mathbb{1}\nu^l(0)\big|\big|_\infty \le K\exp(-\kappa_l\tau)$, any real number $\tau > 0$, where $\nu^l(t) = (\nu^{l1}(t), \nu^{l2}(t), \ldots)$ is the quasi-stationary distribution corresponding to the generator $A^l(t)$, and $\inf_{l\ge 1}\kappa_l > 0$.

THEOREM 5.1. *Under* (A1) *and* (A2′), Theorem 3.7, Theorem 4.2, *and* Theorem 4.4 *hold.*

The proof of this theorem follows the same line of argument as that of Theorems 3.7, 4.2, and 4.4 and is thus omitted. A special case of the theorem is that $Q_\varepsilon(t)$ has the form (5.1) with $A(t)$ having infinitely many blocks such that each $A^i(t)$ is a generator of a finite-state Markov chain.

Example 5.2. Consider a two-station queueing system, where each station has a single server and an exogenous Poisson arrival process. For the first station, having completed service there, customers either proceed to a queue in front of the second station with probability p_1 or depart the system with probability $(1 - p_1)$. For the second station, after completing service there, they depart the system with probability $(1 - p_2)$ and go to a queue in front of the first station with probability p_2. Service is rendered in the order of the arrivals at each station. We assume that the first station can hold at most a total of m_0 customers (including the customer in service) and any further arriving customers from outside will in fact be refused entry and hence will depart immediately without service at the first station, while the second station

has an unlimited waiting room. Furthermore, we assume that the system is a time-dependent Markovian queueing network. Specifically, the rate of station i's customer arriving from outside is $\lambda_i(t)$, and the service rate is $\mu_i(t)$ $(i = 1, 2)$.

Assume that $\lambda_i(t)$ and $\mu_i(t)$ are smooth, positive, real analytic functions of time t. We let $\mathcal{Q}(t) = (\mathcal{Q}_2(t), \mathcal{Q}_1(t))$ be the Markovian queue length process that $\mathcal{Q}_i(t)$ equals the number of customers at station i at time t. Then the state space of the queueing system can be represented by $\{(k, i) : 0 \le k < \infty \text{ and } 0 \le i \le m_0\}$. To determine the generator $Q(t) = (q^{(j,n),(k,i)}(t))$, let

$$
\begin{aligned}
&\pi_1(t) = (1 - p_1)\mu_1(t),\ \pi_2(t) = (1 - p_2)\mu_2(t),\\
&\pi(t) = \lambda_1(t) + \pi_1(t),\ \tilde{\pi}(t) = \lambda_2(t) + p_1\mu_1(t) + \mu_2(t),\\
&A^1(t) = \begin{pmatrix}
-\lambda_1(t) & \lambda_1(t) & & & \\
\pi_1(t) & -\pi(t) & \lambda_1(t) & & \\
& \ddots & \ddots & \ddots & \\
& & \pi_1(t) & -\pi(t) & \lambda_1(t)\\
& & & \pi_1(t) & -\pi_1(t)
\end{pmatrix},\\
&B^1(t) = \operatorname{diag}(-\lambda_2(t), -[\lambda_2(t) + p_1\mu_1(t)], \ldots, -[\lambda_2(t) + p_1\mu_1(t)]),\\
&B^2(t) = \operatorname{diag}(-(\lambda_2(t) + \mu_2(t)), -\tilde{\pi}(t), \ldots, -\tilde{\pi}(t)),\\
&C(t) = \begin{pmatrix}
\lambda_2(t) & & & \\
p_1\mu_1(t) & \lambda_2(t) & & \\
& \ddots & \ddots & \\
& & p_1\mu_1(t) & \lambda_2(t)
\end{pmatrix},\\
&D(t) = \begin{pmatrix}
\pi_2(t) & p_2\mu_2(t) & & \\
& \ddots & \ddots & \\
& & \pi_2(t) & p_2\mu_2(t)\\
& & & \mu_2(t)
\end{pmatrix},
\end{aligned}
$$

where each matrix is a $(m_0 + 1) \times (m_0 + 1)$ matrix. Then some tedious calculations yield

$$
Q(t) = \operatorname{diag}(A^1(t), A^1(t), \ldots) + \begin{pmatrix}
B^1(t) & C(t) & & \\
D(t) & B^2(t) & C(t) & \\
& \ddots & \ddots & \ddots
\end{pmatrix}. \tag{5.2}
$$

For an initial time point t_0 with $t_0 \in [0, T]$, let $P(t_0, t)$ with $t > t_0$ be the transition matrix $P(t_0, t) = (p^{(j,n),(k,i)}(t_0, t))$ with

$$
p^{(j,n),(k,i)}(t_0, t) = P(\mathcal{Q}(t) = (k, i) | \mathcal{Q}(t_0) = (j, n)) \text{ for all } 0 \le j, k < \infty,\ 0 \le n, i \le m_0.
$$

Then we have the following system of equations for a quasi-birth-death process:

$$
\frac{d}{dt}P(t_0, t) = P(t_0, t)Q(t). \tag{5.3}
$$

Assume that

$$
(1 - p_1) \gg p_1, \quad \lambda_1(t) \gg p_1\mu_1(t), \quad \lambda_1(t) \gg \lambda_2(t), \quad \lambda_1(t) \gg \mu_2(t). \tag{5.4}
$$

Introduce $\varepsilon = \frac{1}{\inf_{t_0 \le t \le T} \lambda_1(t)}$. Then $\nu^i(t) = (\nu^{i0}(t), \nu^{i1}(t), \ldots, \nu^{im_0}(t))$ with

$$
\nu^{1j}(t) = \left(\frac{\lambda_1(t)}{\pi_1(t)}\right)^j p_0(t), \quad p_0(t) = 1 \Big/ \left[1 + \sum_{j=1}^{m_0} \left(\frac{\lambda_1(t)}{\pi_1(t)}\right)^j\right].
$$

Let $\widehat{\pi}(t) = [\lambda_2(t) + p_1\mu_1(t)] - p_1\mu_1(t)p_0(t)$. It follows from some tedious calculations that

$$\overline{B}(t) = \begin{pmatrix} -\widehat{\pi}(t) & \widehat{\pi}(t) & & \\ \mu_2(t) & -\widehat{\pi}(t) - \mu_2(t) & \widehat{\pi}(t) & \\ & \ddots & \ddots & \ddots \end{pmatrix}. \tag{5.5}$$

By Theorem 5.1,

$$\begin{aligned} P_\varepsilon(t_0, t) &= \Phi_0(t_0, t) + [I - \Phi(t_0, t_0)] \exp\left(A(t_0)\frac{t - t_0}{\varepsilon}\right) \\ &\quad + O\left(\varepsilon + \exp\left(\frac{t - t_0}{\varepsilon}\right)\right) \end{aligned} \tag{5.6}$$

uniformly in (t_0, t), where for $0 \le t_0 \le t \le T$, $\Phi_0(t_0, t)$ is given by (3.22) with $\Theta(t_0, t)$ satisfying (3.23). Using [23, Theorem 5.2, p. 128], the solution of (3.23) can be explicitly written as $\Theta(t_0, t) = U(t_0, t)$, where $U(t_0, t)$ is a solution operator. Consequently $\Phi(t_0, t)$ given in (5.6) takes the form $\Phi(t_0, t) = \widetilde{\mathbb{1}}U(t_0, t)\nu$. In particular, when the generators are time independent, $U(t_0, t)$ becomes a semigroup [23, Chapter 1] and may be expressed as $U(t_0, t) = \exp((t - t_0)\overline{B})$.

Next we consider the queue length process at the first station, $\mathcal{Q}_1(t)$. In general, $\mathcal{Q}_1(t)$ is not a Markov process. But from Theorem 5.1, $\mathcal{Q}_1(t)$ *can be approximated very well* by a Markov process $\overline{\alpha}(t)$ with generator $\overline{B}(t)$ defined by (5.5). Furthermore, we consider a refined approximation of $\mathcal{Q}(t) = (\mathcal{Q}_2(t), \mathcal{Q}_1(t))$. Define $\varpi_\varepsilon(t) = \sum_{i=0}^{\infty}\sum_{j=0}^{m_0}\int_0^t \left[I_{\{\mathcal{Q}(u)=(i,j)\}} - \nu^{(i,j)}(u)I_{\{\mathcal{Q}_1(u)=j\}}\right] du$. It follows from Theorem 5.1 that $(1/\sqrt{\varepsilon})\varpi_\varepsilon(t)$ can be well approximated by the switching diffusion process $\int_0^t \sigma(u, \overline{\alpha}(u))dw(u)$, where $w(\cdot)$ is a standard one-dimensional Brownian motion, $\sigma^2(t, i) = \sum_{k=0}^{m_0}\sum_{l=0}^{m_0}\left[\nu^{0k}(t)\int_0^\infty \psi_0^{kl}(i, t, u)du + \nu^{0l}(t)\int_0^\infty \psi_0^{lk}(i, t, u)du\right]$ for $0 \le i < \infty$, and $\psi_0^{kl}(i, s, t)$ is the (k, l)th entry of $\Psi_0(i, t, u)$ given by

$$\Psi_0(i, t, u) = \left(I - \begin{pmatrix} \nu^1(t) \\ \vdots \\ \nu^1(t) \end{pmatrix}\right) \exp(A^1(t)u).$$

Remark 5.3. References [12] and [18] establish asymptotic expansions for the queue length distribution of the time inhomogeneous single serve queue, a time-dependent *pure* birth-death process. Here we consider a *quasi*-birth-death process. Furthermore, the queue length process at the first station (generally non-Markovian) can be approximated well by a Markov process with generator $\overline{B}(t)$. In [20], uniform acceleration expansions for time-varying generators were treated, and in [28] diffusion approximation was also considered. In these references, the Markov chains have one ergodic class, whereas in the current paper, multiergodic classes are considered. Compared with the diffusion approximations in [19] and [27], the switching diffusion approximation given here is related to the queue length process on the interval $[0, t]$ and provides the evolution of the scaled sequence. The usual diffusion approximation leads to asymptotic normality, whereas switching diffusion limit yields Gaussian mixture distribution.

The $[\lambda_1(t) + p_2\lambda_2(t)]/[(1 - p_1p_2)\mu_1(t)]$ and $[\lambda_2(t) + p_1\mu_1(t)]/[(1 - p_1p_2)\mu_2(t)]$ are called the traffic intensities at station 1 and station 2, respectively. If $[\lambda_1(t) + p_2\lambda_2(t)]/[(1-p_1p_2)\mu_1(t)]$ and $[\lambda_2(t)+p_1\mu_1(t)]/[(1-p_1p_2)\mu_2(t)]$ are less than but close

to one, the system is regarded as in heavy traffic, whereas if $[\lambda_1(t) + p_2\lambda_2(t)]/[(1 - p_1p_2)\mu_1(t)]$ and $[\lambda_2(t) + p_1\mu_1(t)]/[(1 - p_1p_2)\mu_2(t)]$ are much less than one, the system is considered to be in light traffic. Our approximations are valid for either the heavy traffic case or the light traffic case as long as (5.4) holds. Using the Laplace transform technique, from the balance equation of the system, one may proceed as in [15] to carry out heavy traffic analysis for the sojourn time of time-homogeneous Markovian tandem queues with two servers.

Comparing with (5.2), we can consider a general quasi-birth-death process with state space $\{(i, j), i \geq 0, 1 \leq j \leq m_0\}$ with a generator $G(t)$ given by

$$\begin{pmatrix} A^1(t) - \operatorname{diag}(C^1(t)\mathbb{1}) & C^1(t) & & \\ B^1(t) & A^2(t) - \operatorname{diag}(B^1(t)\mathbb{1} + C^2(t)\mathbb{1}) & C^2(t) & \\ & \ddots & \ddots & \end{pmatrix}, \tag{5.7}$$

where i and j are called level and phase, respectively; see [22]. If the transitions between levels are much less frequent than the transitions between the phases inside the same level, then $Q(t)$ given by (5.7) can be written as

$$\operatorname{diag}(A^1(t), A^2(t), \ldots) + \varepsilon \begin{pmatrix} -\operatorname{diag}(C^1(t)\mathbb{1}) & C^1(t) & & \\ B^1(t) & -\operatorname{diag}(B^1(t)\mathbb{1} + C^2(t)\mathbb{1}) & C^2(t) & \\ & \ddots & \ddots & \ddots \end{pmatrix}.$$

Hence following Example 5.2, we can also get the asymptotic probability distribution of the process.

5.2. Asymptotic properties under γ-norm. For $v = (v^1, \ldots, v^l)$ with $v^i = (v^{i1}, v^{i2}, \ldots)$, the γ-norm (which was named ν-norm in [21]; see also [1]) is defined as $||v||_\gamma = \max_{1\leq i\leq l} \sup_{1\leq k<\infty} |v^{ik}|/\gamma^{ik}$. The corresponding induced γ-norm for any operator is given by $||A||_\gamma = \max_{1\leq k\leq l} \sup_{1\leq i<\infty} [\sum_{j=1}^{\infty} |A^k_{ij}|\gamma^{kj}]/\gamma^{ki}$. Replace (A2) with (A2″).

(A2″) There is a $\kappa > 0$ such that $\big|\big|\exp(A^l(0)\tau) - \mathbb{1}\nu^l(0)\big|\big|_\gamma \leq K \exp(-\kappa_l\tau)$ for any real number $\tau > 0$, where $\nu^l(t) = (\nu^{l1}(t), \nu^{l2}(t), \ldots)$ is the quasi-stationary distribution corresponding to the generator $A^l(t)$, and $\inf_{l\geq 1} \kappa_l > 0$.

Similar to Theorem 3.7, using essentially the same techniques, we obtain the following result.

THEOREM 5.4. *Under* (A1) *and* (A2″), *we construct two sequences* $\{\phi_k(t)\}_{k=0}^n$ *and* $\{\psi_k(t/\varepsilon)\}_{k=0}^n$ *by Theorem* 3.3 *such that* $\phi_k(t) \in C^{n+1-k}$ *and* $\psi_k(t/\varepsilon)$ *decay exponentially fast. Moreover, with* $e_{\varepsilon,n}(t)$ *defined by* (3.20), $\sup_{t\in[0,T]} ||e_{\varepsilon,n}(t)||_\gamma = O(\varepsilon^{n+1})$.

Example 5.5. Consider the queueing system given by Example 5.2, but allow the first station to have unlimited waiting rooms. Then the state space of $(\mathcal{Q}_2(t), \mathcal{Q}_1(t))$

can be represented by $\{(k,i): 0 \le k < \infty \text{ and } 0 \le i < \infty\}$. Let

$$\widehat{A}^1(t) = \begin{pmatrix} -\lambda_1(t) & \lambda_1(t) & & \\ (1-p_1)\mu_1(t) & -\pi(t) & \lambda_1(t) & \\ & \ddots & \ddots & \ddots \end{pmatrix},$$
$$\widehat{B}^1(t) = \operatorname{diag}(-\lambda_2(t), -[\lambda_2(t)+p_1\mu_1(t)], \ldots),$$
$$\widehat{B}^2(t) = \operatorname{diag}(-(\lambda_2(t)+\mu_2(t)), -\tilde{\pi}(t), \ldots),$$
$$\widehat{C}(t) = \begin{pmatrix} \lambda_2(t) & & \\ p_1\mu_1(t) & \lambda_2(t) & \\ & \ddots & \ddots \end{pmatrix}, \ \widehat{D}(t) = \begin{pmatrix} \pi_2(t) & p_2\mu_2(t) & \\ & \pi_2(t) & p_2\mu_2(t) \\ & & \ddots & \ddots \end{pmatrix}.$$

Then, some tedious calculations yield

$$\widehat{Q}(t) = \operatorname{diag}(\widehat{A}^1(t), \widehat{A}^1(t), \ldots) + \begin{pmatrix} \widehat{B}^1(t) & \widehat{C}(t) & & \\ \widehat{D}(t) & \widehat{B}^2(t) & \widehat{C}(t) & \\ & \ddots & \ddots & \ddots \end{pmatrix}. \tag{5.8}$$

Letting $\gamma^{1i}(0) = ((1-p_1)\mu_1(0)/\lambda_1(0))^i$, then we have (A2″) holds; see [21]. Therefore, by Theorem 5.4, we obtain the asymptotic probability distribution of $(\mathcal{Q}_2(t), \mathcal{Q}_1(t))$ under γ-norm.

Appendix A.

Proof of Theorem 4.1. Suppose that $0 \le s \le t$. For each $i = 1, \ldots, l$ and $j \in \mathbb{N}$, define

$$\begin{aligned} \zeta_\varepsilon^{1,ij}(s,t) &= P(\alpha_\varepsilon(s) = s_{ij}, \alpha_\varepsilon(t) = s_{ij}) - \nu^{ij}(t)P(\alpha_\varepsilon(s) = s_{ij}, \alpha_\varepsilon(t) \in \mathcal{M}_i), \\ \zeta_\varepsilon^{2,ij}(s,t) &= \nu^{ij}(s)\nu^{ij}(t)P(\alpha_\varepsilon(s) \in \mathcal{M}_i, \alpha_\varepsilon(t) \in \mathcal{M}_i) \\ &\quad -\nu^{ij}(s)P(\alpha_\varepsilon(s) \in \mathcal{M}_i, \alpha_\varepsilon(t) = s_{ij}). \end{aligned} \tag{A.1}$$

Theorem 3.8 and Remark 3.9 yield $\zeta_\varepsilon^{1,ij}(s,t) = O(\varepsilon + \exp(-\kappa_0(t-s)/\varepsilon))$. Similarly, $\zeta_\varepsilon^{2,ij}(s,t) = O(\varepsilon + \exp(-\kappa_0(t-s)/\varepsilon))$. It follows from $\mu_\varepsilon^{ij}(0) = 0$ that $E(\mu_\varepsilon^{ij}(t))^2 = 2\int_0^t \int_0^s (\zeta_\varepsilon^{1,ij}(r,s) + \zeta_\varepsilon^{2,ij}(r,s))drds$. The desired order estimates then follows. □

Proof of Theorem 4.2. The theorem will be proved in two steps. In the first step, we establish the tightness of $\{\overline{\alpha}_\varepsilon(\cdot)\}$, and in the second step, we characterize the limit process.

(1) Tightness. Note that $\overline{\alpha}_\varepsilon(t) = \sum_{i=1}^l iI_{\{\overline{\alpha}_\varepsilon(t)=i\}} = \sum_{i=1}^l iI_{\{\alpha_\varepsilon(t)\in\mathcal{M}_i\}}$. Define $\chi_\varepsilon(t) = (\chi_\varepsilon^1(t), \ldots, \chi_\varepsilon^l(t))$, $\chi_\varepsilon^i(t) = (I_{\{\alpha_\varepsilon(t)=s_{ij}\}})$, $\overline{\chi}_\varepsilon(t) = (I_{\{\overline{\alpha}_\varepsilon(t)=1\}}, \ldots, I_{\{\overline{\alpha}_\varepsilon(t)=l\}})$. In view of the definition of $\overline{\alpha}_\varepsilon(t)$ and the Cramér–Wold theorem [2, p. 49], to prove the tightness of $\{\overline{\alpha}_\varepsilon(\cdot)\}$, it suffices to derive that of $\{\overline{\chi}_\varepsilon(\cdot)\}$. For any $\delta > 0$, any $t > 0$, and any $s > 0$ with $\delta > s > 0$ and $t + s \in [0,T]$, owing to the Markov property, $E(\chi_\varepsilon(t+s) - \chi_\varepsilon(t) - \int_t^{t+s} \chi_\varepsilon(u)Q_\varepsilon(u)du|\mathcal{F}_{t,\varepsilon}) = 0$, where $\mathcal{F}_{t,\varepsilon}$ denotes the σ-algebra generated by $\{\alpha_\varepsilon(u): u \le t\}$. Postmultiplying by $\widetilde{\mathbb{1}}$ and noting $A(t)\widetilde{\mathbb{1}} = 0$ lead to $E(\overline{\chi}_\varepsilon(t+s) - \overline{\chi}_\varepsilon(t) - \int_t^{t+s} \chi_\varepsilon(u)B(u)\widetilde{\mathbb{1}}du|\mathcal{F}_{t,\varepsilon}) = 0$. Thus by (A1), $E(\overline{\chi}_\varepsilon(t+s)|\mathcal{F}_{t,\varepsilon}) - \overline{\chi}_\varepsilon(t) = O(s)$. This implies that $E(|\overline{\chi}_\varepsilon(t+s) - \overline{\chi}_\varepsilon(t)|^2|\mathcal{F}_{t,\varepsilon}) = E(|\int_t^{t+s} \chi_\varepsilon(u)B(u)\widetilde{\mathbb{1}}du|^2|\mathcal{F}_{t,\varepsilon}) = O(s)$. Consequently, for $0 < \delta \le s$, $\lim_{\delta\to 0} \limsup_{\varepsilon\to 0} E|\overline{\chi}_\varepsilon(t+s) - \overline{\chi}_\varepsilon(t)|^2 = 0$. The desired tightness then follows from [16, p. 47].

(2) Characterization of the limit. Since $\{\overline{\alpha}_\varepsilon(\cdot)\}$ is tight, by Prohorov's theorem, we can extract a weakly convergent subsequence. Select such a sequence and still use ε as its index for notational simplicity; denote its limit by $\overline{\alpha}(\cdot)$. We proceed to

characterize the limit process. We need to show that for any bounded function $\overline{g}(\cdot)$ defined on $\{1,\dots,l\}$,

$$\overline{g}(\overline{\alpha}(t)) - \int_0^t \overline{B}(u)\overline{g}(\cdot)(\overline{\alpha}(u))du \text{ is a martingale.} \tag{A.2}$$

To establish (A.2), it suffices that for any positive integers m, and $k \le m$, any bounded functions $h_k(\cdot)$ defined on $\{1,\dots,l\}$, and any $0 < t_k \le t \le t+s$,

$$E\prod_{k=1}^m h_k(\overline{\alpha}(t_k))\left(\overline{g}(\overline{\alpha}(t+s)) - \overline{g}(\overline{\alpha}(t)) - \int_t^{t+s}\overline{B}(u)\overline{g}(\cdot)(\overline{\alpha}(u))du\right) = 0. \tag{A.3}$$

To proceed, define $g(\alpha_\varepsilon(s)) = \sum_{i=1}^l \overline{g}(i)I_{\{\alpha_\varepsilon(s)\in\mathcal{M}_i\}}$. Then, $g(\alpha_\varepsilon(s)) = \overline{g}(\overline{\alpha}_\varepsilon(s))$. Thus, $g(\alpha_\varepsilon(t)) - \int_0^t Q_\varepsilon(u)g(\cdot)(\alpha_\varepsilon(u))du$ is a martingale. Consequently,

$$\prod_{k=1}^m Eh_k(\overline{\alpha}_\varepsilon(t_k))[g(\alpha_\varepsilon(t+s)) - g(\alpha_\varepsilon(t)) - \int_t^{t+s} Q_\varepsilon(u)g(\cdot)(\alpha_\varepsilon(u))du)] = 0. \tag{A.4}$$

By virtue of the weak convergence and the Skorohod representation, as $\varepsilon \to 0$,

$$E\prod_{k=1}^m h_k(\overline{\alpha}_\varepsilon(t_k))[g(\alpha_\varepsilon(t+s) - g(\alpha_\varepsilon(t))] \to E\prod_{k=1}^m h_k(\overline{\alpha}(t_k))[\overline{g}(\overline{\alpha}(t+s)) - \overline{g}(\overline{\alpha}(t))]. \tag{A.5}$$

The definition of $g(\cdot)$ leads to $A(u)g(\cdot)(\alpha_\varepsilon(u)) = 0$ for all $u \in [0,T]$. As a result,

$$\begin{aligned} &E\prod_{k=1}^m h_k(\overline{\alpha}_\varepsilon(t_k))\left(\int_t^{t+s} Q_\varepsilon(u)g(\cdot)(\alpha_\varepsilon(u))du\right)\\ &= \sum_{i=1}^l\sum_{j=1}^\infty E\prod_{k=1}^m h_k(\overline{\alpha}_\varepsilon(t_k))[\int_t^{t+s}\nu^{ij}(u)I_{\{\overline{\alpha}_\varepsilon(u)=i\}}B(u)g(\cdot)(s_{ij})du]\\ &\quad + \sum_{i=1}^l\sum_{j=1}^\infty E\prod_{k=1}^m h_k(\overline{\alpha}_\varepsilon(t_k))[\int_t^{t+s}[I_{\{\alpha_\varepsilon(u)=s_{ij}\}} - \nu^{ij}(u)I_{\{\overline{\alpha}_\varepsilon(u)=i\}}]B(u)g(\cdot)(s_{ij})du]. \end{aligned} \tag{A.6}$$

Using Theorem 4.1, the last term in (A.6) goes to 0 as $\varepsilon \to 0$. By virtue of the weak convergence of $\overline{\alpha}_\varepsilon(\cdot)$ and the Skorohod representation, (A.6) yields that

$$\begin{aligned} &E\prod_{k=1}^m h_k(\overline{\alpha}_\varepsilon(t_k))\left(\int_t^{t+s} Q_\varepsilon(u)g(\cdot)(\alpha_\varepsilon(u))du\right)\\ &\to E\prod_{k=1}^m h_k(\overline{\alpha}(t_k))\left(\int_t^{t+s}\overline{B}(u)\overline{g}(\cdot)(\overline{\alpha}(u))du\right). \end{aligned} \tag{A.7}$$

Combining (A.4), (A.5), and (A.7), (A.3) is verified. This completes the proof. □

Proof of Lemma 4.3. Note that $Y_\varepsilon(\cdot)$ is a sequence of vector-valued random processes with two components. Again, using the Crámer–Wold theorem, since $\{\overline{\alpha}_\varepsilon(\cdot)\}$ is tight, to prove the tightness of $\{Y_\varepsilon(\cdot)\}$, it suffices to verify the tightness of $\{x_\varepsilon(\cdot)\}$.

We claim that for any $\delta > 0$, $0 \le s \le t$, and $t - s \le \delta$,

$$\sup_{0\le s\le t\le T} E\left(|x_\varepsilon(t) - x_\varepsilon(s)|^2 \big| \mathcal{F}_{s,\varepsilon}\right) \le K(t-s).$$

First, for $u \in [s,t]$, $E(z^{ij}(u,\alpha_\varepsilon(u))\big|\mathcal{F}_{s,\varepsilon}) = O(\varepsilon + \exp(-\kappa_0(t-s)/\varepsilon))$ uniformly in i and j, by (4.2), the Markov property, and Theorem 3.8. Thus

$$\begin{aligned} E(x_\varepsilon(t) - x_\varepsilon(s)\big|\mathcal{F}_{s,\varepsilon}) &= \sum_{i=1}^{l}\sum_{j=1}^{\infty} f(s_{ij})\frac{1}{\sqrt{\varepsilon}}\int_s^t E(z^{ij}(u,\alpha_\varepsilon(u))\big|\mathcal{F}_{s,\varepsilon})du \\ &= \frac{1}{\sqrt{\varepsilon}} O(\varepsilon + \exp(-\kappa_0(t-s)/\varepsilon)) = O(\sqrt{\varepsilon}). \end{aligned} \tag{A.8}$$

Denote

$$\begin{aligned} \Delta_\varepsilon(s,t) &= E((x_\varepsilon(t) - x_\varepsilon(s))^2\big|\mathcal{F}_{s,\varepsilon}) \\ &= \frac{1}{\varepsilon}E\left(\left(\sum_{i=1}^{l}\sum_{j=1}^{\infty} f(s_{ij})\int_s^t z_\varepsilon^{ij}(\alpha_\varepsilon(u))du\right)^2 \Big|\mathcal{F}_{s,\varepsilon}\right). \end{aligned}$$

It then follows $\frac{d}{dt}\Delta_\varepsilon(s,t) = \frac{2}{\varepsilon}\sum_{i=1}^{l}\sum_{j=1}^{\infty} f(s_{ij})\int_s^t E(\zeta_\varepsilon^{1,ij}(u,t) + \zeta_\varepsilon^{2,ij}(u,t)\big|\mathcal{F}_{s,\varepsilon})du$, where $\zeta_\varepsilon^{1,ij}(u,t)$ and $\zeta_\varepsilon^{2,ij}(u,t)$ are defined in (A.1). Thus $(d/dt)\Delta_\varepsilon(s,t) = O(1)$, $\Delta_\varepsilon(s,s) = 0$. An integration then leads to $\Delta_\varepsilon(s,t) = O(t-s)$. Therefore,

$$\lim_{\delta\to 0}\limsup_{\varepsilon\to 0}\sup_{0\le t-s\le\delta} E(E(x_\varepsilon(t) - x_\varepsilon(s))^2\big|\mathcal{F}_{s,\varepsilon}) = 0.$$

Moreover, for each $\delta > 0$, and each rational $t \ge 0$, using the Chebyshev's inequality $\inf_\varepsilon P(|x_\varepsilon(t)| \le K_{t,\delta}) \ge \inf_\varepsilon(1 - E|x_\varepsilon(t)|^2/(K_{t,\delta})^2) \ge 1 - Kt/(K_{t,\delta})^2$. Select $K_{t,\delta} > \sqrt{KT/\delta}$. Then $\inf_\varepsilon P(|x_\varepsilon(t)| \le K_{t,\delta}) \ge 1-\delta$. This inequality, (A.8), and the criterion [16] then yield the desired tightness. □

Proof of Theorem 4.4. To prove the theorem, it suffices to show that

$$g(x(t),\overline{\alpha}(t)) - \int_0^t \mathcal{D}(u)g(x(u),\overline{\alpha}(u))du \text{ is a martingale.} \tag{A.9}$$

To verify this martingale property, we prove that for any positive integer m, any $k \le m$, any bounded and continuous function $h_k(\cdot,\alpha)$ for each α, and any $0 < t_k \le t \le t+s$,

$$\begin{aligned} \prod_{k=1}^{m} Eh_k(x(t_k),\overline{\alpha}(t_k))\ \Big(&g(x(t+s),\overline{\alpha}(t+s)) - g(x(t),\overline{\alpha}(t)) \\ &- \int_t^{t+s}\mathcal{D}(u)g(x(u),\overline{\alpha}(u))du\Big) = 0. \end{aligned}$$

To accomplish this, we begin with the process indexed by ε.

(1) Using an argument similar to [29, pp. 199–200], it can be verified that the martingale problem associated with operator $\mathcal{D}(t)$ has a unique (in the sense of distribution) solution.

(2) In what follows, we often need to carry out estimates involving $x_\varepsilon(t)$ and $\alpha_\varepsilon(t)$ intertwined. To untangle them, as a preparation, let $\eta(t,x)$ be a real-valued function

that is Lipschitz continuous in $(t,x)\in\mathbb{R}^2$. Using an argument similar to that of [29, Lemma 7.4, pp. 189–192], we can show

$$\lim_{\varepsilon\to 0}\sup_{0\le t\le T} E\left|\int_0^t z^{ij}(u,\alpha_\varepsilon(u))\eta(u,x_\varepsilon(u))du\right|^2=0. \tag{A.10}$$

(3) Define an operator $\mathcal{D}_\varepsilon(t)$ by

(A.11)
$$\begin{aligned}\mathcal{D}_\varepsilon(t)\rho(t,x,\alpha)=\frac{\partial\rho(t,x,\alpha)}{\partial t}\ &+\frac{1}{\sqrt{\varepsilon}}\left(\sum_{i=1}^{l}\sum_{j=1}^{\infty}f(s_{ij})z^{ij}(t,\alpha)\right)\frac{\partial\rho(t,x,\alpha)}{\partial x}\\ &+Q_\varepsilon(t)\rho(t,x,\cdot)(\alpha)\ \text{ for }\ \alpha\in\mathcal{M},\ \rho(\cdot,\cdot,\alpha)\in C^{1,1}.\end{aligned}$$

Then $\rho(t,x_\varepsilon(t),\alpha_\varepsilon(t))-\int_0^t\mathcal{D}_\varepsilon(u)\rho(u,x_\varepsilon(u),\alpha_\varepsilon(u))du$ is a martingale; see [6, Chapter 2].

Define $\overline{g}(x,\alpha)=\sum_{i=1}^{l}g(x,i)I_{\{\alpha\in\mathcal{M}_i\}}$. It is easily seen that $A(u)\overline{g}(x,\cdot)(\alpha_\varepsilon(u))=0$. Thus,

(A.12)
$$\mathcal{D}_\varepsilon(u)\overline{g}(x_\varepsilon(u),\alpha_\varepsilon(u))=\frac{1}{\sqrt{\varepsilon}}\widetilde{b}(u,\alpha_\varepsilon(u))\frac{\partial}{\partial x}\overline{g}(x_\varepsilon(u),\alpha_\varepsilon(u))+B(u)\overline{g}(x_\varepsilon(u),\cdot)(\alpha_\varepsilon(u)),$$

where $\widetilde{b}(u,\alpha)=\sum_{i=1}^{l}\sum_{j=1}^{\infty}f(s_{ij})z^{ij}(u,\alpha)$. To obtain the desired limit, we use the techniques of perturbed test function methods [16], which require us to introduce a perturbation that is small in magnitude and that results in desired cancellations. To this end, define the perturbation by $\widehat{g}(t,x,\alpha)$ such that $\widehat{g}(\cdot,x,\alpha)$ is uniform Lipschitz in t; both $\widehat{g}(\cdot)$ and $(\partial/\partial x)\widehat{g}(\cdot)$ are bounded; $(\partial/\partial x)\widehat{g}(\cdot,\cdot,\alpha)$ is Lipschitz in (t,x) such that it is the solution of

$$A(t)\widehat{g}(t,x,\cdot)(\alpha)=-\widetilde{b}(t,\alpha)\frac{\partial\overline{g}(x,\alpha)}{\partial x}. \tag{A.13}$$

It can be shown that such a function exists similar to [29, Remark 7.16].

Next, define the perturbed test function $g_\varepsilon(\cdot)$ by

$$g_\varepsilon(t,x,\alpha)=\overline{g}(x,\alpha)+\sqrt{\varepsilon}\widehat{g}(t,x,\alpha). \tag{A.14}$$

Then using the definition of $\mathcal{D}_\varepsilon(t)$, $g_\varepsilon(t,x_\varepsilon(t),\alpha_\varepsilon(t))-\int_0^t\mathcal{D}_\varepsilon(u)g_\varepsilon(u,x_\varepsilon(u),\alpha_\varepsilon(u))du$ is a martingale. Consequently,

$$\begin{aligned}\prod_{k=1}^{m}Eh_k(x_\varepsilon(t_k),\overline{\alpha}_\varepsilon(t_k))\Bigg[&g_\varepsilon(t+s,x_\varepsilon(t+s),\alpha_\varepsilon(t+s))-g_\varepsilon(t,x_\varepsilon(t),\alpha_\varepsilon(t))\\ &-\int_t^{t+s}\mathcal{D}_\varepsilon(u)g_\varepsilon(u,x_\varepsilon(u),\alpha_\varepsilon(u))du\Bigg]=0.\end{aligned}$$

By the weak convergence of $Y_\varepsilon(\cdot)$ to $Y(\cdot)$, the Skorohod representation, and (A.14),

$$\begin{aligned}&\lim_{\varepsilon\to 0}\prod_{k=1}^{m}Eh_k(x_\varepsilon(t_k),\overline{\alpha}_\varepsilon(t_k))[g_\varepsilon(t+s,x_\varepsilon(t+s),\alpha_\varepsilon(t+s))-g_\varepsilon(t,x_\varepsilon(t),\alpha_\varepsilon(t))]\\ &\quad=\prod_{k=1}^{m}Eh_k(x(t_k),\overline{\alpha}(t_k))[g(x(t+s),\overline{\alpha}(t+s))-g(x(t),\overline{\alpha}(t))].\end{aligned}$$

Note that from (A.11),

$$\begin{aligned}\sqrt{\varepsilon}\mathcal{D}_\varepsilon(u)\widehat{g}(u,x_\varepsilon(u),\alpha_\varepsilon(u)) &= \sum_{i=1}^{l}\sum_{j=1}^{\infty} f(s_{ij})z^{ij}(u,\alpha_\varepsilon(u))\frac{\partial\widehat{g}(u,x_\varepsilon(u),\alpha_\varepsilon(u))}{\partial x}\\ &+\sqrt{\varepsilon}\frac{\partial\widehat{g}(u,x_\varepsilon(u),\alpha_\varepsilon(u))}{\partial u}\\ &+\sqrt{\varepsilon}B(u)\widehat{g}(u,x_\varepsilon(u),\cdot)(\alpha_\varepsilon(u))+\frac{1}{\sqrt{\varepsilon}}A(u)\widehat{g}(u,x_\varepsilon(u),\cdot)(\alpha_\varepsilon(u)).\end{aligned}\tag{A.15}$$

It follows from (A.12), (A.13), and (A.15), upon cancellation, that

(A.16)

$$\begin{aligned}\mathcal{D}_\varepsilon(u)g_\varepsilon(u,x_\varepsilon(u),\alpha_\varepsilon(u)) &= \sum_{i=1}^{l}\sum_{j=1}^{\infty} f(s_{ij})z^{ij}(u,\alpha_\varepsilon(u))\frac{\partial\widehat{g}(u,x_\varepsilon(u),\alpha_\varepsilon(u))}{\partial x}\\ &+B(u)\overline{g}(x_\varepsilon(u),\cdot)(\alpha_\varepsilon(u))\\ &+\sqrt{\varepsilon}\frac{\partial\widehat{g}(u,x_\varepsilon(u),\alpha_\varepsilon(u))}{\partial u}+\sqrt{\varepsilon}B(u)\widehat{g}(u,x_\varepsilon(u),\cdot)(\alpha_\varepsilon(u)).\end{aligned}$$

Thus again, by the weak convergence and the Skorohod representation,

$$\begin{aligned}&\lim_{\varepsilon\to0}\prod_{k=1}^{m}Eh_k(x_\varepsilon(t_k),\overline{\alpha}_\varepsilon(t_k))\left(\int_t^{t+s}\mathcal{D}_\varepsilon(u)g_\varepsilon(u,x_\varepsilon(u),\alpha_\varepsilon(u))du\right)\\ &=\lim_{\varepsilon\to0}\prod_{k=1}^{m}Eh_k(x_\varepsilon(t_k),\overline{\alpha}_\varepsilon(t_k))\left[\int_t^{t+s}\sum_{i=1}^{l}\sum_{j=1}^{\infty}f(s_{ij})z^{ij}(u,\alpha_\varepsilon(u))\frac{\partial\widehat{g}(u,x_\varepsilon(u),\alpha_\varepsilon(u))}{\partial x}du\right.\\ &\qquad\left.+\int_t^{t+s}B(u)\overline{g}(x_\varepsilon(u),\cdot)(\alpha_\varepsilon(u))du\right].\end{aligned}$$

By virtue of the weak convergence of $Y_\varepsilon(\cdot)$ to $Y(\cdot)$, the Skorohod representation, and the boundedness of $\widehat{g}(\cdot)$, $(\partial/\partial x)\widehat{g}(\cdot)$, we have $\int_0^t B(u)\overline{g}(x_\varepsilon(u),\cdot)(\alpha_\varepsilon(u))du\to\int_0^t\overline{B}(u)g(x(u),\cdot)(\overline{\alpha}(u))du$.

Define

$$b(t,x,\alpha)=\widetilde{b}(t,\alpha)\frac{\partial\widehat{g}(t,x,\alpha)}{\partial x}.\tag{A.17}$$

Then

$$\begin{aligned}\int_0^t b(u,x_\varepsilon(u),\alpha_\varepsilon(u))du &= \int_0^t\sum_{i=1}^{l}\sum_{j=1}^{\infty}b(u,x_\varepsilon(u),s_{ij})\nu^{ij}(u)I_{\{\overline{\alpha}_\varepsilon(u)=i\}}du\\ &+\int_0^t\sum_{i=1}^{l}\sum_{j=1}^{\infty}[I_{\{\alpha_\varepsilon(u)=s_{ij}\}}-\nu^{ij}(u)I_{\{\overline{\alpha}_\varepsilon(u)=i\}}]b(u,x_\varepsilon(u),s_{ij})du.\end{aligned}\tag{A.18}$$

By (A.10), the last term in (A.18) goes to 0 in mean squares, so $\int_0^t b(u,x_\varepsilon(u),\alpha_\varepsilon(u))du$ converges to $\int_0^t\overline{b}(u,x(u),\overline{\alpha}(u))du$, where $\overline{b}(u,x,i)=\sum_{j=1}^{\infty}\nu^{ij}(u)b(u,x,s_{ij})$. Combining the estimates obtained so far, we arrive at

$$\begin{aligned}&Eh_k(x_\varepsilon(t_k),\overline{\alpha}_\varepsilon(t_k))\Big(\int_t^{t+s}\mathcal{D}_\varepsilon(t)g_\varepsilon(u,x_\varepsilon(u),\alpha_\varepsilon(u))du\Big)\\ &\quad\to Eh_k(x(t_k),\overline{\alpha}(t_k))\Big(\int_t^{t+s}\overline{b}(u,x(u),\overline{\alpha}(u))du+\int_t^{t+s}\overline{B}(u)g(x(u),\cdot)(\overline{\alpha}(u))du\Big).\end{aligned}$$

(4) In view of (A.13), for $i = 1, \ldots, l$,

$$A^i(t) \begin{pmatrix} \widehat{g}(t,x,s_{i1}) \\ \widehat{g}(t,x,s_{i2}) \\ \vdots \end{pmatrix} = \begin{pmatrix} \widetilde{b}(t,s_{i1}) \frac{\partial \overline{g}(x,i)}{\partial x} \\ \widetilde{b}(t,s_{i2}) \frac{\partial \overline{g}(x,i)}{\partial x} \\ \vdots \end{pmatrix},$$

which has a unique solution. This implies that $\widehat{g}(t,x,s_{ij})$ is a function of $(\partial/\partial x)\overline{g}(x,i)$. In view of (A.17), $b(t,x,s_{ij})$ is a function of $(\partial/\partial x)\widehat{g}(t,x,s_{ij})$. Thus $b(t,x,s_{ij})$ is a function of $(\partial^2/\partial x^2)\overline{g}(x,i)$.

Denote $\overline{b}(t,x,i) = (1/2)a(t,i)(\partial^2/\partial x^2)\overline{g}(x,i)$, where $a(t,i)$ is an appropriate function. Using an argument similar to that [29, pp. 200–203], it can be shown that $a(t,i) \geq 0$. Thus, $a(t,i)$ can be written as $a(t,i) = \sigma^2(t,i)$, where for each $i = 1, 2, \ldots, l$,

$$\sigma^2(t,i) = \sum_{k=1}^{\infty} \sum_{m=1}^{\infty} f(s_{ik}) f(s_{im}) \left[\nu^{ik}(t) \int_0^{\infty} \psi_0^{km}(i,u,t)du + \nu^{im}(t) \int_0^{\infty} \psi_0^{mk}(i,u,t)du \right], \tag{A.19}$$

and $\psi_0^{km}(i,s,t)$ is the (k,m)th entry of $\Psi_0(i,s,t)$ given by

$$\Psi_0(i,s,t) = (I - (\nu^i(s), \nu^i(s), \ldots)') \exp(A^i(s)t).$$

This specifies the covariance structure of the limit process and establishes the desired martingale property and hence the theorem follows. □

Acknowledgment. We thank the editor and the reviewers for detailed comments and suggestions on an early version of the manuscript, which have led to much improvement of the paper.

REFERENCES

[1] E. Altman, K. E. Avrachenkov, and R. Nunez-Queija, *Pertrubation analysis for denumerable Markov chains with applications to queueing models*, Adv. Appl. Probab., 36 (2004), pp. 839–853.
[2] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.
[3] G. Bitran and D. Tirupati, *Multiproduct queueing networks with deterministic routing: Decomposition approach and the notion of interference*, Management Sci., 34 (1988), pp. 75–100.
[4] N. N. Bogoliubov and Y. A. Mitropolskii, *Asymptotic Methods in the Theory of Nonlinear Oscillator*, Gordon and Breach, New York, 1961.
[5] J. Daigle and J. Langford, *Models for analysis of packet voice communication systems*, IEEE J. Sel. Areas Commun., SAC-4 (1986), pp. 847-855.
[6] M. H. A. Davis, *Markov Models and Optimization*, Chapman & Hall, London, 1993.
[7] J. L. Doob, *Stochastic Processes*, Wiley Classic Library Edition, Wiley, New York, 1990.
[8] S. G. Eick, W. A. Massey, and W. Whitt, *The physics of the $M_t/G/\infty$ queue*, Oper. Res., 41 (1993), pp. 731–742.
[9] V. Hutson and J. S. Pym, *Applications of Functional Analysis and Operator Theory*, Academic Press, London, 1980.
[10] A. M. Il'in, A. S. Kalashnikov, and O. A. Oleinik, *Linear equations of the second order of parabolic type*, Russian Math. Surveys, 17 (1962), pp. 1–143.
[11] A. M. Il'in, R. Z. Khasminskii, and G. Yin, *Asymptotic expansions of solutions of integro-differential equations for transition densities of singularly perturbed switching diffusions: Rapid switchings*, J. Math. Anal. Appl., 238 (1999), pp. 516–539.

[12] J. Keller, *Time-dependent queues*, SIAM Rev., 24 (1982), pp. 401–412.
[13] R. Z. Khasminskii, G. Yin, and Q. Zhang, *Asymptotic expansions of singularly perturbed systems involving rapidly fluctuating Markov chains*, SIAM J. Appl. Math., 56 (1996), pp. 277–293.
[14] C. Knessl, B. J. Matkowsky, Z. Schuss, and C. Tier, *Asymptotic analysis of a state-dependent $M/G/1$ queueing system*, SIAM J. Appl. Math., 46 (1986), pp. 483–505.
[15] C. Knessl and J. A. Morrison, *Heavy traffic analysis of the sojourn time in tandem queues with overtaking*, SIAM J. Appl. Math., 51 (1991), pp. 1740–1763.
[16] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.
[17] G. E. Ladas and V. Lakshmikantham, *Differential Equations in Abstract Spaces*, Academic Press, New York, 1972.
[18] W. A. Massey, *Asymptotic analysis of the time dependent $M/M/1$ queue*, Math. Oper. Res., 10 (1985), pp. 305–327.
[19] W. A. Massey and W. Whitt, *Unstable asymptotics for nonstationary queues*, Math. Oper. Res., 19 (1994), pp. 267–291.
[20] W. A. Massey and W. Whitt, *Uniform acceleration expansions for Markov chains with time-varying rates*, Ann. Appl. Probab., 8 (1998), pp. 1130–1155.
[21] S. P. Meyn and R. L. Tweedie, *Computable bounds for geometric convergence rates of Markov chains*, Ann. Appl. Probab., 4 (1994), pp. 981–1021.
[22] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models*, Johns Hopkins University Press, Baltimore, 1981.
[23] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[24] M. I. Reiman, *Asymptotically exact decomposition approximations for open queueing networks*, Oper. Res. Lett., 9 (1990), pp. 363–370.
[25] A. B. Vasil'eava and V. F. Butuzov, *Asymptotic Methods in Singular Perturbations Theory*, Vysshaya Shkola, Moscow, 1990 (in Russian).
[26] W. Whitt, *The queueing network analyzer*, Bell Syst. Tech. J., 62 (1983), pp. 2779–2815.
[27] W. Whitt, *Stochastic-Process Limits*, Springer-Verlag, New York, 2001.
[28] G. Yin and H. Zhang, *Countable-state-space Markov chains with two-time scales and applications to queueing systems*, Adv. Appl. Probab., 34 (2002), pp. 662–688.
[29] G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications: A Singular Perturbation Approach*, Springer-Verlag, New York, 1998.
[30] G. Yin, Q. Zhang, and G. Badowski, *Asymptotic properties of a singularly perturbed Markov chain with inclusion of transient states*, Ann. Appl. Probab., 10 (2000), pp. 549–572.

SIAM J. APPL. MATH.
Vol. 65, No. 2, pp. 587–617

ACOUSTIC-ROTATIONAL INTERNAL FLOW CAUSED BY TRANSIENT SIDEWALL MASS ADDITION*

P. L. STAAB†, M. J. REMPE‡, AND D. R. KASSOY§

Abstract. Asymptotic and numerical methods are used to describe thermal transients in an internal flow caused by time-dependent, spatially distributed sidewall mass addition. Solutions are obtained for the temperature distribution and wall heat transfer, as well as the vorticity, in a high Reynolds number (Re), low Mach number (M), *compressible* flow in a cylinder. A multiple-scale analysis, valid in the limit of $M \to 0$ and $Re \to \infty$, is used to derive an equation for $O(M)$ acoustic disturbances arising from transient injection. The limit process is also used to obtain reduced equations for the rotational axial velocity field and the nonacoustic temperature variation arising from a balance of convection and transverse diffusion. The convection-diffusion equations are characterized by viscous and conductive transport on an $O(M)$ radial scale relative to the $O(1)$ nondimensional cylinder radius. These small-scale diffusive effects are pervasive throughout the entire cylinder in this large Re, injected flow. The thermal analysis presented here shows that the transient temperature disturbance consists of an $O(M)$ acoustic component and an $O(M)$ radially dependent "rotational" component arising from the convection and diffusion of large radial gradients. The latter are generated on the injection surface and then gradually fill the cylinder as time elapses. The radial gradient of temperature is $O(1)$, although the temperature disturbance is only $O(M)$. Results for the radial and axial variations of the instantaneous temperature distribution imply that the $O(M)$ acoustic and "rotational" temperature components make the largest contributions to the total energy transient. Smaller kinetic energy effects appear only at $O(M^2)$. These results emphasize the importance of modeling intense thermal transients, including the sidewall heat transfer, in addition to the more familiar vorticity distributions.

Key words. fluids mechanics, heat transfer, finite difference methods, perturbation methods

AMS subject classifications. 76N99, 76M45, 80A20, 65M06

DOI. 10.1137/S0036139903421560

1. Introduction.

Time-dependent mass injection from the boundaries of cylinders and channels induces a transient internal flow characterized by the presence of coexisting rotational and irrotational (acoustic) phenomena. The velocity and vorticity dynamics in these high Reynolds number and low Mach number flows have been elucidated through the use of both linear and nonlinear mathematical models.

Linear analysis was initiated by Flandro [1] and later reviewed in Flandro [2]. More recent, related research advances are noted by Majdalani and Rienstra [3]. These studies focus on the evolution of very small, linear disturbances to the basic steady flow field induced by uniform mass addition at the boundary. Flow transients are driven by an assumed small quasi-steady pressure disturbance, independent of conditions on the flow boundary. The asymptotic methodology used to derive the disturbance equations implies that the magnitude of the assumed pressure field is smaller than the injection Mach number, typically $O(10^{-3})$. It follows that the disturbance is about 0.1 of commensurately small vorticity at the injection surface, resulting primarily from

*Received by the editors January 20, 2003; accepted for publications (in revised form) January 26, 2004; published electronically January 5, 2005. This work was supported by the Air Force Office of Scientific Research.

http://www.siam.org/journals/siap/65-2/42156.html

†Department of Mathematics, Tufts University, Medford, MA 02155 (peter.staab@tufts.edu).

‡Department of Applied Mathematics, Northwestern University, Evanston, IL 60208 (michael.rempe@northwestern.edu).

§Department of Mechanical Engineering, University of Colorado, Boulder, CO 80309-0427 (david.kassoy@colorado.edu).

an inviscid interaction between the imposed pressure gradient transients and the fluid injected from the boundary, as well as subsequent penetration of vorticity into the flow field. These models supplement classical acoustic stability theory for solid rocket motors, reviewed by Culick and Yang [4], valid only for irrotational flow dynamics. Related, more general linear stability analyses, including the rotational flow effects, are given by Casalis, Avalon, and Pineau [5] and Venugopal, Najjar, and Moser [6, 7].

Nonlinear velocity transients in channel and cylinder configurations have also been investigated [8, 9, 10, 11, 12, 13, 14]. These asymptotic and numerical studies use an initial-boundary value problem (IBVP) approach to investigate the effect of imposed time-dependent disturbances, located on a boundary surface, on the internal flow dynamics. The asymptotic methodology is based on the limit $M \to 0$, where M is the characteristic axial Mach number. In this case, the $O(M)$ pressure gradient transients, typically 1% to 10% of the baseline value and 10 to 100 times larger than those in the linear analyses, arise directly from the time-dependent boundary condition, often associated with variable mass addition. The evolution of $O(1/M)$ transient vorticity, generated by an inviscid mechanism at the boundary, is described by a fully nonlinear convection-diffusion equation for a relatively large rotational axial velocity disturbance.

The vorticity generation concept was used originally by Cole and Aroesty [15] to describe high Reynolds number, external steady flow past a flat plate with massive injection (a magnitude larger than that permitted in traditional boundary layer theory). This concept has been generalized in [8, 9, 10, 11, 12, 13, 14] to describe high Reynolds number, low Mach number, compressible transient internal flows. In contrast to the Cole and Aroesty theory, where viscosity is confined to a thin, separated shear layer, viscous effects are pervasive throughout the transient internal flow, affecting the diffusion of vorticity in a fundamental way.

An early asymptotic formulation for flow in a cylinder is described by Zhao [8] and more fully by Zhao et al. [9]. Significant transient endwall mass addition is the source of large disturbances in the flow field. A related computational study is given by Kirkkopru, Kassoy, and Zhao [10], where the source of large disturbances is an assumed pressure transient on the exit plane. Staab and Kassoy [11] extend the study in [9] to describe multidimensional flow in a cylinder arising from time-dependent endwall mass addition with an amplitude varying in the transverse direction.

The impact of sidewall mass addition transients are considered by Staab et al. [12]. In this case the acoustic field in a cylinder is directly attributable to the time-dependent component of the axially distributed injection velocity. A related computational study by Kirkkopru et al. [13] predicts flow transients arising from a similar injection distribution on the walls of a cylinder. Staab and Kassoy [14] impose a sidewall transient injection velocity with axial and azimuthal dependence to induce a fully three-dimensional transient internal flow disturbance. Results include an axial component of the vorticity vector, in addition to the more familiar azimuthal component found in cylinders with symmetric injection.

The consequences of sidewall vorticity generation in transient internal flows have been observed in many related computational studies. Vuillot and Avalon [16] and Vuillot [17] provide early examples of vorticity penetration into channel flow. Lupoglazoff and Vuillot [18] consider the presence of vorticity shed from the injection surface of a channel. Many other examples are cited in [8, 9, 10, 11, 12, 13, 14]. More recently, Venugopal, Najjar, and Moser [6, 7] described the appearance of vorticity striations on a channel sidewall and their penetration into a turbulent internal flow.

These results support the predictions of vorticity spreading discussed in [8, 9, 10, 11, 12, 13, 14] and [19].

Solutions in [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14] and [16, 17, 18] provide results for the pressure and velocity transients in the flow field, with little consideration of thermal effects. The latter have probably been relegated to secondary importance because the injected fluid is assumed to be isothermal. However, Staab et al. [12] show that unexpectedly large transient radial temperature gradients are generated at the injection surface by a nonconductive interaction between the local acoustic pressure transients and the injected fluid. These gradients are the source of unanticipated heat transfer at the surface, just as the radial gradients of axial velocity (vorticity) are the source of unanticipated shear stress on the surface. The radial temperature gradients will be convected into the internal flow field and diffused by conduction, analogous to the flow physics affecting the vorticity distribution in the internal flow field. Hegab's [19] computational model for channel flow with transient isothermal sidewall mass addition provides solutions for the time-history of the spatially varying temperature distribution in the channel and for the injection surface heat transfer.

The current work provides an asymptotic description of the temperature distribution and wall heat transfer for high Reynolds number (Re), low Mach number (M), compressible flow in a cylinder. An initial-boundary value approach is used to develop the model. When $t \leq 0$, a steady flow is generated by a steady, axially dependent, isothermal mass injection from the sidewall. Then for $t > 0$, a similar magnitude, isothermal component of mass is added with oscillatory time-dependence and spatial variation in the axial direction.

A formal multiple-scale analysis in the limit of $M \to 0$ and $Re \to \infty$ is used to derive an equation for the acoustic response to the transient injection. The axial acoustic velocity is of the same order of magnitude as that of the initial steady flow. The limit process is also used to obtain reduced equations for the rotational axial velocity field and the associated "rotational" temperature variation above and beyond the acoustic response. The latter is described by two compatible equations. The first is a wave equation which describes a nonconductive interaction between the $O(M)$ local pressure transient and the injected fluid, leading to the generation of $O(1)$ radial gradients in the "rotational" temperature field. The second is a *linear* convection-diffusion equation with coefficients dependent on both the acoustic and rotational velocities. Velocities are obtained from a solution to an analogous *nonlinear* convection-diffusion equation. Diffusion in either case occurs on a small radial scale, $O(M)$ compared to the $O(1)$ nondimensional cylinder radius. Diffusive effects in this large Re injected flow are pervasive throughout the entire cylinder.

The present temperature and velocity analysis incorporates an integral scaling transformation for the crucial small radial variable, used by Zhao [8] and Zhao et al. [9], in place of the linear scaling transformation used in [12]. Solutions written in terms of the former variable are valid for much longer time-scales than those in the latter.

The analysis presented here shows that the transient temperature disturbance consists of an acoustic component and a radially dependent "rotational" component arising from the convection and diffusion of large radial gradients. The latter are generated on the injection surface and then gradually fill the cylinder as time elapses. Results are given for the radial and axial variations of the instantaneous temperature distribution. Spatial waves in the radial coordinate are observed between the sidewall and an identifiable front, which moves toward the cylinder axis with increasing time.

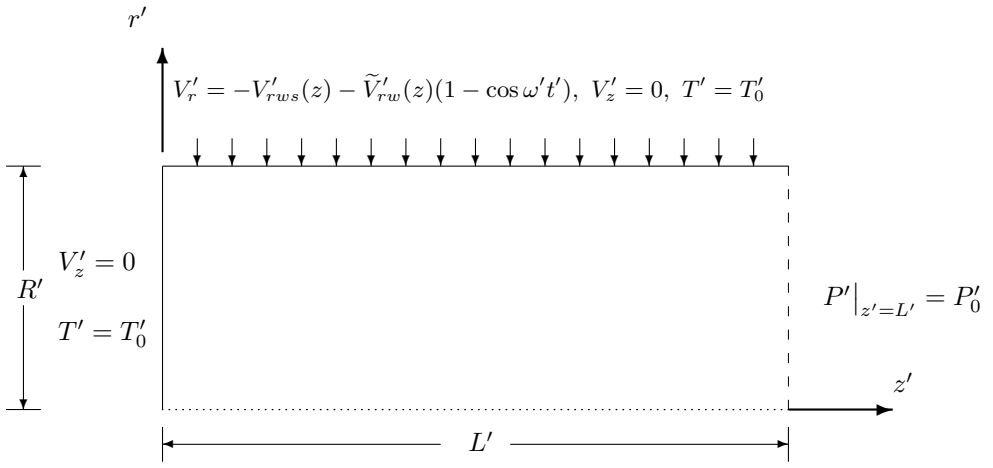


FIG. 1. *The geometric domain of the model is a cylinder of length L' and radius R', where the aspect ratio $\delta = \frac{L'}{R'} \gg 1$. Along the sidewall, $r' = R'$, a steady velocity, $V_{rws}'(z')$, and an unsteady velocity, $\widetilde{V}_{rw}(z')(1 - \cos\omega' t')$, are imposed, and the no-slip condition, $V_z' = 0$, is satisfied. A temperature T_0' is imposed along the sidewall and the endwall. A pressure node is imposed at the exit plane, $z' = L'$.*

Temperature solutions are developed for two different spatial distributions of the mass addition. A comparison shows that the details of the internal flow are sensitive to the characteristics of the mass addition.

A comparison is made of new, long-time solutions for the spatially distributed temperature and velocity. Differences in characteristic spatial waves are observed and attributed to basic properties of the linear (for temperature) and nonlinear (for velocity convection-diffusion) equations.

A study of energy partitioning shows that the $O(M)$ acoustic and "rotational" temperature components make the largest contributions to the total energy transient. Smaller kinetic energy effects appear only at (M^2). In addition, the long-time average of the largest acoustic temperature transient vanishes, while that associated with the "rotational" component is nonzero. This result emphasizes the importance of modeling the thermal transients, including the sidewall heat transfer in the cylinder.

2. Mathematical formulation. The flow occurs in a right circular cylinder of length L' and radius R'. A pressure node is imposed at the exit plane, as shown in Figure 1, to simplify the calculation of the acoustic field arising from unsteady wall injection. Other, more physically viable pressure conditions at the exit plane could be employed to find alternative acoustic responses. However, the basic modeling concepts, in particular, the role of acoustics in generating intense transient vorticity and heat transfer, will be unaffected.

The mathematical model is based on the nondimensional, compressible Navier–Stokes equations in cylindrical coordinates:

$$\frac{\partial \rho}{\partial t} + M\left(\frac{1}{r}\,\frac{\partial}{\partial r}\,\left(r\rho V_r\right) + \frac{\partial}{\partial z}\,\left(\rho V_z\right)\right) = 0, \tag{1}$$

$$\rho\frac{DV_r}{Dt} = -\frac{\delta^2}{\gamma\ M}\frac{\partial P}{\partial r} + \frac{M\ \delta^2}{Re}\left(2\frac{\partial}{\partial r}\left(\mu\left(\frac{\partial V_r}{\partial r} - \frac{1}{3}\nabla\cdot\overline{\mathbf{V}}\right)\right)\right.$$
$$\left. + \frac{\partial}{\partial z}\left(\mu\left(\frac{1}{\delta^2}\frac{\partial V_r}{\partial z} + \frac{\partial V_z}{\partial r}\right)\right) + 2\frac{\mu}{r}\left(\frac{\partial V_r}{\partial r} - \frac{V_r}{r}\right)\right), \tag{2}$$

$$\rho\frac{DV_z}{Dt} = -\frac{1}{\gamma M}\frac{\partial P}{\partial z} + 2\frac{M}{Re}\frac{\partial}{\partial z}\left(\mu\left(\frac{\partial V_z}{\partial z} - \frac{1}{3}\nabla\cdot\overline{\mathbf{V}}\right)\right)$$
$$+ \frac{M\ \delta^2}{Re}\frac{\partial}{\partial r}\left(\mu\left(\frac{1}{\delta^2}\frac{\partial V_r}{\partial z} + \frac{\partial V_z}{\partial r}\right)\right), \tag{3}$$

$$\rho C_V\frac{DT}{Dt} = -M(\gamma-1)P\nabla\cdot\overline{\mathbf{V}} + \frac{M^3(\gamma-1)\gamma}{Re}\Phi$$
$$+ \frac{M}{Pr}\frac{\delta^2}{Re}\left(\frac{1}{r}\frac{\partial}{\partial r}\left(r\kappa\frac{\partial T}{\partial r}\right) + \frac{1}{\delta^2}\frac{\partial}{\partial z}\left(\kappa\frac{\partial T}{\partial z}\right)\right), \tag{4}$$

$$P = \rho T, \tag{5}$$

where

$$\nabla\cdot\overline{\mathbf{V}} = \frac{\partial V_r}{\partial r} + \frac{V_r}{r} + \frac{\partial V_z}{\partial z},$$
$$\frac{D}{Dt} = \frac{\partial}{\partial t} + M\left(V_r\frac{\partial}{\partial r} + V_z\frac{\partial}{\partial z}\right),$$

and Φ is the dissipation function. The nondimensionalized variables in (1)–(5) are defined by the following:

$$\rho = \frac{\rho'}{\rho_0'}, \quad P = \frac{P'}{P_0'}, \quad T = \frac{T'}{T_0'}, \quad V_r = \frac{V_r'}{V_{r0}'}, \quad V_z = \frac{V_z'}{V_{z0}'},$$
$$r = \frac{r'}{R'}, \quad z = \frac{z'}{L'}, \quad t = \frac{t'}{t_a'}, \quad \kappa = \frac{\kappa'}{\kappa_0'}, \quad \mu = \frac{\mu'}{\mu_0'}, \quad C_V = \frac{C_V'}{C_{V0}'},$$

where P_0' is the initial static pressure in the cylinder and ρ_0' and T_0' are the density and temperature of the fluid being injected from the sidewall. The aspect ratio, given by $\delta = \frac{L'}{R'}$, where $\delta \gg 1$, is chosen to reflect the large aspect ratios found in typical large solid rocket motors. The induced characteristic axial velocity, V_{z0}', is defined with respect to the injection reference sidewall velocity, V_{r0}', by $\frac{V_{z0}'}{V_{r0}'} = \delta$, which is a global characteristic mass conservation statement.

Time is nondimensionalized by the axial acoustic time-scale, $t_a' = \frac{L'}{C_0'}$, where $C_0' = (\gamma\mathcal{R}'T_0')^{\frac{1}{2}}$ is the speed of sound, $\mathcal{R}'$ is the gas constant, and γ is the ratio of specific heats. Thermal diffusivity, viscosity, and specific heat at constant volume, κ_0', μ_0', and C_{V0}', are characteristic properties of the single-species injected fluid. Also the axial Reynolds number, Prandtl number, and axial Mach number are defined as

$$Re = \frac{\rho_0' V_{z0}' L'}{\mu_0'}, \quad Pr = \frac{\mu_0' C_{p0}'}{\kappa_0'}, \quad M = \frac{V_{z0}'}{C_0'}, \tag{6}$$

where $Re \gg 1, M \ll 1$, and $Pr = O(1)$. It is also assumed that hard blowing, defined in Cole and Aroesty [15] by $\delta^2/Re << 1$ prevails. The nondimensional parameter δ^2/Re represents the inverse of the radial Reynolds number with characteristic speed V_{r0}' and length R'.

2.1. Boundary conditions. Initially, for $t \leq 0$, a steady internal flow is generated by an axisymmetric time-independent sidewall radial velocity, $V_r = -V_{rws}(z)$. Subsequently, $t > 0$, a transient disturbance is added to the steady value such that $V_r = -V_{rws}(z) - \widetilde{V}_{rw}(z)(1 - \cos\omega t)$ for $\omega = O(1)$. The transient mass addition is of the same magnitude as the steady component and is always positive, e.g., $\tilde{V}_{rw}(z) > 0$.

The full boundary conditions for this axisymmetric problem are

$$z = 0; \quad V_z = 0, \quad V_r = 0, \quad T = 1, \tag{7}$$

$$z = 1; \quad P = 1, \tag{8}$$

$$r = 0; \quad V_r = 0, \quad \frac{\partial V_z}{\partial r} = \frac{\partial T}{\partial r} = \frac{\partial \rho}{\partial r} = \frac{\partial P}{\partial r} = 0, \tag{9}$$

$$r = 1; \quad V_r = \begin{cases} -V_{rws}(z), & t \leq 0, \\ -V_{rws}(z) - \widetilde{V}_{rw}(z)(1 - \cos\omega t), & t > 0, \end{cases} \tag{10}$$

$$r = 1; \quad V_z = 0, \quad T = 1. \tag{11}$$

The unsteady radial boundary velocity is chosen to ensure that the axial acoustic velocity, driven by the fluctuations in the radial injection velocity, is of the same amplitude as the mean axial velocity flow. As a consequence, the model accommodates nonlinear phenomena.

3. Steady flow solutions. The solution to the steady equations is a base flow arising from the first boundary condition in (10). The steady flow is then altered by imposing the second unsteady boundary condition in (10). The initial step in finding solutions is to divide the velocities and thermodynamic variables,

$$(V_r, V_z, P, \rho, T) = (V_{rs}, V_{zs}, P_s, \rho_s, T_s) + (\widetilde{V}_r, \widetilde{V}_z, \widetilde{P}, \widetilde{\rho}, \widetilde{T}), \tag{12}$$

where the subscript "s" represents the steady part of the flow and ($\tilde{}$) represents the unsteady flow.

The steady variables are then expanded as

$$\begin{aligned} (V_{rs}, V_{zs}) &\sim (V_{r0s}, V_{z0s}) + O(M), \\ (P_s, \rho_s, T_s) &\sim 1 + M^2(P_{0s}, \rho_{0s}, T_{0s}) + O(M^3) \end{aligned} \tag{13}$$

for the limit $M \to 0$ and $\delta^2/Re \to 0$ (see [9, 12]). The leading-order steady equations for the velocity and pressure are found by substitution of (13) into (1)–(5),

$$\frac{1}{r}\frac{\partial (rV_{r0s})}{\partial r} + \frac{\partial V_{z0s}}{\partial z} = 0, \tag{14}$$

$$P_{0s} = P_{0s}(z), \tag{15}$$

$$V_{r0s}\frac{\partial V_{z0s}}{\partial r} + V_{z0s}\frac{\partial V_{z0s}}{\partial z} = -\frac{1}{\gamma}\frac{\partial P_{0s}}{\partial z}, \tag{16}$$

$$V_{r0s}\frac{\partial T_{0s}}{\partial r} + V_{z0s}\frac{\partial T_{0s}}{\partial z} = 0, \tag{17}$$

$$P_{0s} = \rho_{0s} + T_{0s}, \tag{18}$$

and must satisfy the steady boundary conditions,

$$z = 0; \quad V_{z0s} = 0, \quad T_{0s} = 0, \tag{19}$$

$$z = 1; \quad P_{0s} = 0, \tag{20}$$

$$r = 0; \quad V_{r0s} = 0, \quad \frac{\partial V_{z0s}}{\partial r} = \frac{\partial P_{0s}}{\partial r} = \frac{\partial T_{0s}}{\partial r} = \frac{\partial \rho_{0s}}{\partial r} = 0, \tag{21}$$

$$r = 1; \quad V_{r0s} = -V_{rws}(z), \quad V_{z0s} = 0, \quad T_{0s} = 0. \tag{22}$$

Equation (14) shows that the steady flow is incompressible, and the lack of viscous terms in (16) is noted. As shown by Staab et al. [12], evaluation of (16) along the sidewall, $r = 1$, and use of the no-slip condition in (22) shows that a nonzero steady vorticity, Ω_{0s}, is generated on the sidewall,

$$\Omega_{0s}(r = 1, z) = \left(\frac{\partial V_{z0s}}{\partial r} - \frac{\partial V_{r0s}}{\partial z} \right)(r = 1, z) = -\frac{1}{\gamma\, V_{r0s}} \frac{\partial P_{0s}}{\partial z} + \frac{\partial V_{rws}(z)}{\partial z}. \tag{23}$$

In contrast, the steady heat transfer along the sidewall, found from (17) and (22), $\frac{\partial T_{0s}}{\partial r}(r = 1, z)$, is zero. Last, (15) shows that the steady pressure varies only with z, a result of the large aspect ratio assumption, $\delta \gg 1$.

The solutions to (14)–(22) can be written as

$$V_{r0s} = -\frac{V_{rws}(z)}{r} \sin\left(\frac{\pi}{2} r^2\right), \tag{24}$$

$$V_{z0s} = \left(\pi \int_0^z V_{rws}(\widehat{z}) d\widehat{z} \right) \cos\left(\frac{\pi}{2} r^2\right), \tag{25}$$

$$P_{0s} = \gamma \pi^2 \int_z^1 V_{rws}(\widehat{z}) \int_0^{\widehat{z}} V_{rws}(\tau) d\tau d\widehat{z}, \tag{26}$$

$$T_{0s} = 0, \tag{27}$$

$$\rho_{0s} = \gamma \pi^2 \int_z^1 V_{rws}(\widehat{z}) \int_0^{\widehat{z}} V_{rws}(\tau) d\tau d\widehat{z}. \tag{28}$$

The results in (24), (25), and (26), derived by Zhao et al. [9], reduce to those of Culick [20] and Taylor [21] when $V_{rws} = 1$. In the absence of conduction, (17) implies that the steady flow is isothermal.

The r-dependence of the steady radial velocity in (24) will be used in section 4 to define a useful coordinate transformation [9, 22], where

$$s(r) \equiv -\frac{\sin(\pi r^2/2)}{r}. \tag{29}$$

The steady solutions in (24)–(28) provide the initial conditions for the full transient flow. Thus, from (12),

$$t = 0, \quad (\widetilde{V}_r, \widetilde{V}_z, \widetilde{P}, \widetilde{\rho}, \widetilde{T}) = 0. \tag{30}$$

4. Integral transform. Zhao et al. [9] and Zhao and Kassoy [22] show that an integral transform of the radial variable facilitates solution development, particularly for the nonirrotational, transient flow components. The transformed radial variable,

$$\sigma(r) = \int_1^r \frac{1}{s(\hat{r})}\, d\hat{r} = -\frac{1}{\pi} \log\Big(\tan\left(\frac{\pi r^2}{4}\right)\Big), \tag{31}$$

is used to calculate partial derivatives in r,

$$\frac{\partial}{\partial r} = \frac{1}{s}\,\frac{\partial}{\partial \sigma}, \qquad \frac{\partial^2}{\partial r^2} = \frac{1}{s^2}\,\frac{\partial^2}{\partial \sigma^2} + \frac{d}{dr}\left(\frac{1}{s}\right)\frac{\partial}{\partial \sigma}, \tag{32}$$

where s is defined in (29). The transform maps the radial domain $0 \leq r \leq 1$ to $0 \leq \sigma \leq \infty$. The sidewall of the cylinder is at $\sigma = 0$, and the centerline is at $\sigma \to \infty$. The inverse of (31),

$$r(\sigma) = \frac{1}{\sqrt{\pi}}\,\sqrt{\tan^{-1}\left(e^{-\pi\sigma}\right)}, \tag{33}$$

is used to convert σ-dependent functions to the original variable dependence.

Zhao et al. [9] show that two disparate length scales are needed to describe the solution dynamics since physical phenomena are occurring simultaneously on two disparate radial length scales. A multiple-scale analysis can be carried out in terms of the independent variables σ_1 and σ_2 defined by

$$\sigma_1 = \sigma, \qquad \sigma_2 = \frac{\sigma}{M}. \tag{34}$$

The variable σ in (31) can be interpreted physically as the nondimensional time required for a fluid particle injected from a z-location on the wall ($r = 1$) to reach a radial location $r \geq 0$. As the centerline is approached ($r \to 0$), the value of σ becomes unbounded because the steady radial speed in (24) is proportional to $O(r)$. It follows that the σ-variable is used to describe physical variations experienced by an injected fluid particle as it traverses the entire radius of the cylinder, R'. In contrast, the second variable, σ_2, in (34) is used to describe physical variations occurring on the much shorter acoustic time $t'_A = L'/Co'$, during which a fluid particle will move only a small radial distance relative to R'. These simple physical interpretations of (31) and (34) stand in contrast to unsupported criticisms by Majdalani and Flandro [23] suggesting that the transformations are a "...conjectured set of scales found by intuition" and that they are "...different from the uniformly valid scales...prescribed by the problem's solvability condition." The latter remark refers to analysis found in their small disturbance theory and is not relevant to the nonlinear model considered in the present work.

In order to derive accurate reduced forms of the basic descriptive equations in the limit $M \to 0$, the radial variables in (34) must be introduced into (1)–(4) through the use of a chain rule

$$\frac{\partial}{\partial \sigma} = \frac{\partial}{\partial \sigma_1} + \frac{1}{M}\,\frac{\partial}{\partial \sigma_2} \tag{35}$$

prior to taking limits. This approach differs from that of Majdalani and Van Moorhem [23, 24], who introduce the short scale only after deriving reduced equations based on an asymptotic limit for large Reynolds number. This nested expansion approach makes it difficult to account for all appropriate terms in each order of reduced equations.

Figure 2 shows a comparison of the new radial variable, $\sigma_2(r)$, to r_2. Equation (34) defines $\sigma_2(r)$, while r_2 is defined in Staab et al. [12] as $r_2 = \frac{1-r}{M}$. The use of the radial variable r_2 in Staab et al. [12] restricts the maximum domain to $r_2 \leq 1/M$. The new variable's semi-infinite domain has no restriction. Both r_2 and $\sigma_2(r)$ are 0 at the sidewall and increase as they approach the centerline.

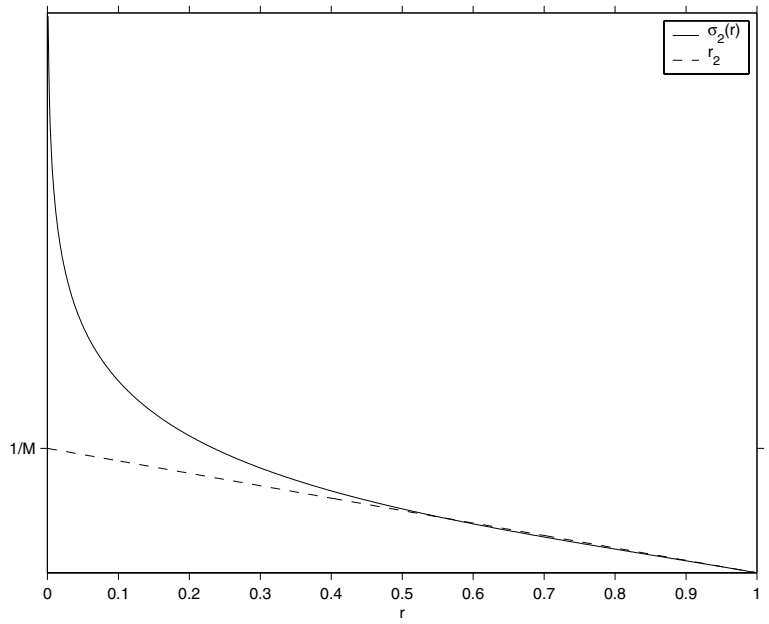


FIG. 2. *Comparison of the new radial variable, $\sigma_2(r)$, to r_2, the radial variable used by Staab et al.* [12].

Staab et al. [12] explain that the coupling between the time and radial direction, found from a method of characteristics solution (see section 5.5), implies a time restriction for a given spatial restriction. Analytic and numerical results in [12] show that $t \leq O(1/M)$. This restriction is absent in the present study.

5. Unsteady flow. The mathematical analysis described here is a generalization and extension of that in Staab et al. [12]. First, the integral transformation in (31) and (34) facilitates the development of solutions valid for extended time values relative to the solutions in [12]. Second, the primary objective of the present modeling is to describe thermal effects including density variations occurring in the flow field. In contrast, the results in [12] describe only the velocity and pressure responses to transient sidewall mass injection for limited values of the time variable.

The modeling paradigm and asymptotic concepts used in [12] are informative with respect to the current work. At the same time, the mathematical analyses differ considerably because the use of the integral transform alters the equations in a substantive way. In this respect, a reader is likely to benefit from a systematic exposition of the asymptotic analysis.

The boundary conditions for the unsteady axisymmetric flow are found using (12), (13), and the difference between (7)–(11) and (19)–(22):

$$z = 0; \quad \widetilde{V}_z = 0, \quad \widetilde{V}_r = 0, \tag{36}$$

$$z = 1; \quad \widetilde{P} = 0, \tag{37}$$

$$\sigma = 0; \quad \widetilde{V}_r = -\widetilde{V}_{rw}(1 - \cos \omega t), \quad \widetilde{V}_z = 0, \quad \tilde{T} = 0, \tag{38}$$

$$\sigma \to \infty; \quad \widetilde{V}_r \to 0, \quad \frac{\partial \widetilde{V}_z}{\partial \sigma} = \frac{\partial \widetilde{T}}{\partial \sigma} = \frac{\partial \widetilde{P}}{\partial \sigma} = \frac{\partial \widetilde{\rho}}{\partial \sigma} \to 0. \tag{39}$$

Asymptotic expansions for the velocity and thermodynamic variables, in the limit $M \to 0$, are

$$\begin{aligned} V_z &\sim V_{z0s}(z,r) + \sum_{n=0}^{\infty} M^n \widetilde{V}_{zn}(z,\sigma,t), \\ V_r &\sim V_{r0s}(z,r) + \sum_{n=0}^{\infty} M^n \widetilde{V}_{rn}(z,\sigma,t), \\ (P,\rho,T) &\sim 1 + \sum_{n=0}^{\infty} M^{n+1}(\widetilde{P}_n, \widetilde{\rho}_n, \widetilde{T}_n), \end{aligned} \tag{40}$$

where the tilde $(\widetilde{\ })$ denotes a time-dependent variable, as in (12). V_{r0s} and V_{z0s} are the steady solutions in (24) and (25).

The large aspect ratio and Reynolds number parameters are connected to the master asymptotic parameter, M, through the relations

$$\delta = \frac{k}{M}, \qquad Re = \frac{1}{CM^4}, \tag{41}$$

where k and C are $O(1)$ constants. The first relationship defines the size of the large aspect ratio chamber relative to a small Mach number. It can be rewritten as $\delta^2 = k/M_i$, where the typical size of the characteristic injection Mach number $M_i = V'_{ro}/C'_o = O(10^{-3})$ (see [2]). This implies that the theory is valid for cylindrical rocket motors with aspect ratios roughly in the range 10 to 30, values that are certainly found in practice.

The second relationship in (41) arises from the limit process analysis itself and is chosen to assure that the effects of viscosity and conductivity are pervasive in the cylinder and maximized with respect to the asymptotic limit, $M \to 0$.

Equation (41) is compatible with the hard blowing limit $\delta^2/Re \to 0$ as $M \to 0$. The relations in (41) permit one to accurately assess the order of magnitude of each and every term in (1)–(4) in the limit $M \to 0$, so that accurate reduced equation systems are derived at each order of M.

5.1. Leading-order equations in the limit $M \to 0$. The derivatives related to the transformation in (31) and (35), the multiple-scale relationships in (34), the unsteady expansions in (40), and the relationship between the parameters δ, Re, and M in (41) are substituted into (1)–(5) prior to applying the limit $M \to 0$. The resulting $O(M^0)$ to $O(M^2)$ approximations of (1) are

$$\frac{\partial \widetilde{V}_{r0}}{\partial \sigma_2} = 0, \tag{42}$$

$$\frac{\mathcal{D}\widetilde{\rho}_0}{\mathcal{D}t} = -\frac{\partial \widetilde{V}_{z0}}{\partial z} - \frac{1}{s}\frac{\partial \widetilde{V}_{r1}}{\partial \sigma_2} - \frac{\widetilde{V}_{r0}}{r} - \frac{1}{s}\frac{\partial \widetilde{V}_{r0}}{\partial \sigma_1}, \tag{43}$$

$$\begin{aligned} \frac{\mathcal{D}\widetilde{\rho}_1}{\mathcal{D}t} = \widetilde{\rho}_0 \frac{\mathcal{D}\widetilde{\rho}_0}{\mathcal{D}t} - \frac{\widetilde{V}_{r1}}{r} - \frac{1}{s}\left(\frac{\partial \widetilde{V}_{r1}}{\partial \sigma_1} + \frac{\partial \widetilde{V}_{r2}}{\partial \sigma_2} + \widetilde{V}_{r1}\frac{\partial \widetilde{\rho}_0}{\partial \sigma_2} + (V_{r0s} + \widetilde{V}_{r0})\frac{\partial \widetilde{\rho}_0}{\partial \sigma_1}\right) \\ - \frac{\partial \widetilde{V}_{z1}}{\partial z} - (V_{z0s} + \widetilde{V}_{z0})\frac{\partial \widetilde{\rho}_0}{\partial z}, \end{aligned} \tag{44}$$

where

$$\frac{\mathcal{D}}{\mathcal{D}t} \equiv \frac{\partial}{\partial t} + \frac{V_{r0s} + \widetilde{V}_{r0}}{s}\frac{\partial}{\partial \sigma_2}. \tag{45}$$

The $\mathcal{D}$ symbol differentiates (45) from the substantial derivative below (5).

The $O(M^0)$ to $O(M^2)$ versions of the radial momentum equation (2),

$$\frac{\partial \widetilde{P}_0}{\partial \sigma_2} = 0, \tag{46}$$

$$\frac{\partial \widetilde{P}_1}{\partial \sigma_2} + \frac{\partial \widetilde{P}_0}{\partial \sigma_1} = 0, \tag{47}$$

$$\frac{\partial \widetilde{P}_2}{\partial \sigma_2} + \frac{\partial \widetilde{P}_1}{\partial \sigma_1} = 0, \tag{48}$$

result from the large aspect ratio assumption.

The $O(M^0)$ and $O(M)$ terms of the axial momentum equation (3) result in

$$\frac{\mathcal{D}\widetilde{V}_{z0}}{\mathcal{D}t} = -\frac{1}{\gamma}\frac{\partial \widetilde{P}_0}{\partial z}, \tag{49}$$

$$\begin{aligned}\frac{\mathcal{D}\widetilde{V}_{z1}}{\mathcal{D}t} = &-\frac{1}{\gamma}\frac{\partial \widetilde{P}_1}{\partial z} - \frac{1}{\gamma}\frac{\partial P_{0s}}{\partial z} + \frac{1}{\gamma}\widetilde{\rho}_0\frac{\partial \widetilde{P}_0}{\partial z} + \frac{Ck^2}{s^2}\frac{\partial^2 \widetilde{V}_{z0}}{\partial \sigma_2^2} - \frac{\widetilde{V}_{r1}}{s}\frac{\partial \widetilde{V}_{z0}}{\partial \sigma_2} \\ &-\frac{\widetilde{V}_{r0} + V_{r0s}}{s}\left(\frac{\partial \widetilde{V}_{z0}}{\partial \sigma_1} + \frac{\partial V_{z0s}}{\partial \sigma_1}\right) - \left(\widetilde{V}_{z0} + V_{z0s}\right)\left(\frac{\partial \widetilde{V}_{z0}}{\partial z} + \frac{\partial V_{z0s}}{\partial z}\right),\end{aligned} \tag{50}$$

where (49) has been used to simplify (50).

The $O(M)$ and $O(M^2)$ approximations to the energy equation (4) are

$$\frac{\mathcal{D}\widetilde{T}_0}{\mathcal{D}t} = \frac{\gamma - 1}{\gamma}\frac{\partial \widetilde{P}_0}{\partial t}, \tag{51}$$

$$\begin{aligned}\frac{\mathcal{D}\widetilde{T}_1}{\mathcal{D}t} = &-\left(\frac{V_{r0s} + \widetilde{V}_{r0}}{s}\,\frac{\partial \widetilde{T}_0}{\partial \sigma_1} + \frac{\widetilde{V}_{r1}}{s}\,\frac{\partial \widetilde{T}_0}{\partial \sigma_2} + (V_{z0s} + \widetilde{V}_{z0})\frac{\partial \widetilde{T}_0}{\partial z}\right) \\ &+\frac{\gamma - 1}{\gamma}\left(\frac{\partial \widetilde{P}_1}{\partial t} + (V_{z0s} + \widetilde{V}_{z0})\frac{\partial \widetilde{P}_0}{\partial z}\right) + \frac{Ck^2}{\partial\ Pr}\,\frac{1}{s^2}\,\frac{\gamma^2 \widetilde{T}_0}{\gamma \sigma_2^2} \\ &-\frac{(\gamma - 1)^2}{\gamma^2}\left(\widetilde{P}_0\frac{\partial \widetilde{P}_0}{\partial t} - \widetilde{T}_0\frac{\partial \widetilde{P}_0}{\partial t}\right),\end{aligned} \tag{52}$$

The somewhat complicated derivation of the latter appears in the appendix.

The $O(M)$ and $O(M^2)$ equations of state, found from (5), are

$$\widetilde{P}_0 = \widetilde{\rho}_0 + \widetilde{T}_0, \tag{53}$$

$$\widetilde{P}_1 = \widetilde{\rho}_1 + \widetilde{T}_1 + \widetilde{\rho}_0\widetilde{T}_0. \tag{54}$$

The σ_1-dependence of the leading-order pressure, $\widetilde{P}_0$, is found by integration of (47) with respect to σ_2,

$$\widetilde{P}_1 = -\frac{\sigma_2}{s(\sigma_1)}\frac{\partial \widetilde{P}_0}{\partial \sigma_1} + \phi(\sigma_1, z, t).$$

The first term on the right-hand side contains unacceptable secular growth with respect to σ_2, typical of a multiple-scale analysis. It must be suppressed by choosing

$$\frac{\partial \widetilde{P}_0}{\partial \sigma_1} = 0, \tag{55}$$

which, together with (46), implies that the pressure, $\widetilde{P}_0 = \widetilde{P}_0(z,t)$, is planar throughout the cylinder. This result is typical of large aspect ratio cylinder geometries.

A similar analysis can be performed for $\widetilde{P}_1$ to show that

$$\frac{\partial \widetilde{P}_1}{\partial \sigma_2} = \frac{\partial \widetilde{P}_1}{\partial \sigma_1} = 0, \tag{56}$$

and hence the pressure $\widetilde{P}_1$ is also planar.

5.2. Leading-order solutions. Complete solutions for the variables $\widetilde{V}_{z0}$, $\widetilde{V}_{r0}$, $\widetilde{P}_0$, and $\widetilde{T}_0$, described by (42)–(53) and the boundary conditions in (36)–(39), can now be obtained.

The solution development is analogous to that of Zhao et al. [9], in which the variables are split into weak ($\overline{\ }$, implies radial dependence only on σ_1) and strong ($\widehat{\ }$, implies both σ_1 and σ_2 radial dependence) rotational components,

$$\begin{aligned}
\widetilde{V}_{r0} &= \overline{V}_{r0}(z,t,\sigma_1),\\
\widetilde{V}_{r1} &= \overline{V}_{r1}(z,t,\sigma_1) + \widehat{V}_{r1}(z,t,\sigma_1,\sigma_2),\\
\widetilde{V}_{z0} &= \overline{V}_{z0}(z,t,\sigma_1) + \widehat{V}_{z0}(z,t,\sigma_1,\sigma_2),\\
\widetilde{P}_0 &= \overline{P}_0(z,t),\\
\widetilde{\rho}_0 &= \overline{\rho}_0(z,t,\sigma_1) + \widehat{\rho}_0(z,t,\sigma_1,\sigma_2),\\
\widetilde{T}_0 &= \overline{T}_0(z,t,\sigma_1) + \widehat{T}_0(z,t,\sigma_1,\sigma_2).
\end{aligned} \tag{57}$$

The radial velocity $\widehat{V}_{r0}$ has no strong component due to (42), while (46) and (55) preclude any radial dependence for pressure.

The naming convention "strong" and "weak" rotational flow has been used in Staab et al. [12]. Upon substitution of (57) into (42)–(53), two sets of equations can be derived. The weak terms are dependent only on σ_1, z, and t. In contrast, the strong rotational terms depend on σ_1, σ_2, z, and t.

5.3. Weak rotational equations. The leading-order weak rotational unsteady equations, derived from (43), (49), (51), and (53), are

$$\frac{\partial \overline{\rho}_0}{\partial t} = -\frac{1}{r}\frac{\partial (r\overline{V}_{r0})}{\partial r} - \frac{\partial \overline{V}_{z0}}{\partial z}, \tag{58}$$

$$\frac{\partial \overline{V}_{z0}}{\partial t} = -\frac{1}{\gamma}\frac{\partial \overline{P}_0}{\partial z}, \tag{59}$$

$$\frac{\partial \overline{T}_0}{\partial t} = \frac{\gamma - 1}{\gamma}\frac{\partial \overline{P}_0}{\partial t}, \tag{60}$$

$$\overline{P}_0 = \overline{\rho}_0 + \overline{T}_0, \tag{61}$$

where the radial variable σ_1 has been replaced with the original radial variable, r, for convenience.

The boundary conditions for these equations are subsets of those in (36)–(39),

$$z = 0; \quad \overline{V}_{z0} = 0, \tag{62}$$

$$z = 1; \quad \overline{P}_0 = 0, \tag{63}$$

$$r = 0; \quad \overline{V}_{r0} = 0, \qquad \frac{\partial \overline{V}_{z0}}{\partial r} = \frac{\partial \overline{T}_0}{\partial r} = \frac{\partial \overline{P}_0}{\partial r} = \frac{\partial \overline{\rho}_0}{\partial r} = 0, \tag{64}$$

$$r = 1; \quad \overline{V}_{r0} = -\widetilde{V}_{rw}(z)(\cos \omega t - 1). \tag{65}$$

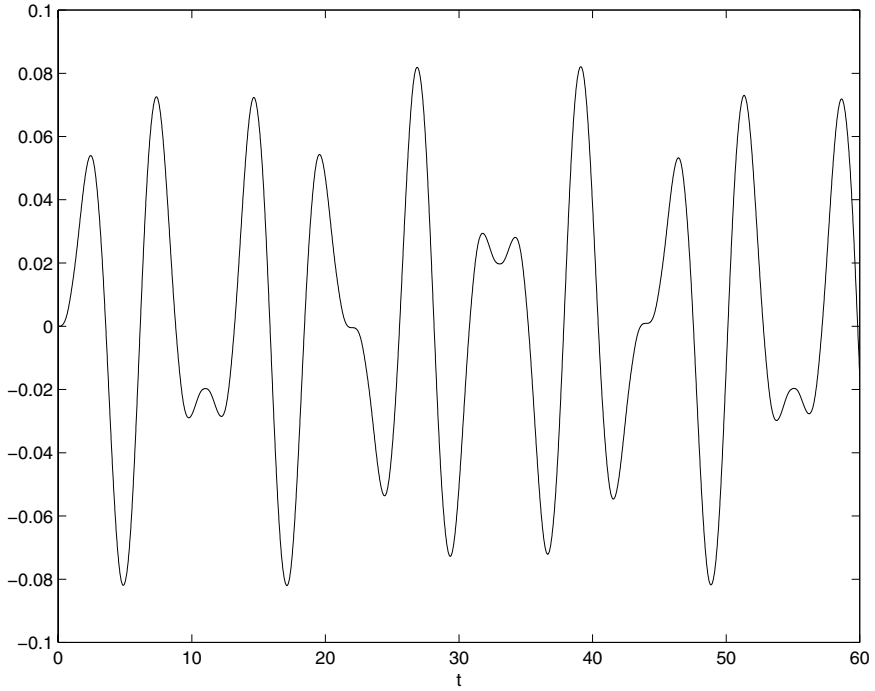


FIG. 3. *The temperature solution of* $\overline{T}_0$ *vs.* t *at* $z = 0.5$ *for* $\omega = 1$ *has a period roughly equal to* 45 *time units.*

The solutions to (58)–(65) are those found in Staab et al. [12],

$$\overline{P}_0 = \sum_{n=0}^{\infty} \frac{2c_n\gamma\omega}{\omega^2 - b_n^2} \left(\frac{\omega}{b_n} \sin b_n t - \sin \omega t \right) \cos b_n z, \tag{66}$$

$$\overline{V}_{z0} = \sum_{n=0}^{\infty} \frac{2\omega c_n}{\omega^2 - b_n^2} \left(\frac{\omega}{b_n}(1 - \cos b_n t) - \frac{b_n}{\omega}(1 - \cos \omega t) \right) \sin b_n z, \tag{67}$$

$$\overline{T}_0 = \sum_{n=0}^{\infty} \frac{2c_n(\gamma - 1)\omega}{\omega^2 - b_n^2} \left(\frac{\omega}{b_n} \sin b_n t - \sin \omega t \right) \cos b_n z, \tag{68}$$

$$\overline{\rho}_0 = \sum_{n=0}^{\infty} \frac{2\omega c_n}{\omega^2 - b_n^2} \left(\frac{\omega}{b_n} \sin b_n t - \sin \omega t \right) \cos b_n z, \tag{69}$$

$$\overline{V}_{r0} = -r\widetilde{V}_{rw}(z)(1 - \cos \omega t), \tag{70}$$

where

$$c_n = 2 \int_0^1 \widetilde{V}_{rw}(z) \cos b_n z \, dz, \tag{71}$$

and $b_n = (n + 1/2)\pi$.

The $\overline{T}_0$ solution is planar and contains both a boundary condition response ($\sin \omega t$) and eigenfunction ($\sin b_n t$) relevant to the exit plane pressure node boundary condition. The latter can be rewritten as a sum of two counter-propagating planar waves, while the boundary condition response represents a standing wave. A time

series plot of $\overline{T}_0$ is found in Figure 3 at $z = 0.5$ for $\omega = 1$. The period is roughly 45 time units. The relatively small positive and negative peaks near $t = 10, 33$, and 55 will be seen in the results for the full temperature field, showing a connection between the acoustic field and the overall temperature variation.

5.4. Steady parts of the weak rotational solutions. Equations (67) and (70) reveal that $\overline{V}_{z0}$ and $\overline{V}_{r0}$ contain both mean and fluctuating parts. It will be shown in section 6 that the means of $\overline{V}_{z0}$ and $\overline{V}_{r0}$, important in describing the temperature dynamics, can be written [12] as

$$\overline{V}_{z0s} = 2\int_0^z \widetilde{V}_{rw}(\hat{z})d\hat{z}, \tag{72}$$

$$\overline{V}_{r0s} = -r\widetilde{V}_{rw}(z). \tag{73}$$

5.5. Strong rotational equations. The leading-order strong rotational equations are found by subtracting (58)–(61) from (43), (49), (51), (53),

$$\frac{\mathcal{D}\widehat{\rho}_0}{\mathcal{D}t} = -\frac{\partial \widehat{V}_{z0}}{\partial z} - \frac{1}{s}\frac{\partial \widehat{V}_{r1}}{\partial \sigma_2}, \tag{74}$$

$$\frac{\mathcal{D}\widehat{V}_{z0}}{\mathcal{D}t} = 0, \tag{75}$$

$$\frac{\mathcal{D}\widehat{T}_0}{\mathcal{D}t} = 0, \tag{76}$$

$$\widehat{\rho}_0 + \widehat{T}_0 = 0, \tag{77}$$

where $\frac{\mathcal{D}}{\mathcal{D}t}$ is defined in (45).

Equations (76) and (77) can be combined to show that

$$\frac{\mathcal{D}\widehat{\rho}_0}{\mathcal{D}t} = 0, \tag{78}$$

and therefore (74) can be written as

$$\frac{\partial \widehat{V}_{z0}}{\partial z} + \frac{1}{s}\frac{\partial \widehat{V}_{r1}}{\partial \sigma_2} = 0. \tag{79}$$

Equations (75), (76), and (78) are first-order wave equations defining invariance along characteristics defined by the radial velocity field, $V_{r0s} + \widetilde{V}_{r0}$. The wave-like solutions to (75), (76), and (78) originate on the sidewall and convect toward the centerline with the local radial velocity. The solution to (76) for the temperature, $\widehat{T}_0$, is obtained following the procedures in Staab et al. [12].

The relevant initial/boundary conditions for (76) are

$$\begin{aligned} t = 0; \qquad & \widehat{T}_0 = 0, \\ \sigma_1 = 0 \quad \text{and} \quad \sigma_2 = 0; \qquad & \widehat{T}_0 = -\overline{T}_0, \end{aligned} \tag{80}$$

where $\overline{T}_0$ is the solution in (68). The second condition demonstrates that the acoustic solution in (68) is a forcing function for $\hat{T}_0$. Equation (76), with the substantial derivative represented by (45), can be solved along a set of characteristics found by integration of

$$\frac{d\sigma_2}{dt} = \frac{V_{r0s}(\sigma_1, z) + \overline{V}_{r0}(\sigma_1, z, t)}{s(\sigma_1)} = V_{rws}(z) + \frac{\overline{V}_{r0}}{s}, \tag{81}$$

where σ_1 and z are treated as constants in the multiple-scale analysis. It follows that (81) can be integrated from 0 to t with the integration constant found from $\sigma_2 = \xi$ at $t = 0$, and σ_1 and σ_2 treated as independent variables. The equation characteristics are described by the result

$$\xi - \sigma_2 = -tV_{rws}(z) - \frac{r^2\widetilde{V}_{rw}(z)}{\sin(\pi r^2/2)}\left(t - \frac{\sin\omega t}{\omega}\right), \tag{82}$$

where r is taken to be $r(\sigma_1)$ as defined by the inverse of the function in (31).

The characteristic, $\xi = 0$, represents the front of the strongly rotational temperature distribution described by (76), which convects away from the sidewall toward the centerline with the speed defined by (81).

An alternative representation of the characteristic curves in (82) can be found by integration of (81) from $t^\star$ to t with the integration constant found from $\sigma_2 = 0$ when $t = t^\star$,

$$\sigma_2 = V_{rws}(z)(t - t^\star) + \frac{r^2\widetilde{V}_{rw}(z)}{\sin(\pi r^2/2)}\left(t - t^\star + \frac{\sin\omega t^\star - \sin\omega t}{\omega}\right). \tag{83}$$

This representation is useful because the solution to (76) along $\sigma_1 = 0$ can be found by integration and the application of the boundary/initial conditions in (80),

$$\widehat{T}_0(t^\star; z) = \begin{cases} 0, & t^\star \le 0, \\ -\overline{T}_0(t^\star, z), & t^\star > 0. \end{cases} \tag{84}$$

The characteristic variable $t^\star$ is the solution of (83) for a given σ_2, z, and t when $\sigma_1 = 0$.

This analysis yields only $\widehat{T}_0(\sigma_1 = 0, \sigma_2, z, t)$, given in (84). As is typical of a multiple-scale analysis, higher-order equations must be considered to resolve behavior on the longer scale, in this case, σ_1.

6. Higher-order considerations. The $O(M^2)$ energy equation in (52) is used in the higher-order analysis to determine the σ_1-dependence of the temperature, $\widehat{T}_0(\sigma_1, \sigma_2, z, t)$.

Similar to the procedure used for the lower-order variables, the $O(M^2)$ unsteady pressure and $O(M)$ unsteady axial velocity are written in the form

$$\begin{aligned} \widetilde{P}_1 &= \overline{P}_1(z,t), \\ \widetilde{V}_{z1} &= \overline{V}_{z1}(\sigma_1, z, t) + \widehat{V}_{z1}(\sigma_1, \sigma_2, z, t), \\ \widetilde{T}_1 &= \overline{T}_1(\sigma_1, z, t) + \widehat{T}_1(\sigma_1, \sigma_2, z, t), \end{aligned} \tag{85}$$

where the weak and strong rotational components of the temperature are denoted by the bar and caret superscripts, respectively. The pressure $\widetilde{P}_1$ is shown to be planar in (56), and therefore no radially dependent pressure term exists.

Substitution of (57) and (85) into (52) yields two sets of energy equations. The first set contains all of the terms of (52) with dependence on σ_1, z, and t,

$$\begin{aligned} \frac{\partial\overline{T}_1}{\partial t} =& \frac{\gamma-1}{\gamma}\frac{\partial\overline{P}_1}{\partial t} - (V_{z0s} + \overline{V}_{z0})\frac{\partial\overline{T}_0}{\partial z} + \frac{\gamma-1}{\gamma}(V_{z0s} + \overline{V}_{z0})\frac{\partial\overline{P}_0}{\partial z} \\ & - \frac{(\gamma-1)^2}{\gamma^2}\left(\overline{P}_0\frac{\partial\overline{P}_0}{\partial t} - \overline{T}_0\frac{\partial\overline{P}_0}{\partial t}\right), \end{aligned} \tag{86}$$

while the second equation is the difference between (52) and (86),

$$\frac{\mathcal{D}\widehat{T}_1}{\mathcal{D}t} = -\left(\frac{V_{r0s}+\overline{V}_{r0}}{s}\frac{\partial\widehat{T}_0}{\partial\sigma_1} + \frac{\widetilde{V}_{r1}}{s}\frac{\partial\widehat{T}_0}{\partial\sigma_2} + (V_{z0s}+\overline{V}_{z0}+\widehat{V}_{z0})\frac{\partial\widehat{T}_0}{\partial z} + \widehat{V}_{z0}\frac{\partial\overline{T}_0}{\partial z}\right)$$
$$+\frac{\gamma-1}{\gamma}\widehat{V}_{z0}\frac{\partial\overline{P}_0}{\partial z} - \frac{(\gamma-1)^2}{\gamma^2}\widehat{T}_0\frac{\partial\overline{P}_0}{\partial t} + \frac{Ck^2}{\gamma\, Pr}\frac{1}{s^2}\frac{\partial^2\widehat{T}_0}{\partial\sigma_2^2}. \tag{87}$$

When (87) is integrated in time, several terms on the right-hand side are sources of unbounded solution behavior in time [12]. The unphysical secular behavior can be suppressed if

$$0 = \frac{V_{r0s}+\overline{V}_{r0s}}{s}\frac{\partial\widehat{T}_0}{\partial\sigma_1} + (V_{z0s}+\overline{V}_{z0s})\frac{\partial\widehat{T}_0}{\partial z} - \frac{Ck^2}{\gamma\, Pr}\frac{1}{s^2}\frac{\partial^2\widehat{T}_0}{\partial\sigma_2^2}, \tag{88}$$

which describes the complete solution $\widehat{T}_0(\sigma_1,\sigma_2,z,t)$. This convection-diffusion equation for the strongly rotational temperature field is analogous to that for the axial speed $\hat{V}_{z0}$,

$$0 = \frac{V_{r0s}+\overline{V}_{r0s}}{s}\frac{\partial\widehat{V}_{z0}}{\partial\sigma_1} + \frac{\partial}{\partial z}\left(\widehat{V}_{z0}\left(\frac{1}{2}\widehat{V}_{z0}+V_{z0s}+\overline{V}_{z0s}\right)\right) - \frac{Ck^2}{\gamma}\frac{1}{s^2}\frac{\partial^2\widehat{V}_{z0}}{\partial\sigma_2^2}. \tag{89}$$

Both the first and third terms in (88) have analogues in (89) for $Pr = 1$. However, the axial convection term in (89) is nonlinear, while that in (88) is linear. Also, the solutions to the two equations will have different properties arising from the boundary conditions imposed on each.

The boundary conditions for (88) have the form

$$\begin{aligned} z = 0, &\quad \widehat{T}_0 = 0; \\ \sigma_2 = 0, &\quad \widehat{T}_0 = \widehat{T}_0(t^\star, z); \\ \sigma_1,\sigma_2 \to \infty, &\quad \widehat{T}_0 \to 0. \end{aligned} \tag{90}$$

The physical solution, $\widehat{T}_0(r,z,t)$, is found by evaluating the function $\widehat{T}_0(\sigma_1,\sigma_2,z,t)$ along $\sigma_1 = M\sigma_2$ and returning to the original radial variable using (33). A numerical solution to (88) is described in section 11.

7. Density field dynamics. The complete leading-order temperature solution can be used to obtain the density. As with the temperature, the density is composed of a steady field, ρ_s, and an unsteady field, $\widetilde{\rho}$. The latter is defined by an asymptotic expansion in powers of M as in (40). In short, the leading-order density field is $\rho = 1 + M(\widehat{\rho}_0 + \overline{\rho}_0) + O(M^2)$, where $\overline{\rho}_0$ is the solution in (69) and $\widehat{\rho}_0$ is explained here.

The strongly rotational density, $\widehat{\rho}_0$, is found from (77), once $\widehat{T}_0$ is found from (88). The qualitative behavior of the density field is analogous to that of the temperature field. A large radial gradient of the density arising along the sidewall of the cylinder will be convected downstream and toward the centerline.

8. Heat transfer along the sidewall. The strongly rotational energy equation, (76) can be used to determine the heat transfer along the sidewall. The dimensional heat transfer per unit area,

$$q' = \kappa'\frac{\partial T'}{\partial r'}, \tag{91}$$

can be nondimensionalized using the definition

$$q_0' = \frac{\kappa_0' T_0'}{R'}$$

to find the first approximation

$$q' \sim \frac{q_0'}{V_{r0s}} \frac{\partial \widehat{T}_0}{\partial \sigma_2}, \tag{92}$$

where (32), (35), and (40) have been used. The characteristic radial heat flux q_0' is far larger than one might expect in a flow where the typical temperature variation is $O(MT_0')$. This $O(1)$, rather than $O(M)$, dependence arises from the presence of large radial gradients on the scale $O(MR')$.

The heat transfer along the sidewall can be found using the leading-order conservation of energy in (51), the sidewall boundary condition in (10), and the pressure solution in (66),

$$\begin{aligned}\left(\frac{q_{\text{wall}}'}{q_0'}\right) = q_{\text{wall}} \equiv \frac{1}{s}\frac{\partial \widehat{T}_0}{\partial \sigma_2}\bigg|_{\text{wall}} &= \frac{1}{\left(V_{r0s} + \widetilde{V}_{r0}\right)} \frac{\partial \overline{T}_0}{\partial t}\bigg|_{\text{wall}} \\ &= \frac{1}{V_{rws} + \widetilde{V}_{rw}(z)\sin\omega t} \sum_{n=0}^{\infty} \frac{2c_n(\gamma-1)\omega}{\omega^2 - b_n^2} (\cos b_n t - \cos\omega t)\cos b_n z. \end{aligned} \tag{93}$$

This surprising result shows that the basic heat transfer is determined by a nonconductive interaction between the acoustic field and the injected fluid.

It should be noted that the heat flux is found from a transport equation (51) in which conduction is absent. This property of the equations is analogous to that of the inviscid momentum equation in (75) which is compatible with the no-slip condition. Hard blowing problems ($\delta^2/Re \ll 1$) are characterized by relatively diminished influence of transport effects on near-surface gradients.

Related wall heat transfer analyses have been done by Staab et al. [12], who present results for different frequencies of the sidewall mass addition.

9. Properties of $\widehat{T}_0$. The effective conductivity for the $\widehat{T}_0$-equation (88) is proportional to $s^{-1}(V_{r0s} + \overline{V}_{r0s})^{-1}$. As the centerline is approached, $r \to 0$, the conductivity becomes unbounded like $1/r^2$. In a linear conduction problem, such conductivity behavior implies that in the limit $r- > 0$, $\widehat{T}_0 \to 0$.

Qualitatively, the linear convection-conduction equation (88) describes convection of thermal energy toward the centerline and downstream, and conduction in the radial direction. However, much more can be said about the properties of the equation. For example, (88) can be written as a linear conduction equation

$$\frac{\partial \widehat{T}_0}{\partial \sigma_1} = \frac{Ck^2}{\gamma Pr} \frac{1}{s(V_{r0s} + \overline{V}_{r0s})} \frac{\partial^2 \widehat{T}_0}{\partial \sigma_2^2} \tag{94}$$

along characteristic curves $z = z(\sigma_1)$ defined by

$$\frac{dz}{d\sigma_1} = \frac{s(V_{z0s} + \overline{V}_{z0s})}{V_{r0s} + \overline{V}_{r0s}}. \tag{95}$$

These curves are streamlines determined by the lowest-order total axial and radial speeds. The slope $dz/d\sigma_1 > 0$, as determined from the right-hand side of (95), which is positive for all σ_1 and z. The streamlines start on the sidewall, $\sigma_1 = 0$, and leave through the exit plane, $z = 1$. All solutions for $\widehat{T}_0$ can be described along these characteristics. Such curves will not cross due to the lack of $\widehat{T}_0$ in (95), a result of the linear form of (88). Equation (95) can be evaluated for $r \to 0$ to show that the streamlines become parallel to the axis ($r = 0$) as $r \to 0$.

The heat equation in (94) can be solved using Fourier transform techniques. A simplified version is solved in Staab et al. [12] and used to show that solutions near the centerline decay with form

$$\lim_{r\to 0} \widehat{T}_0 \sim \left(r\ \frac{\pi}{4}\right)^{\frac{Ck^2\omega^2}{\pi(1+A)^2}}, \tag{96}$$

where the original variable r has been used, C and k are the constants found in (41), and A is an average transverse vorticity wave speed used to simplify the boundary condition. Equation (96) shows that the strongly rotational temperature $\widehat{T}_0$ goes to zero at the centerline. Only the acoustic solution, $\overline{T}_0$ in (68), contributes to the centerline temperature.

10. Solution development and numerical methodology. Solutions to (88) and (89) are found using a procedure similar to that employed by Staab et al. [12]. However, in this case there is no upper limit on the transformed radial variable, σ_2, so the solution can be solved for a much larger time domain.

First the results in (80)–(84) are used to find $\widehat{T}_0(\sigma_1 = 0, \sigma_2 = 0, z, t)$, which is the initial condition for (88). The numerical solution of $\widehat{T}_0(\sigma_1, \sigma_2, z, t)$ at each z and t is evaluated along $\sigma_1 = M\sigma_2$ to yield the solution in (σ, z, t)-space. Finally, (33) is used to return to (r, z, t)-space.

It is noted that t is an implicit variable in (88), with σ_1 playing the role of the timelike variable. For a given value of t the solution to the wave equation (76) on $\sigma_1 = 0$ penetrates a distance defined by the front $\xi = 0$ in (82). Equation (88) describes the convection and diffusion effects on $\widehat{T}_0$ between the wall and the front.

Equation (88) is solved via the method of lines, a numerical technique which employs a spatial discretization to reduce a time-dependent partial differential equation (PDE) to a system of ordinary differential equations (ODEs) in the timelike variable. Since convection is downstream, a second-order backward finite difference formula in the z-direction is the basis for the axial spatial discretization. In this manner the finite differencing is stable since all information comes from the upstream direction. A fourth-order centered finite difference formula is used for the $\frac{\partial^2 \widehat{T}_0}{\partial \sigma_2^2}$ term. The discretization leads to a set of $N_{\sigma_2} \times N_z$ coupled ODEs, where N_z and N_{σ_2} are the number of gridpoints chosen in the z- and σ_2-directions, respectively. This set of ODEs is solved using an adaptive fourth-order Runge–Kutta solver, a stable method for solving a set of ODEs that arise from the discretization of equations with both parabolic and hyperbolic terms.

The computational results in the following section are produced using 150 grid points in the radial direction and 50 in the axial direction. The figures shown in that section were produced with 300 radial gridpoints, where the new points are found using a spline interpolation based on the original 150 points. This was done in order to save computing time in making the plots since running the solver with high grid density takes much longer than performing a spline interpolation. Solving (88)

with 300 radial gridpoints took 89 minutes of computing time, whereas 150 radial gridpoints and then splines took only 10 minutes. All computations were carried out on a SuperSparc workstation.

Before discussing the error generated at each extra gridpoint from using splines, it is necessary to define the following variables:

$$U_{150} = \widehat{T}_0 \quad \text{with } N_{\sigma_2} = 150, \tag{97}$$

$$U_{300} = \widehat{T}_0 \quad \text{with } N_{\sigma_2} = 300, \tag{98}$$

$$U_s = \text{spline}(U_{150}), \tag{99}$$

where U_s comes from taking U_{150} and spline interpolating it to 300 points. The notation N_{σ_2} denotes the number of gridpoints used in the σ_2-direction for the computational solution.

The error at each extra gridpoint is calculated using the following formula:

$$\epsilon = \frac{\|U_{300} - U_s\|_\infty}{\|U_{300}\|_\infty}. \tag{100}$$

The maximum error, ϵ, computed using (100) on a representative test problem is 1.78×10^{-2}. This is considered to be acceptable since the next term in the asymptotic expansion for the temperature is $O(M)$ smaller than $\widehat{T}_0$ and the Mach numbers used in this study are slightly larger than the error. Therefore, the error associated with the asymptotic series is larger than that due to the spline interpolation process.

11. Results. The spatial distribution of the instantaneous temperature and temperature gradient are discussed in this section in order to explain the origin and evolution of thermal disturbances for the sidewall boundary condition in (10), with $V_{rws}(z) = 1$ and $\tilde{V}_{rw}(z) = 0.2\cos\left(\frac{\pi z}{2}\right)$;

$$V_r = -1 - 0.2\cos\left(\frac{\pi z}{2}\right)(1 - \cos\omega t) \qquad \text{at} \quad r = 1. \tag{101}$$

It is noted that the additional unsteady mass increase is on the order of 20% of that due to uniform injection.

Figure 4 shows the instantaneous rotational temperature $\widehat{T}_0$ variation with r and z, for $M = 0.02$, $\delta = 20$, $Re = 3 \times 10^5$, $\omega = 1$, $Pr = 1$, and $\gamma = 1.4$ at $t = 20$. The temperature disturbances, which begin on the sidewall, are driven by the third condition in (90). Three spatial waves characterized by large radial gradients have propagated into the cylinder interior to about $r = 0.5$. Beyond this first wave the solution is $\widehat{T}_0 = 0$. The 0.001-contour nearest the centerline will hereafter be referred to as the "$\widehat{T}_0$-front."

Two mechanisms give the surface in Figure 4 its characteristic morphology. First, along the sidewall of the cylinder ($r = 1$) the solution is given in (84). The time-dependence of the solution along the edge is the negative of $\widehat{T}_0$ shown in Figure 3. The waves are generated along the sidewall and convected into the flow field by the radial velocity field as shown in (76), where the substantial derivative is defined in (45). The other mechanisms governing the shape of the solution are the convective and conductive effects in (88). The waves are convected toward the centerline and downstream and diffused on the length-scale associated with the σ_2-variable.

Observation of each temperature wave in Figure 4 shows that the magnitude decreases in the downstream direction. In contrast, the analogous rotational axial velocity increases with z (see [12]). The difference can be attributed to the mathematical

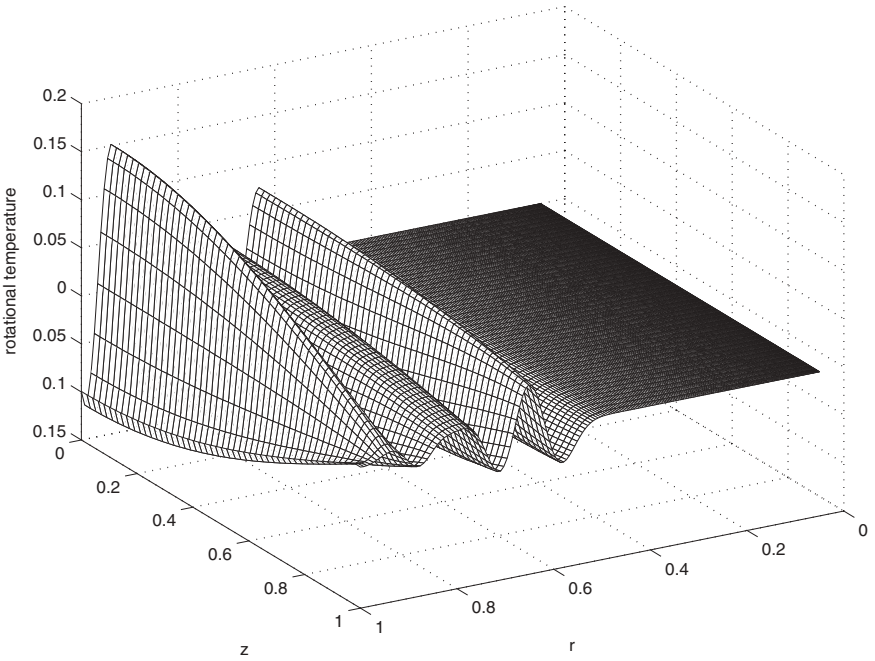


FIG. 4. *The variation of* $\widehat{T}_0$ *with* r *and* z *for* $M = 0.02$, $\delta = 20$, $Re = 3 \times 10^5$, $\omega = 1$, $Pr = 1$, $\gamma = 1.4$, *and at* $t = 20$. *Temperature disturbances originate on the sidewall and convect toward the centerline. The temperature front has reached about* $r = 0.5$ *at this time. The fluid ahead of the front remains at its initial temperature of* $T = 1$.

form of the acoustic solutions in (67) and (68). The acoustic axial velocity contains a $\sin\left(\frac{\pi z}{2}\right)$ term which increases with z, while the acoustic temperature has a $\cos\left(\frac{\pi z}{2}\right)$ term which decreases in the same region. These terms arise from the z-dependence of the sidewall radial velocity in (10). The radial sidewall velocity drives the acoustic temperature field that ultimately generates the strongly rotational temperature.

Figure 5 shows the (r, z) variation of the complete temperature $T \sim 1 + M(\widehat{T}_0 + \overline{T}_0) + O(M^2)$, where $\widehat{T}_0$ is the solution to (76) and (88) and $\overline{T}_0$ is defined in (68). The results are given at $t = 20$ and 30 for the parameter values in Figure 4. Since the boundary condition in (10) has a period of 2π when $\omega = 1$, at $t = 30$, one should expect to find about five spatial waves, as observed in Figure 5(b).

The instantaneous acoustic temperature, $\overline{T}_0$, is described by the "ramp" in Figures 5(a) and 5(b), where $\widehat{T}_0 = 0$. Also, since it is proportional to $\cos\left(\frac{\pi z}{2}\right)$, it always approaches 0 at $z = 1$. The variations with z and t differ according to the dependencies in (68). The impact of the $\widehat{T}_0$-solution can be observed in the region between the front location and $r = 1$. Note the different vertical scales used in Figures 5(a) and 5(b). The amplitude of the solution in Figure 5(a) is almost entirely greater than 1, while that in Figure 5(b) is mostly less than 1. Again, this difference arises from the instantaneous acoustic temperature field. It is noted from Figure 3 that $\widehat{T}_0$ is positive at $t = 20$ and negative at $t = 30$. Figure 6, a contour plot of Figure 5, shows clearly the axial shape of the waves and the radial location of the $\widehat{T}_0$-front near $r = 0.45$. The pure acoustic field is associated with horizontal contours on the left side of the figure. To the right of the $\widehat{T}_0$-front and extending to the sidewall, tightly packed contours

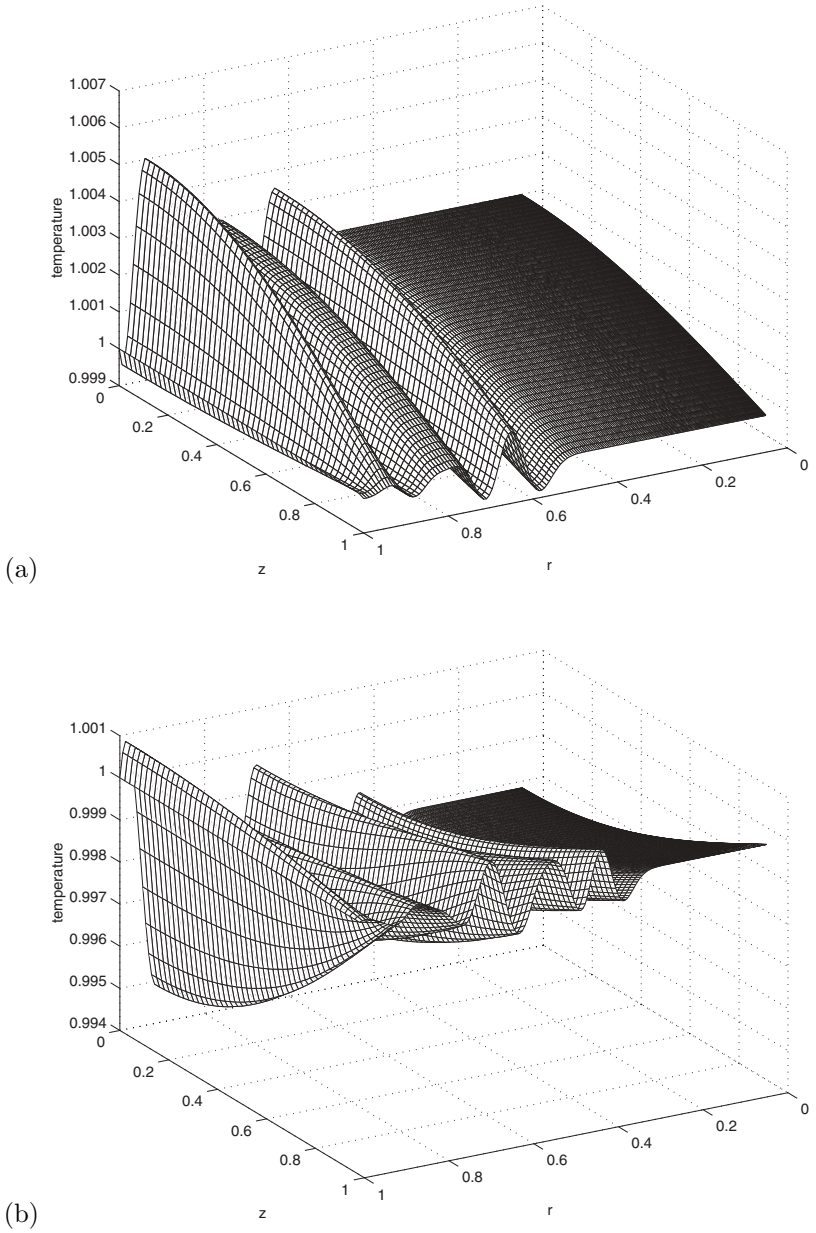


FIG. 5. *Variation of the complete temperature* $T \sim 1 + M(\widehat{T}_0 + \overline{T}_0)$ *with* r *and* z *for* $M = 0.02$, $\delta = 20$, $Re = 3 \times 10^5$, $\omega = 1$, $Pr = 1$, *and* $\gamma = 1.4$ *at* (a) $t = 20$ *and* (b) $t = 30$. *The characteristics of the field near* $r = 1$ *are dominated by* $\widehat{T}_0$, *while that near the centerline is dominated by* $\overline{T}_0$. *Note that different vertical scales have been used for the two plots.*

represent regions of large radial gradients. The shading depicts variations from the background of $T = 1$.

A surface plot of the radial temperature gradient, $\frac{\partial T}{\partial r}$, is shown in Figure 7 using the parameters in Figure 5(a). Pronounced gradients persist up to the location of

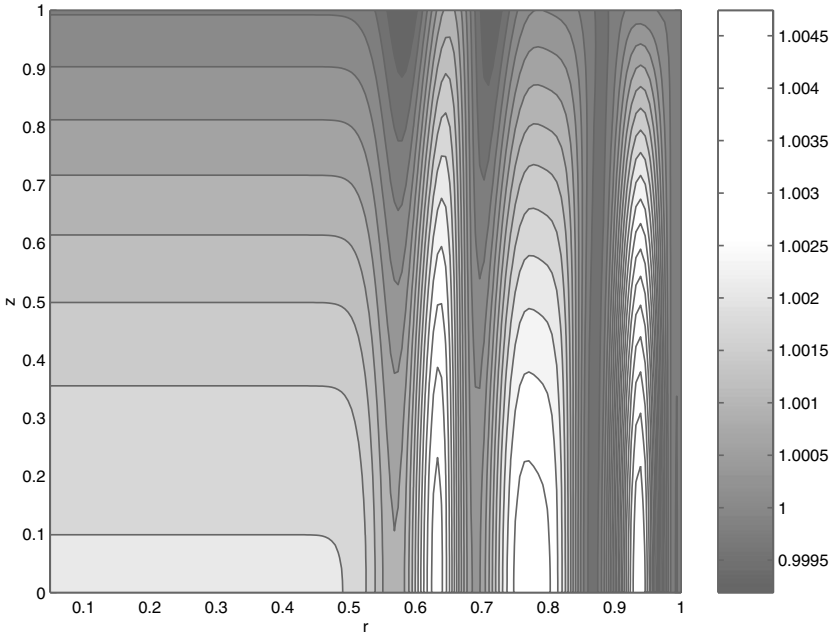


FIG. 6. *Contour plot of the complete temperature T for the parameters in Figure* 5(a).

the $\widehat{T}_0$-front. The variation in the temperature gradient, typically about 0.1, is of the same magnitude as that for $\widetilde{V}_{rw}$. This is in contrast to the leading-order temperature variation, $T-1 \sim M(\widehat{T}_0+\overline{T}_0)+O(M^2)$, roughly .005, as can be seen in Figure 5(a). The variation in the temperature gradient is roughly twenty times larger than that of the temperature in this case, a result of the short wavelength of the spatial oscillations in Figure 5(a). Since the conductive heat transfer is proportional to the temperature gradient, this means that surprisingly large amounts of heat transfer exist in the cylinder even though the temperature perturbations are small.

Figure 8(a) shows the (r,z) variation of the rotational temperature, $\widehat{T}_0$, at $t=60$ with all the other parameters the same as those in Figure 5. At any given axial location, it can be seen that the amplitude of the radial oscillations varies considerably as one moves between the sidewall and the axis of the cylinder. The monotonic variations arise because the solution to (88) along $r=1$ is $-\overline{T}_0$, as shown in (84). Since $-\overline{T}_0$ is not monotonic in time, as shown in Figure 3, the waves that originate on the sidewall are not strictly decreasing in time. However, if one follows a given spatial wave originating at the sidewall and convecting into the cylinder, it will be observed that the wave magnitude is diminished. Staab et al. [12] have shown that this is the result of accumulated convection and diffusion. One may note the smaller amplitude spatial waves near the cylinder axis.

Figure 8(b) describes the spatial dependence of the strong rotational axial velocity, $\overline{V}_{z0}$, for $t=60$ and the same parameter values as those for Figure 8(a). This long-time solution could not be obtained in the earlier work by Staab et al. [12] because of the limitations described earlier in section 4. The results in Figures 8(a) and 8(b) show that the $\widehat{V}_{z0}$ solution is damped more quickly in the radial direction than is the $\widehat{T}_0$ solution. This difference can be attributed to the characteristic of the

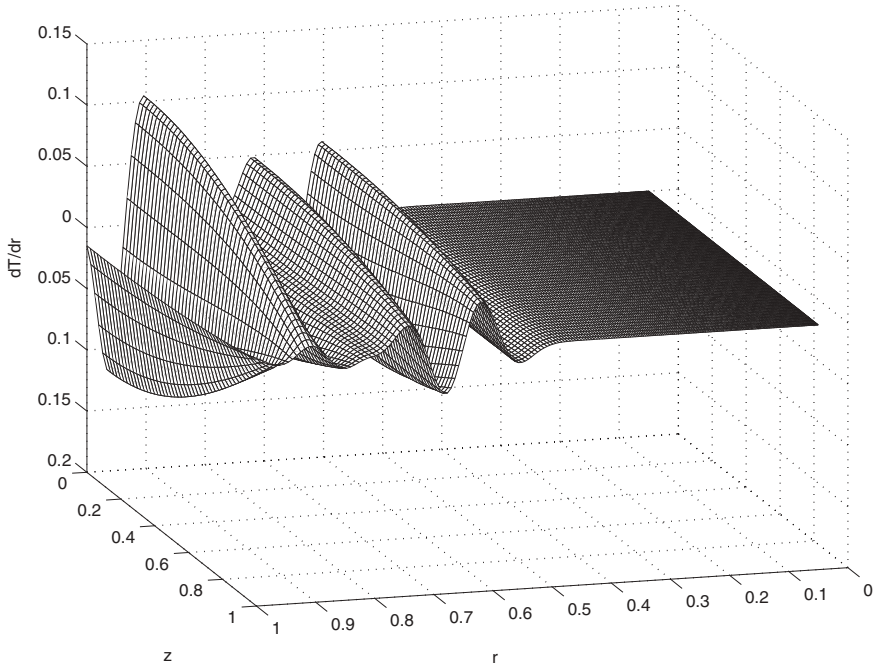


FIG. 7. *The spatial distribution of the temperature gradient,* $\frac{\partial T}{\partial r}$*, for the same parameter values as in Figure* 5(a)*. The temperature gradient is of the same magnitude as the radial velocity at the sidewall. Note that relatively large temperature gradients persist into the flow field, not just at the wall.*

different convective terms in (88) and (89), particularly the nonlinear properties of the latter.

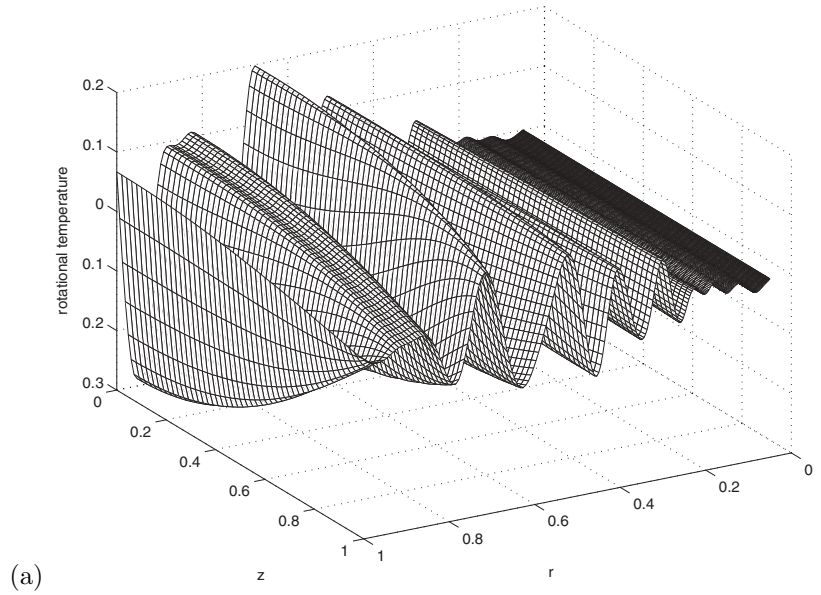
The radial spatial wave variation for $\widehat{T}_0$ differs from that of $\widehat{V}_{z0}$, as shown at $t = 20$ and $z = 0.5$ in Figure 9. In part, this occurs because the acoustic solutions, $\overline{T}_0$ and $\overline{V}_{z0}$, the drivers for the strong rotational solutions, are out of phase, as can be seen from (67) and (68). Here again it can be seen that there is a difference between the damping of spatial oscillations in $\widehat{T}_0$ and those in the $\widehat{V}_{z0}$-waves as the centerline is approached.

Figures 10 and 11 describe the spatial variation of the complete temperature T for the parameter values used in Figure 5(a) when the injection boundary condition is given by

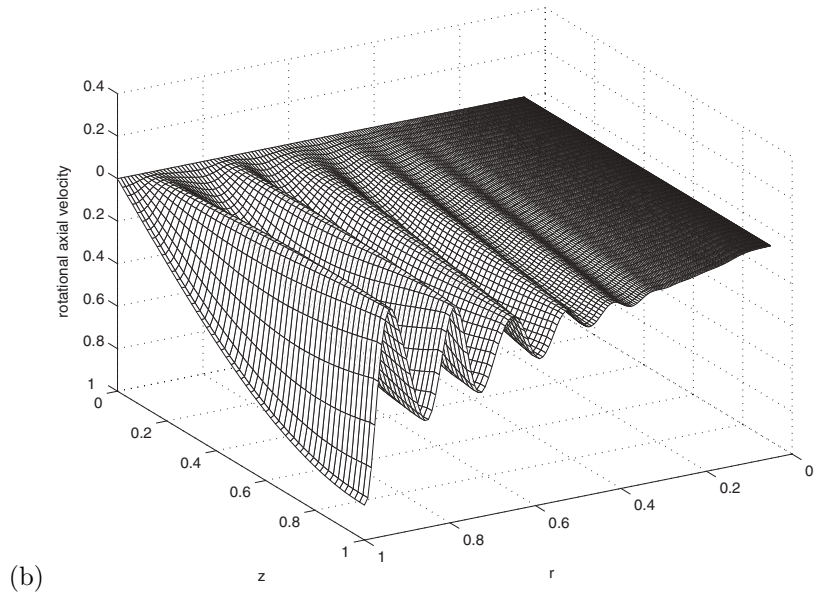
$$V_r = -1 - 0.2\cos^2\left(\frac{3\pi z}{2}\right)(1 - \cos\omega t) \qquad \text{at} \quad r = 1. \tag{102}$$

This injection distribution is employed to explore the impact of higher axial wave number variations on the internal flow dynamics. The z-dependence of this boundary condition is shown on the right-hand side of Figure 11, where positive velocity values are given to the right on the horizontal axis. A new acoustic solution was derived to satisfy (102).

The leading-order complete temperature solution is shown in Figure 10. The parameter values correspond to those in Figure 5(a). Three waves have propagated into the cylinder just as in Figure 5(a). Now, however, the waves contain a little more



(a)



(b)

FIG. 8. (a) *Variation of the strong rotational temperature* $\widehat{T}_0$ *with* r *and* z *at time* $t = 60$ *for the parameters used in Figure* 5. *The use of a radial variable integral transformation in* (31) *removes restrictions on the maximum time for which the solution can be found.* (b) *Variation of* $\hat{V}_{z0}$ *with* r *and* z, *the rotational axial velocity, for* $t = 60$.

axial variation, resulting from the higher wave number axial variation in the sidewall mass addition. The rate of radial wave propagation is axially dependent and directly related to variation in the mass addition, as implied by (76) and (81).

A comparison of the contour plots in Figures 6 and 11 is useful for observing the altered shape of the temperature waves for a given set of parameter values. The first

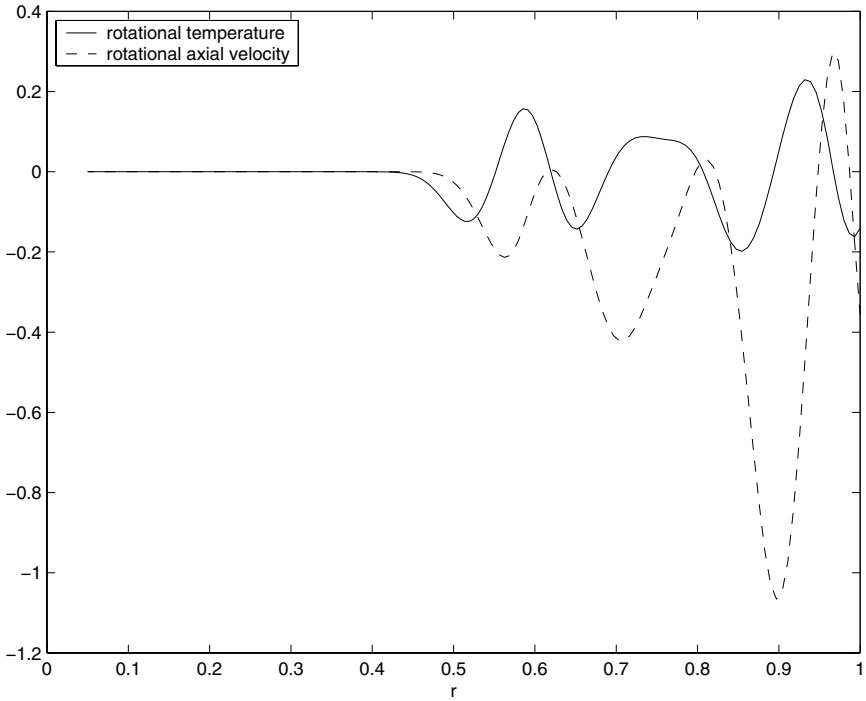


FIG. 9. *Comparison of the radial variation of* $\widehat{T}_0$ *and* $\widehat{V}_{z0}$ *at* $z = 0.5$ *for* $M = 0.02$, $\delta = 20$, $Re = 3 \times 10^5$, $\omega = 1$, $Pr = 1$, *and* $\gamma = 1.4$ *at* $t = 20$. *The radial locations of the peaks differ because the acoustic solutions,* $\overline{T}_0$ *and* $\overline{V}_{z0}$, *are qualitatively different, and the linear convective term in* (88) *differs from the nonlinear effect in* $\widehat{V}_{z0}$ ((see (89)).

waves adjacent to the sidewall have very little z-variation. However, the second and third waves in Figure 11 have crests with radial locations that vary considerably in the z-direction relative to those in Figure 6.

Figure 12 shows the location and shape of the .001-contour of the rotational temperature for 12 different times up to $t = 60$ reading from right to left. As the front propagates toward the centerline, regions of relatively large curvature are convected downstream and swept out of the cylinder. Downstream convection can be followed by noting that the local minimum in the front location, marked with an X near $z = 0.3$ at $t = 5$, moves to a larger axial location at each new time value. It moves out of the cylinder by the time $t = 30$ and leaves behind a front profile that has very little axial dependence.

The lack of axial variation near the centerline is similar to a result of Staab and Kassoy [14], who find that azimuthal variations along the sidewall driven by nonaxisymmetric wall injection are damped out near the centerline. Both the model in Staab and Kassoy [14] and that used here are without azimuthal and axial diffusion terms, respectively. These results show that axial convective terms are sufficient to produce these phenomena.

12. Summary of results and conclusions. A multiple-scale asymptotic analysis has provided the mathematical foundation for investigating the thermal properties of the compressible flow in a large aspect ratio cylinder with unsteady, axially

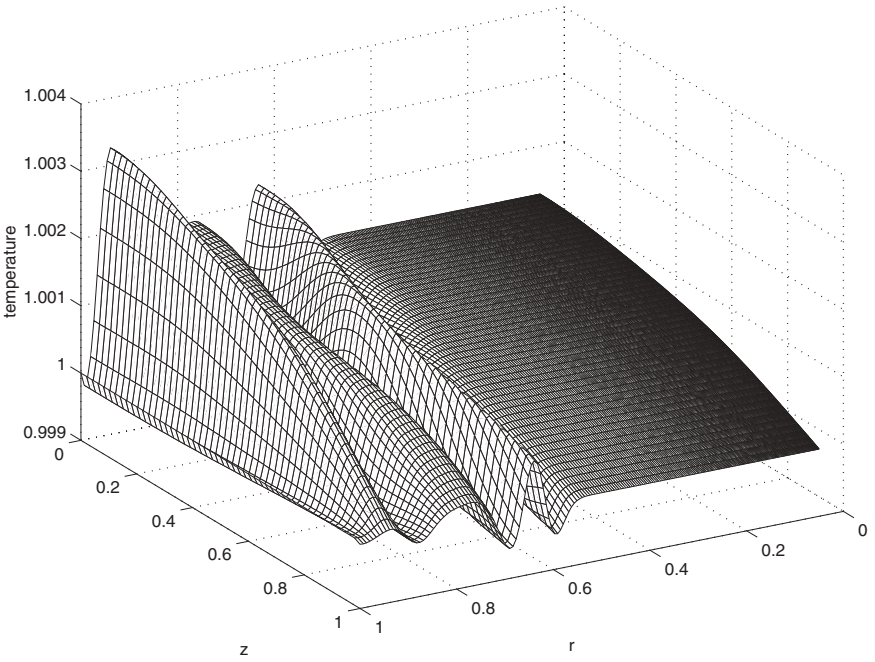


FIG. 10. *The* (r, z) *variation of* $T \sim 1 + M(\widehat{T}_0 + \overline{T}_0) + O(M^2)$ *for* $M = 0.02$, $\delta = 20$, $Re = 3 \times 10^5$, $\omega = 1$, $Pr = 1$, *and* $\gamma = 1.4$ *at* $t = 20$ *using the new radial sidewall velocity,* $\widetilde{V}_{rw}(z) = 0.2\left(\cos^2(\frac{3\pi z}{2})\right)$. *The curvature changes of the wave crests in the z-direction arise from the higher wave number axial variation of the sidewall mass addition in* (102).

distributed sidewall mass addition. The latter drives significant planar acoustic disturbances in the low Mach number and high Reynolds number fluid flow. Although the temperature variations in the flow field are only $O(M)$, one can find larger-than-expected $O(1)$ temperature gradients on the sidewall because the relevant radial length scale for gradients is $O(M)$ rather than $O(1)$. It follows that the transient heat transfer on the sidewall is rather more intense than one might expect and is controlled entirely by the injection-induced acoustic disturbances.

The current work extends the earlier study of Staab et al. [12], who predicted velocity and vorticity solutions for a limited range of time. This limitation is overcome by employing an integral transformation for the short-length–scale radial variable. When this transform is used in the asymptotic analysis, the time-scale for solution validity is extended considerably compared to the limitations found in Staab et al. [12]. As a result, long-time velocity and temperature results have been included here that were not possible previously.

Several interesting thermal phenomena have been observed in the cylinder.

1. The temperature gradient on the sidewall is $O(1)$, even though the fluctuations in the temperature are $O(M)$. This results in an unexpectedly large heat transfer on the sidewall.
2. The decay rate of the strongly rotational axial velocity waves is more pronounced and systematic than that of the corresponding rotational temperature waves. Since the results are for $Pr = 1$, the characteristic difference

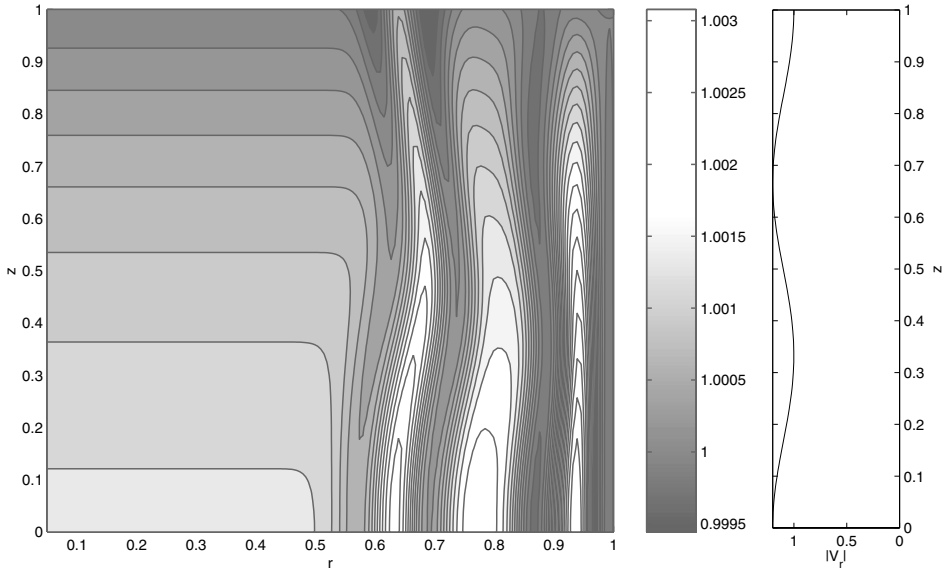


FIG. 11. *The contour plot of T for the parameter values in Figure* 10*, showing the curvature in the crest locations of the waves. The crest shapes become more varied in the z-direction as one moves away from the wall. The z-dependence of the sidewall mass addition in* (102) *is shown on the right-hand side. Axial variations in the sidewall mass addition create the curvature in the temperature contours.*

in structure must be due to the impact of the differing axial convection process.

3. Increasing the wave number of the axial variation in mass addition causes axial variations in the front separating the irrotational and "rotational" temperature distributions when the front is not too close to the center line. However, the axial structure is lost when the front approaches the center line. Since axial diffusion is not present in the model, the result can be explained by the axial convection that transports fluid downstream and out of the cylinder.

The transient mass addition on the boundary of the cylinder causes work to be done on the system as well as producing acoustic disturbances. The latter interact with the fluid at the injection surface to create vorticity. As a result, the energy of the fluid can be partitioned into acoustic and rotational components. A careful analysis of the asymptotic results shows that the $O(M)$ acoustic and "rotational" temperature disturbances make the largest contributions to the total energy transient. Kinetic energy effects appear only at $O(M^2)$. The presence of thermal energy in the "rotational" disturbance, and of course kinetic energy associated with the rotational velocity disturbance, suggest that a purely acoustical analysis cannot accurately describe the partitioning of the energy in the flow.

The wall injection boundary condition in (10) simulates the transverse speed of hot gaseous products of reaction exiting the very thin combustion zone, typically on the order of millimeters (O(mm.)) adjacent to the burning rocket motor propellant. The speed is known to be compatible with the hard blowing approximation [15] used in section 2. The axial gas speed at the edge of the combustion zone is likely to be small compared to the transverse value because the zone thickness is small in absolute

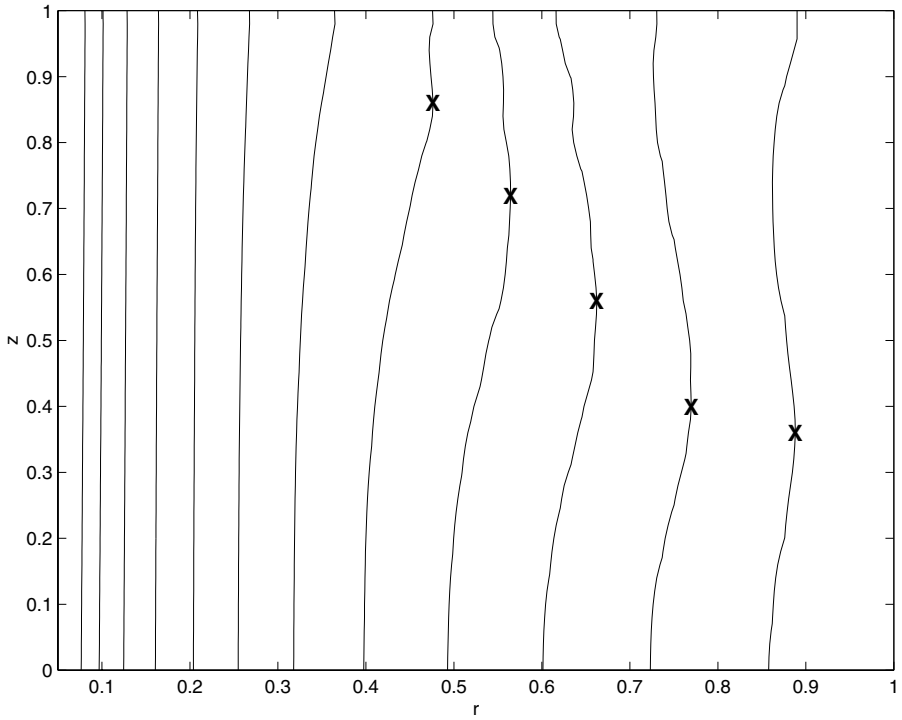


FIG. 12. *Front locations for* $\widehat{T}_0 = 0.001$ *at* $t = 5, 10, 15, 20, 23, \ldots, 60$. *The front near* $r = 0.85$ *corresponds to* $t = 5$, *and each successive front is to the left of the one before it. The X's mark a local minimum in the front which gets convected downstream.*

terms and the local transient, axial pressure gradient is $O(M)$. It follows that the no-slip condition imposed by (7) provides a reasonable first approximation to the model.

The isothermal boundary condition in (11) represents the temperature of the hot gaseous products, typically 2500K–3000K in dimensional value. The axial variation of the combustion zone flame temperature is likely to be small compared to the flame temperature because it is controlled primarily by gas phase exothermic chemical kinetics. Hence, the use of a spatially uniform temperature in (11) is a reasonable modeling approximation. In this respect the isothermal injection model provides a reasonable approximation of the physics occurring at the edge of the combustion zone.

The boundary condition on the radial speed in (7) is not satisfied by the solution provided. In fact a thin viscous boundary is required to provide a transition between the solutions given here and the endwall of the cylinder. This issue, discussed in [12], will not affect the solutions described here.

The model presented here does not include the effect of turbulent transition in the downstream portion of the cylinder, arising from hydrodynamic instability. Vorticity generated by turbulent flow will add to that arising from the deterministic mechanism explored in the current work. It would be worthwhile to ascertain the relative intensity of both components of the vorticity field. It should also be recognized that transient surface pressure fluctuations associated with the turbulent field will initiate additional vorticity via the deterministic process.

Appendix: Derivation of the higher-order energy equation. Derivations of (51) and (52) are given here to describe the algebraic manipulations. The complete set of $O(M)$ terms for the $\widetilde{T}_0$-equation is

$$\frac{\partial \widetilde{T}_0}{\partial t} + \frac{V_{r0s} + \widetilde{V}_{r0}}{s}\,\frac{\partial \widetilde{T}_0}{\partial \sigma_2} = -(\gamma-1)\left(\frac{\widetilde{V}_{r0}}{r} + \frac{1}{s}\,\frac{\partial \widetilde{V}_{r0}}{\partial \sigma_1} + \frac{1}{s}\,\frac{\partial \widetilde{V}_{r1}}{\partial \sigma_2} + \frac{\partial \widetilde{V}_{r0}}{\partial z}\right). \tag{A1}$$

Equation (43) can be used to reduce (A1) to

$$\frac{\mathcal{D}\widetilde{T}_0}{\mathcal{D}t} = (\gamma-1)\,\frac{\mathcal{D}\widetilde{\rho}_0}{\mathcal{D}t}, \tag{A2}$$

which can be further simplified by introducing (53),

$$\frac{\mathcal{D}\widetilde{T}_0}{\mathcal{D}t} = \frac{\gamma-1}{\gamma}\,\frac{\partial \widetilde{P}_0}{\partial t}. \tag{A3}$$

The $O(M^2)$ energy equation for $\widetilde{T}_1$ is derived in much the same way. The complete set of $O(M^2)$ terms is

$$\begin{aligned}
&\frac{\partial \widetilde{T}_1}{\partial t} + \frac{V_{r0s} + \widetilde{V}_{r0}}{s}\,\frac{\partial \widetilde{T}_1}{\partial \sigma_2} + \frac{V_{r0s} + \widetilde{V}_{r0}}{s}\,\frac{\partial \widetilde{T}_0}{\partial \sigma_1} + \frac{\widetilde{V}_{r1}}{s}\,\frac{\partial \widetilde{T}_0}{\partial \sigma_2} + (V_{z0s} + \widetilde{V}_{z0})\,\frac{\partial \widetilde{T}_0}{\partial z} \\
&= \frac{Ck^2}{Pr}\,\frac{1}{s^2}\,\frac{\partial^2 \widetilde{T}_0}{\partial \sigma_2^2} - (\gamma-1)\left(\frac{\widetilde{V}_{r1}}{r} + \frac{1}{s}\,\frac{\partial \widetilde{V}_{r1}}{\partial \sigma_1} + \frac{\partial \widetilde{V}_{r2}}{s} + \frac{\partial \widetilde{V}_{z1}}{\partial z}\right) \\
&-(\gamma-1)\widetilde{P}_0\left(\frac{\widetilde{V}_{r0}}{r} + \frac{1}{s}\,\frac{\partial \widetilde{V}_{r0}}{\partial \sigma_1} + \frac{1}{s}\,\frac{\partial \widetilde{V}_{r1}}{\partial \sigma_2} + \frac{\partial \widetilde{V}_{z0}}{\partial z}\right).
\end{aligned} \tag{A4}$$

Then (43) and (44) can be used in (A4) to find

$$\begin{aligned}
&\frac{\partial \widetilde{T}_1}{\partial t} + \frac{V_{r0s} + \widetilde{V}_{r0}}{s}\,\frac{\partial \widetilde{T}_1}{\partial \sigma_2} + \frac{V_{r0s} + \widetilde{V}_{r0}}{s}\,\frac{\partial \widetilde{T}_0}{\partial \sigma_1} + \frac{\widetilde{V}_{r1}}{s}\,\frac{\partial \widetilde{T}_0}{\partial \sigma_2} + (V_{z0s} + \widetilde{V}_{z0})\,\frac{\partial \widetilde{T}_0}{\partial z} \\
&= (\gamma-1)\left(\frac{\mathcal{D}\widetilde{\rho}_1}{\mathcal{D}t} - \widetilde{\rho}_0\,\frac{\mathcal{D}\widetilde{\rho}_0}{\mathcal{D}t} + \widetilde{P}_0\,\frac{\mathcal{D}\widetilde{\rho}_0}{\mathcal{D}t} + \frac{V_{r0s} + \widetilde{V}_{r0}}{s}\,\frac{\partial \widetilde{\rho}_0}{\partial \sigma_1}\right. \\
&\qquad \left. + \frac{\widetilde{V}_{r1}}{s}\,\frac{\partial \widetilde{\rho}_0}{\partial \sigma_2} + (V_{z0s} + \widetilde{V}_{z0})\,\frac{\partial \widetilde{\rho}_0}{\partial z}\right) \\
&+ \frac{Ck^2}{Pr}\,\frac{1}{s^2}\,\frac{\partial^2 \widetilde{T}_0}{\partial \sigma_2^2}.
\end{aligned} \tag{A5}$$

Next, (53) and (54) are used to replace $\widetilde{\rho}_0$ and $\widetilde{\rho}_1$, respectively:

$$\begin{aligned}
&\frac{\partial \widetilde{T}_1}{\partial t} + \frac{V_{r0s} + \widetilde{V}_{r0}}{s}\,\frac{\partial \widetilde{T}_1}{\partial \sigma_2} + \frac{V_{r0s} + \widetilde{V}_{r0}}{s}\,\frac{\partial \widetilde{T}_0}{\partial \sigma_1} + \frac{\widetilde{V}_{r1}}{s}\,\frac{\partial \widetilde{T}_0}{\partial \sigma_2} + (V_{z0s} + \widetilde{V}_{z0})\,\frac{\partial \widetilde{T}_0}{\partial z} \\
&= (\gamma-1)\left(\frac{\partial \widetilde{P}_1}{\partial t} - \frac{\mathcal{D}\widetilde{T}_1}{\mathcal{D}t} - \frac{\mathcal{D}\widetilde{\rho}_0\widetilde{T}_0}{\mathcal{D}t} + \widetilde{T}_0\,\frac{\mathcal{D}\widetilde{\rho}_0}{\mathcal{D}t}\right) \\
&+(\gamma-1)\left(-\frac{V_{r0s} + \widetilde{V}_{r0}}{s}\frac{\partial \widetilde{T}_0}{\partial \sigma_1} - \frac{\widetilde{V}_{r1}}{s}\frac{\partial \widetilde{T}_0}{\partial \sigma_2} - (V_{z0s} + \widetilde{V}_{z0})\frac{\partial \widetilde{T}_0}{\partial z} + (V_{z0s} + \widetilde{V}_{z0})\frac{\partial \widetilde{P}_0}{\partial z}\right) \\
&+ \frac{Ck^2}{Pr}\,\frac{1}{s^2}\,\frac{\partial^2 \widetilde{T}_0}{\partial \sigma_2^2}.
\end{aligned} \tag{A6}$$

Further simplification leads to

$$\frac{\mathcal{D}\widetilde{T}_1}{\mathcal{D}t} = -\left(\frac{V_{r0s}+\widetilde{V}_{r0}}{s}\,\frac{\partial\widetilde{T}_0}{\partial\sigma_1}+\frac{\widetilde{V}_{r1}}{s}\,\frac{\partial\widetilde{T}_0}{\partial\sigma_2}+(V_{z0s}+\widetilde{V}_{z0})\frac{\partial\widetilde{T}_0}{\partial z}\right) + \frac{\gamma-1}{\gamma}\left(\frac{\partial\widetilde{P}_1}{\partial t}+(V_{z0s}+\widetilde{V}_{z0})\frac{\partial\widetilde{P}_0}{\partial z}\right)+\frac{Ck^2}{\gamma\ Pr}\,\frac{1}{s^2}\,\frac{\partial^2\widetilde{T}_0}{\partial\sigma_2^2} - \frac{(\gamma-1)^2}{\gamma^2}\left(\widetilde{P}_0\frac{\partial\widetilde{P}_0}{\partial t}-\widetilde{T}_0\frac{\partial\widetilde{P}_0}{\partial t}\right), \tag{A7}$$

which appears as (52) in section 5.1.

Acknowledgments. The authors acknowledge the patience of Mitat Birkan. In addition, P. L. Staab and M. J. Rempe appreciate the research assistantships provided by the AFOSR grants. Finally, the suggestions of two referees have been helpful in providing greater clarity to the manuscript.

REFERENCES

[1] G. A. Flandro, *Solid propellant acoustic admittance corrections*, J. Sound Vibration, 36 (1974), pp. 297–312.

[2] G. A. Flandro, *Effects of vorticity on rocket combustion stability*, J. Prop. Power, 11 (1995), pp. 607–625.

[3] J. Majdalani and S. W. Rienstra, *Two asymptotic forms of the rotational solution for wave propagation inside viscous channels with transpiring walls*, Quart. J. Mech. Appl. Math., 55 (2002), pp. 141–162.

[4] F. E. C. Culick and V. Yang, *Prediction of the stability of unsteady motions in solid-propellant rocket motors*, in Nonsteady Burning and Combustion Stability of Solid Propellants, AIAA Progress in Astronautics and Aeronautics 143, L. De Luca, E. W. Price, and M. Summerfeld, eds., AIAA, Washington, DC, 1992, pp. 719–779.

[5] G. Casalis, G. Avalon, and J. P. Pineau, *Spatial instability of planar channel flow with fluid injection through porous walls*, Phys. Fluids, 10 (1998), pp. 2558–2568.

[6] P. Venugopal, F. M. Najjar, and R. D. Moser, *DNS and LES computations of model solid rocket motors*, AIAA 2000-3571, in Proceedings of the 36th AIAA/ASME/SAE/ASEE Joint Propulsion Conference, Huntsville, AL, 2000.

[7] P. Venugopal, F. M. Najjar, and R. D. Moser, *Numerical simulations of model solid rocket motor flows*, AIAA 2001-3950, in Proceedings of the 37th AIAA/ASME/SAE/ASEE Joint Propulsion Conference, Salt Lake City, UT, 2001.

[8] Q. Zhao, *Nonlinear Acoustic Processes in Solid Rocket Engines*, Ph.D. thesis, Department of Mechanical Engineering, University of Colorado, Boulder, CO, 1994.

[9] Q. Zhao, P. L. Staab, D. R. Kassoy, and K. Kirkkopru, *Acoustically generated vorticity in an internal flow*, J. Fluid Mech., 413 (1999), pp. 247–285.

[10] K. Kirkkopru, D. R. Kassoy, and Q. Zhao, *Unsteady vorticity generation and evolution in a model of a solid rocket motor*, J. Prop. Power, 12 (1996), pp. 646–654.

[11] P. L. Staab and D. R. Kassoy, *Three dimensional, unsteady, acoustic-shear flow dynamics in a cylinder with sidewall mass addition*, Phys. Fluids, 9 (1997), pp. 3753–3763.

[12] P. L. Staab, Q. Zhao, D. R. Kassoy, and K. Kirkkopru, *Co-existing acoustic-rotational flow in a cylinder with axisymmetric sidewall mass addition*, Phys. Fluids, 11 (1999), pp. 2935–2951.

[13] K. Kirkkopru, D. R. Kassoy, Q. Zhao, and P. L. Staab, *Acoustically generated unsteady vorticity field in a long narrow cylinder with sidewall injection*, J. Engrg. Math., 42 (2002), pp. 65–90.

[14] P. L. Staab and D. R. Kassoy, *Three-dimensional flow in a cylinder with sidewall mass addition*, Phys. Fluids, 14 (2002), pp. 3141–3159.

[15] J. D. Cole and J. Aroesty, *The blowhard problem: Inviscid flows with surface injection*, Intl. J. Heat Mass Trans., 11 (1968), pp. 1167–1183.

[16] F. Vuillot and G. Avalon, *Acoustic boundary layers in solid propellant rocket motors using Navier-Stokes equations*, J. Prop. Power, 7 (1991), pp. 231–239.

[17] F. Vuillot, *Vortex-shedding phenomena in solid rocket motors*, J. Prop. Power, 11 (1995), pp. 626–639.

[18] N. Lupoglazoff and F. Vuillot, *Numerical simulations of parietal vortex-shedding in a cold flow set-up*, AIAA 98-3220, in Proceedings of the 34th AIAA/ASME/SAE/ASEE Joint Propulsion Conference, Cleveland, OH, 1998.

[19] A. K. Hegab, *A Study of Acoustic Phenomena in Solid Rocket Engines*, Ph.D. thesis, Department of Mechanical Power, Menoufia University, Menoufia, Egypt, 1998.

[20] F. E. C. Culick, *Rotational axisymmetric mean flow and damping of acoustic waves in a solid propellant*, AIAA Journal, 4 (1966), pp. 1462–1463.

[21] G. I. Taylor, *Fluid flow in regions bounded by porous surfaces*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 234 (1956), pp. 456–475.

[22] Q. Zhao and D. R. Kassoy, *The generation and evolution of unsteady vorticity in a model of a solid rocket engine chamber*, AIAA 94-0779, in Proceedings of the 32nd Aerospace Sciences Meeting, Reno, NV, 1994.

[23] J. Majdalani and G. A. Flandro, *The oscillatory pipe flow with arbitrary wall injection*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 458 (2002), pp. 1621–1651.

[24] J. Majdalani and W. Van Moorhem, *Multiple-scales solution to the acoustic boundary layer in solid rocket motors*, J. Prop. Power, 13 (1997), pp. 186–193.

[25] J. Majdalani and W. Van Moorhem, *Improved time-dependent flow field solution for solid rocket motors*, AIAA J., 36 (1998), pp. 241–248.

SIAM J. APPL. MATH.
Vol. 65, No. 2, pp. 618–643

ON THE TURING PATTERNS IN ONE-DIMENSIONAL GRADIENT/SKEW-GRADIENT DISSIPATIVE SYSTEMS*

MASATAKA KUWAMURA†

Abstract. In this article, fundamental properties concerning the Turing patterns are considered in one-dimensional dissipative systems with gradient/skew-gradient structure introduced in [M. Kuwamura and E. Yanagida, *Phys. D*, 175 (2003), pp. 185–195]. It is a natural extension of free energy, which covers reaction-diffusion systems of activator-inhibitor type. The theory based on this concept provides a new perspective on a fundamental problem of what unique Turing pattern is to be selected among many.

Key words. (skew) free energy, Turing pattern, pattern selection

AMS subject classifications. 35K57, 35B32, 92C15

DOI. 10.1137/S0036139903424898

1. Introduction.

There are many self-organized spatial patterns appearing from uniform steady state as various organs with complex structures are formed and developed from a simple cell. The problem of clarifying the essential mechanism of such a pattern formation process has attracted much attention. For this problem, Rashevsky [33] and Turing [35] suggested that, under certain conditions, chemicals can react and diffuse in such a way as to produce steady state heterogeneous spatial patterns of chemical and morphogen concentration. This is, currently, known as the Turing instability mechanism, which provides a fundamental and universal concept for the study of pattern formation in dissipative systems; we refer to standard texts Nicolis and Prigogine [27], Haken [15], Manneville [22], Mori and Kuramoto [24], Murray [25], and the references therein. This theoretical concept was also demonstrated by real experiments and careful observations by Ouyang and Swinney [29], Kondo and Asai [18], De Kepper, Perraud, Rudovics, and Dulos [8], Yamaguchi [36], and so on. The progress of study of the Turing instability gives us a useful viewpoint to understand pattern formation in various phenomena; however, some natural questions still remain. A typical question is this: when one observes a pattern formation process due to the Turing instability, why does a particular pattern often selected though various patterns appear? This problem is known as pattern selection, one of the main subjects in various fields of sciences, and there are some approaches to study the pattern selection problem; we refer to Cross and Hohenberg [7] and Nishiura [28]. One useful approach is to find a free energy (potential system or variational principle), which determines direction of time evolution of systems. It is quite natural, however, and there are few examples to apply this approach explicitly. The purpose of this paper is to study fundamental properties of spatially periodic patterns induced by the Turing instability under the one-dimensional gradient/skew-gradient dissipative structure introduced in Kuwamura and Yanagida [20]. It is a natural extension of the notion of free energy, which covers reaction-diffusion systems of activator-inhibitor type such as modified FitzHugh–Nagumo systems and Gierer–Meinhardt systems. Our results

*Received by the editors March 24, 2003; accepted for publication (in revised form) January 27, 2004; published electronically January 5, 2005. This work was supported in part by grant-in-aid for Scientific Research 13740066 and 15654018 from the Japan Society for the Promotion of Science.
http://www.siam.org/journals/siap/65-2/42489.html
†Faculty of Human Development, Kobe University, Tsurukabuto 3-11, Nada-ku, Kobe 657-8501, Japan (kuwamura@main.h.kobe-u.ac.jp).

clarify universal properties of the Turing patterns in dissipative systems, and become a first step to study a fundamental problem of what unique Turing pattern is to be selected among many. The organization of this paper is as follows. In the next section, we introduce the concept of gradient/skew-gradient dissipative structure following [20]. In section 3, we give a necessary condition for existence of nontrivial spatially periodic patterns due to the Turing instability. In section 4, we give a sufficient condition for instability of spatially periodic patterns in terms of convexity of (skew) free energy per unit length of periodic patterns in its wavenumber; this is known as the Eckhaus instability criterion. In section 5, we consider a quantity introduced in the previous section, which plays a crucial role in determining the instability of the Turing patterns. It is shown that this quantity is closely related to the validity of amplitude equation which describes dynamics in a sufficiently small neighborhood of the Turing bifurcation point. In section 6, we consider structures of the (skew) free energy per unit length of the Turing patterns. Under a certain nondegeneracy condition, uniqueness of the extremum of (skew) free energy is shown. Moreover, we give a simple formula for computing the wavenumber corresponding to the unique extremum near the Turing bifurcation point. In section 7, we apply our results to concrete examples which are helpful to understand usefulness of the results obtained in previous sections. In section 8, combining the analytical results with numerical experiments, we study what Turing pattern is to be selected near a bifurcation point. The results suggest that the wavenumber corresponding to the unique extremum of (skew) free energy gives an upper bound for the wavenumber of the selected Turing pattern. Section 9 is devoted to a summary of this paper.

2. Gradient/skew-gradient dissipative structure. In this section, we introduce the concept of *gradient/skew-gradient dissipative structure* following [20]. Let us consider an n-component system on $\mathbf{R}$

$$Tu_t = Du_{xx} + f(u), \tag{2.1}$$

where $u(x,t) = (u_1, u_2, \dots, u_n)^T \in \mathbf{R}^n$. We assume that T is a nonnegative diagonal matrix and D is a regular matrix such that (2.1) is well posed in an appropriate sense. As for a nonlinear term, we assume that

$$f(u) = Q \bigtriangledown_u F(u), \tag{2.2}$$

where Q is a symmetric matrix with $Q^2 = I_n$, and that $F = F(u) : \mathbf{R}^n \to \mathbf{R}$ is a smooth function. In addition, we assume that D satisfies the condition

$$D^T Q = QD \tag{2.3}$$

which guarantees that QD is a nondegenerate symmetric matrix. We notice that the Jacobian matrix f_u of f satisfies

$$f_u(u)^T Q = Q f_u(u). \tag{2.4}$$

Under these assumptions, we introduce an energy-like functional

$$\mathcal{E}[u] = \int \left\{ \frac{1}{2} \langle Du_x, Qu_x \rangle - F(u) \right\} dx, \tag{2.5}$$

where $\langle\ ,\ \rangle$ stands for a usual inner product on $\mathbf{R}^n$. In fact, we can easily (formally) check

$$\frac{d}{dt}\mathcal{E}[u(x,t)] = -\int \langle u_t, QTu_t \rangle \, dx.$$

DEFINITION 2.1. *The system* (2.1) *is said to have gradient structure when* QT *is nonnegative symmetric and skew-gradient structure otherwise.*

Although (2.5) is not monotone decreasing when QT is not nonnegative definite, according to Definition 2.1, we naturally think of the functional (2.5) as an extended kind of free energy. This framework covers not only usual gradient (potential) systems such as the real Ginzburg–Landau equation but also systems without (in the usual sense) potential such as reaction-diffusion systems of activator-inhibitor type.

The equation for stationary solutions of (2.1) admits structures of Hamiltonian dynamical systems (e.g., Arnold [2], Goldstein [13]). In fact,

$$Du_{xx} + f(u) = 0 \tag{2.6}$$

is rewritten in the canonical form

$$JZ_x = \frac{\partial H(Z)}{\partial Z}, \tag{2.7}$$

where $Z = (u, u_x)^T$,

$$J = \begin{pmatrix} 0 & -QD \\ QD & 0 \end{pmatrix}$$

is a skew-symmetric matrix by virtue of (2.3), and $H(Z)$ is a first integral (Hamiltonian) given by

$$H(Z) = H(u, u_x) := \frac{1}{2}\langle Du_x, Qu_x\rangle + F(u).$$

We give typical examples of dissipative systems with gradient/skew-gradient structure.

Real Ginzburg–Landau equation:

$$u_t = u_{xx} + u(\mu - u^2 - v^2), \quad v_t = v_{xx} + v(\mu - u^2 - v^2),$$
$$T = D = Q = I_2, \qquad F = \mu(u^2 + v^2)/2 - (u^2 + v^2)^2/4.$$

Modified FitzHugh–Nagumo systems:

$$\tau_1 u_t = d_1 u_{xx} + \alpha u - u^3 - v, \quad \tau_2 v_t = d_2 v_{xx} + u - \gamma v,$$
$$T = \begin{pmatrix} \tau_1 & 0 \\ 0 & \tau_2 \end{pmatrix}, \quad D = \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$
$$F = F(u, v) = \frac{1}{2}\alpha u^2 - \frac{1}{4}u^4 - uv + \frac{1}{2}\gamma v^2.$$

Gierer–Meinhardt system:

$$u_t = \varepsilon^2 u_{xx} - \alpha u + \frac{u^p}{v^q} + \sigma, \quad \tau v_t = dv_{xx} - v + \frac{u^r}{v^s} \quad (p + 1 = r,\ q + 1 = s),$$
$$T = \begin{pmatrix} r & 0 \\ 0 & q\tau \end{pmatrix}, \quad D = \begin{pmatrix} r\varepsilon^2 & 0 \\ 0 & qd \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$
$$F = F(u, v) = -\frac{\alpha r}{2}u^2 + \frac{q}{2}v^2 + \frac{u^r}{v^q} + r\sigma u.$$

As is verified in these two examples, in reaction-diffusion systems, diagonal entries of Q determine types of components—either activator or inhibitor type. An entry 1 indicates an activator and -1 an inhibitor. The following example is a fourth order differential equation with gradient structure.

Swift–Hohenberg equation:

$$u_t = \mu u - (1 + \partial_{xx})^2 u - u^3.$$

In fact, by putting $v = u + u_{xx}$, we find that the above equation is rewritten as (2.1) with

$$T = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

$$F = F(u, v) = \frac{\mu}{2}u^2 - \frac{1}{4}u^4 - uv + \frac{1}{2}v^2.$$

For other examples, see [20].

3. Existence of Turing patterns. In this section, we briefly review construction of Turing patterns from a standard bifurcation theory (e.g., Iooss and Joseph [17], Murray [25]) and consider a necessary condition for existence of the Turing patterns.

A strategy was initiated by Turing [35] to study self-organized spatial patterns emerging from spatially homogeneous steady states and may be described as follows, in the framework of bifurcation theory: Let $\bar{u}$ be a spatially uniform steady state of (2.1), which is an onset of bifurcating patterns. We consider a situation in which $\bar{u}$ loses its stability and gives rise to the appearance of spatially periodic stationary solutions. Therefore, we begin with a linear stability analysis for $\bar{u}$. Let us consider the linearized equation of (2.1) at $u = \bar{u}$

$$Tw_t = Lw, \tag{3.1}$$

where

$$L = D\partial_x^2 + B \tag{3.2}$$

and $B = f_u(\bar{u})$. Let us choose a bifurcation parameter μ in the linear operator L, i.e., D or B. As usual, we are looking for solutions of the following form:

$$w = \Psi_k \exp(\lambda t + ikx), \qquad \Psi_k \in \mathbf{C}^n. \tag{3.3}$$

That is to say, we look for solutions that have a temporal growth rate $\exp(\mathrm{Re}(\lambda)t)$ for perturbations with wavenumber k. Substituting (3.3) into (3.1), we have a system of linear equations in Ψ_k

$$(\lambda T + k^2 D - B)\Psi_k = 0. \tag{3.4}$$

According to a standard bifurcation theory, in order for a stationary solution to bifurcate from $\bar{u}$ at $\mu = \mu_k$, it is necessary that (3.4) possesses a nontrivial solution for $\lambda = 0$. Hence it follows from

$$\det(k^2 D - B) = 0 \tag{3.5}$$

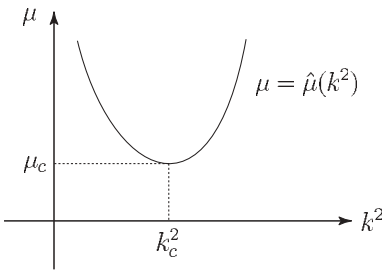


FIG. 1. *The graph of* $\mu = \hat{\mu}(k^2)$. *The bifurcating patterns appear for* $\mu > \hat{\mu}(k^2)$.

that μ_k is a function of k^2, denoted by

$$\mu_k = \hat{\mu}(k^2). \tag{3.6}$$

The bifurcation point μ_c is determined by

$$\frac{\partial}{\partial\theta}\,\hat{\mu}(\theta)|_{\theta=k_c^2} = 0, \quad \mu_c = \hat{\mu}(k_c^2), \tag{3.7}$$

where k_c is called the *critical wavenumber*. In what follows, we suppose that bifurcating patterns appear for $\mu > \mu_c$ without loss of generality (Figure 1). That is, in addition to (3.7), we assume

$$\frac{\partial}{\partial\theta^2}\,\hat{\mu}(\theta)|_{\theta=k_c^2} > 0,$$

and

$$\mathrm{Re}\lambda = \mathrm{Re}\lambda(k^2;\mu) < 0 \tag{3.8}$$

for $\mu < \mu_c$ and arbitrary k, where $\lambda = \lambda(k^2;\mu)$ is the dispersion relation defined by $\det(\lambda T + k^2 D - B) = 0$. Condition (3.8) means that the spatially homogeneous steady state $\bar{u}$ must be stable before bifurcation occurs.

Under suitable conditions, we can construct small bifurcating stationary solutions of (2.1) with wavenumber k close to k_c near the bifurcation point μ_c. From the above arguments, we expect that they are given by

$$\phi(x;k,\mu) = ae^{ikx}\Psi_k + c.c. + h.o.t., \tag{3.9}$$

where *c.c.* and *h.o.t.* denote the complex conjugate and higher order terms with respect to a, respectively. Moreover, $\Psi_k \in \mathbf{C}^n$ is defined by

$$(k^2 D - B)\Psi_k = 0 \tag{3.10}$$

for $\mu = \hat{\mu}(k^2)$ and $\langle \Psi_k, \Psi_k \rangle = 1$, where $\langle\ ,\ \rangle$ denotes the usual inner product in $\mathbf{C}^n$, and $a = a(k, \mu - \hat{\mu}(k^2)) \geq 0$ is sufficiently small with $a(k,0) = 0$. In what follows, $\langle\ ,\ \rangle$ denotes the usual inner product in $\mathbf{R}^n$ or $\mathbf{C}^n$, and we do not explicitly mention this notation unless any confusion occurs. (3.9) is a family of spatially periodic stationary solutions parameterized by its wavenumber k close to k_c.

DEFINITION 3.1. (3.9) *is called the Turing patterns when* $k_c \neq 0$.

From a standard bifurcation theory, a is determined by solving the following bifurcation equation in a:

$$[e^{ikx}\Psi_k^* + c.c., Lv + N(v)] = 0,$$

where $[\ ,\ ]$ denotes the usual inner product of periodic functions, $v = ae^{ikx}\Psi_k + c.c.$, $N(v) = f(\bar{u} + v) - Bv$, and $\Psi_k^* \in \mathbf{C}^n$ corresponds to the kernel of L^*, the adjoint operator of L for $\mu = \hat{\mu}(k^2)$, i.e.,

$$(k^2\, D^T - \, B^T)\Psi_k^* = 0 \quad \text{for} \quad \mu = \hat{\mu}(k^2).$$

Under the gradient/skew-gradient structure, we can verify

$$\Psi_k^* = Q\Psi_k. \tag{3.11}$$

In fact, noting (2.3), (2.4), and $B = f_u(\bar{u})$, it follows from (3.10) that

$$(k^2\, D^T - \, B^T)Q\Psi_k = Q(k^2 D - B)\Psi_k = 0.$$

Thus (under suitable approximation) we can solve

$$[e^{ikx}Q\Psi_k + c.c., Lv + N(v)] = 0$$

with the aid of (3.10) and obtain $a = a(k, \mu - \hat{\mu}(k^2))$. In this paper, we do not use any information about $a(k, \mu - \hat{\mu}(k^2))$.

We now consider a condition concerning the critical wavenumber k_c. The following theorem shows that when QD is definite, Turing patterns (characterized by $k_c \neq 0$) can never be observed under the gradient/skew-gradient structure.

THEOREM 3.2. *If QD is definite, then $k_c = 0$.*

As far as reaction-diffusion systems are concerned, it is well recognized, theoretically and experimentally, that the existence of *two types* of components such as activator and inhibitor with different diffusion rates is a sufficient condition to produce Turing patterns with spatial structure ([1], [12], [25], [33], [35], [37]). On the other hand, this theorem proves that the existence of two types of components such as activator and inhibitor (cf. section 2) is a necessary condition to produce Turing patterns. In other words, Turing patterns with spatial structure cannot be produced by using only one type of reacting component corresponding to activator or inhibitor.

Proof of Theorem 3.2. We consider the case when QD is positive because a negative case can be similarly treated. Since (3.4) has a nontrivial solution for $\lambda = 0$ at the bifurcation point, we consider a condition such that $\det(\nu D - B) = 0$ holds for some $\nu \geq 0$. From elementary linear algebra, there exists a symmetric regular matrix A with $A^2 = QD$ for positive QD. By using $Q^2 = I_n$, $B = f_u(\bar{u})$, and (2.2), we have

$$\begin{aligned}\det(\nu D - B) &= \det Q\ \det(\nu QD - \bigtriangledown_u^2 F(\bar{u})) \\ &= \pm \det(\nu A^2 - \bigtriangledown_u^2 F(\bar{u})) = \pm \det A\ \det(\nu I - A^{-1}\ \bigtriangledown_u^2\ F(\bar{u})A^{-1})\ \det A.\end{aligned}$$

It is easily verified that $A^{-1}\ \bigtriangledown_u^2\ F(\bar{u})A^{-1}$ is symmetric because A and $\bigtriangledown_u^2 F(\bar{u})$ are symmetric, so that all eigenvalues of $A^{-1}\ \bigtriangledown_u^2\ F(\bar{u})A^{-1}$ are real. On the other hand, there exists no wavenumber k satisfying $\det(k^2 D - B) = 0$ when uniform steady state $\bar{u}$ is stable. Hence, all eigenvalues of $A^{-1}\nabla_u^2 F(\bar{u})A^{-1}$ must be negative for a stable steady state $\bar{u}$. Thus, noting the fact that all eigenvalues of $A^{-1}\ \bigtriangledown_u^2\ F(\bar{u})A^{-1}$ are

continuous real functions in parameters, we see that critical wavenumber k_c at the bifurcation point must be equal to zero for any path in parameter space. □

From the above theorem, QD is not definite when we observe nontrivial spatially periodic patterns. Our next question is what periodic pattern is selected near the Turing bifurcation point. As a first step in considering this problem, in the next section we consider linear instability of the Turing patterns because selected patterns are not unstable.

4. Instability of Turing patterns. First, we give a general instability criterion for a family of spatially periodic stationary solutions of (2.1).

Let $u = \varphi(x;k)$ be a family of spatially periodic stationary solutions of (2.1) parameterized by its wavenumber k; that is, $\varphi(x;k)$ satisfies

$$\begin{aligned} D\varphi_{xx}(x;k) + f(\varphi(x;k)) &= 0, \\ \varphi(x;k) &= \varphi(x + l(k);k), \end{aligned} \tag{4.1}$$

where $l(k) = 2\pi/k$ denotes the minimal spatial period of $\varphi(x;k)$.

THEOREM 4.1. *$\varphi(x;k)$ is unstable if $sgn(I(k) \cdot d^2E(k)/dk^2) < 0$, where*

$$I(k) := \int_0^{l(k)} \langle T\varphi_x(x;k), Q\varphi_x(x;k)\rangle \, dx \tag{4.2}$$

and

$$E(k) := \frac{1}{l(k)} \int_0^{l(k)} \left\{ \frac{1}{2}\langle D\varphi_x(x;k), Q\varphi_x(x;k)\rangle - F(\varphi(x;k)) \right\} dx. \tag{4.3}$$

It should be noted that the above instability criterion does not require any other assumptions such as smallness and symmetry for $\varphi(x;k)$. Since $I(k) > 0$ from the definition of gradient systems, we immediately find that the following is true.

COROLLARY 4.2. *For gradient systems, $\varphi(x;k)$ is unstable if*

$$\frac{d^2E(k)}{dk^2} < 0.$$

Recalling (2.5), we regard $E(k)$ as (skew) free energy per unit length for $\varphi(x;k)$. Under the gradient/skew-gradient dissipative structure, instability of $\varphi(x;k)$ is dependent on the convex property of $E(k)$ with respect to wavenumber k.

On the other hand, $I(k)$ is related to the ODE dynamics of (2.1) as well as the spatial dynamics (2.6). $I(k)$ may be reminiscent of the (time) action functional of systems with multisymplectic structures [4]. However, this is not correct. In fact, (2.6) can be rewritten as (2.7) which expresses symplectic (Hamiltonian) structure, while (2.1) cannot be rewritten as a system with multisymplectic structure. For more details, see [4] and the references therein.

From a viewpoint of Hamiltonian dynamical systems theory, Theorem 4.1 corresponds to the well-recognized (in)stability criterion for solutions with periodic structure in terms of the convexity of the averaged Lagrangian with respect to its wavenumber. In fact, we can regard $E(k)$ as the averaged Lagrangian of $\varphi(x;k)$ because (2.6) (i.e., (4.1)) is the Euler–Lagrange equation for the Hamiltonian system (2.7) with Lagrangian

$$L(\varphi, \varphi_x) = \frac{1}{2}\langle D\varphi_x, Q\varphi_x\rangle - F(\varphi).$$

There is a wide range of literature concerning this topic; see well organized works by Bridges [4], Grillakis, Shatah, and Strauss [14], and the references therein.

Proof of Theorem 4.1. Let us consider the linearized eigenvalue problem

$$\lambda TW = DW_x x + f_u(\varphi(x;k))W.$$

We investigate the (local) dispersion relation with respect to the Fourier mode $e^{i\nu x}$ that describes the behavior of critical eigenvalues to determine sideband instability of $\varphi(x;k)$. According to [20], it is given by

$$\lambda = -D_{//}\nu^2 + h.o.t., \tag{4.4}$$

where

$$D_{//} = -\frac{l^2}{I}\frac{dH}{dl}, \tag{4.5}$$

$l = l(k) = 2\pi/k$ is the minimal spatial period, I is given by (4.2), and H is a first integral for (4.1) (see also (2.7)) given by

$$H := \frac{1}{2}\langle D\varphi_x, Q\varphi_x\rangle + F(\varphi). \tag{4.6}$$

Moreover, by [20], we have

$$\frac{dH}{dl} = \frac{1}{l}\frac{dK}{dl} \quad \text{and} \quad \frac{dE}{dl} = -\frac{K}{l^2}, \tag{4.7}$$

where E is given by (4.3), and

$$K := \int_0^l \langle D\varphi_x, Q\varphi_x\rangle\, dx.$$

Applying the chain rule of differentiation to (4.7), we have

$$\frac{dH}{dk} = \frac{1}{l}\frac{dK}{dk} \quad \text{and} \quad \frac{dE}{dk} = \frac{K}{2\pi}, \tag{4.8}$$

which yields

$$\frac{d^2E}{dk^2} = \frac{1}{2\pi}, \quad \frac{dK}{dk} = \frac{l}{2\pi}, \quad \frac{dH}{dk} = -\frac{l}{k^2}\frac{dH}{dl}. \tag{4.9}$$

Thus, it follows from (4.4) and (4.5) that $\varphi(x;k)$ is unstable if $sgn(I(k)\cdot d^2E(k)/dk^2) < 0$. □

We now apply Theorem 4.1 to $\phi(x;k,\mu)$ defined by (3.9). To do so, setting $\varphi(x;k) = \phi(x;k,\mu)$ for any fixed μ, we define

$$I(k,\mu) := \int_0^{l(k)} \langle T\phi_x(x;k,\mu), Q\phi_x(x;k,\mu)\rangle\, dx \tag{4.10}$$

and

$$E(k,\mu) := \frac{1}{l(k)}\int_0^{l(k)} \left\{\frac{1}{2}\langle D\phi_x(x;k,\mu), Q\phi_x(x;k,\mu)\rangle - F(\phi(x;k,\mu))\right\} dx, \tag{4.11}$$

where $\phi(x;k,\mu)$ is given by (3.9). Then it follows from Theorem 4.1 that $\phi(x;k,\mu)$ is unstable if $sgn(I(k,\mu)\cdot\partial_k^2 E(k,\mu))<0$. This is a general result regardless of $k_c\neq 0$ which gives the definition of the Turing patterns. Thus it turns out that $\mu=\mu_E(k)$ derived from $\partial_k^2 E(k,\mu)=0$ plays a crucial role in determining the instability of $\phi(x;k,\mu)$. In fact, as seen in section 7, it coincides with the *Eckhaus instability criterion* [10] with respect to perturbations having a large spatial period. Recalling the fact that Theorem 4.1 does not require smallness of $\phi(x;k,\mu)$, we see that the Eckhaus instability criterion is valid away from a bifurcation point in gradient/skew-gradient dissipative systems. There is a wide range of literature concerning the Eckhaus instability, and main references are given in [20].

To conclude this section, we prepare the following lemma for the subsequent analysis.

LEMMA 4.3. *For* (4.2) *and* (4.3)*, we have*

$$\frac{dE(k)}{dk}=\frac{1}{2\pi}\int_0^{l(k)}\langle D\varphi_x(x;k),\ Q\varphi_x(x;k)\rangle dx,$$

$$\frac{d^2E(k)}{dk^2}=\frac{l(k)}{2\pi}\left(\langle D\varphi_x(x;k),\ Q\varphi_{xk}(x;k)\rangle-\langle D\varphi_{xx}(x;k),\ Q\varphi_k(x;k)\rangle\right).$$

Proof. The first equality directly follows from (4.8). Moreover, it follows from (2.2), (2.3), (4.1), and (4.6) that

$$\begin{aligned}\frac{dH}{dk}&=\frac{1}{2}\langle D\varphi_{xk},Q\varphi_x\rangle+\frac{1}{2}\langle D\varphi_x,Q\varphi_{xk}\rangle+\langle\bigtriangledown F(\varphi),\varphi_k\rangle\\&=\frac{1}{2}\langle\varphi_{xk},\ D^TQ\varphi_x\rangle+\frac{1}{2}\langle D\varphi_x,Q\varphi_{xk}\rangle+\langle Q\bigtriangledown F(\varphi),Q\varphi_k\rangle\\&=\frac{1}{2}\langle\varphi_{xk},QD\varphi_x\rangle+\frac{1}{2}\langle D\varphi_x,Q\varphi_{xk}\rangle+\langle f(\varphi),Q\varphi_k\rangle\\&=\langle D\varphi_x,Q\varphi_{xk}\rangle-\langle D\varphi_{xx},Q\varphi_k\rangle,\end{aligned}$$

where we used $Q=Q^T$ and $Q^2=I_n$. Thus, by (4.9) we see that the second equality is true. ∎

5. Properties of $I(k,\mu)$. In this section, we consider properties of $I(k,\mu)$ defined by (4.10). The contents in this section are regardless of whether or not $k_c\neq 0$, which concerns the definition of the Turing patterns. First, we introduce the notion of the sign of $I(k,\mu)$ at the bifurcation point (k_c,μ_c).

DEFINITION 5.1.

$$I_c=\mathrm{Re}\langle T\Psi_k,\ Q\Psi_k\rangle|_{k=k_c},\tag{5.1}$$

where Ψ_k *is defined by* (3.10).

Since it follows from (3.9) and (4.10) that $\lim_{\mu\downarrow\mu_c} sgn I(k,\mu)|_{k=k_c}=sgn I_c$, we call the sign of I_c the sign of $I(k,\mu)$ at the bifurcation point (k_c,μ_c).

Remark 5.2. In gradient systems, $I_c>0$ from its definition.

In order to study the sign of I_c, we now derive the amplitude equation which describes dynamics of (2.1) in a sufficiently small neighborhood of the bifurcation point (k_c,μ_c). It was first derived by Newell and Whitehead [26] in the study of dynamics sufficiently close to the onset of thermal convection. The derivation presented here is

a standard one [7]. Let $\bar{u}$ be a spatially uniform steady state of (2.1). Since μ is a bifurcation parameter which is included in the linearized operator (3.2), we write

$$L = L_\mu(\partial_x) = D\partial_x^2 + B.$$

Then $L_c = L_{\mu_c}(\partial_x)$ has the following property:

$$L_c(e^{ik_c x}\Psi_{k_c}) = 0,$$

where

$$\det(k_c^2 D - B) = 0,$$

and

$$(k_c^2 D - B)\Psi_{k_c} = 0, \qquad \Psi_{k_c} \in \mathbf{C}^n. \tag{5.2}$$

On the other hand, substituting $u = \bar{u} + v$ into (2.1), we have

$$Lv = Kv - g(v), \tag{5.3}$$

where $K = T\partial_t$ and $g(v) = f(\bar{u} + v) - Bv = a_2 v^2 + a_3 v^3 + \cdots$. Let us set

$$\varepsilon = \frac{\mu - \mu_c}{\mu_c}$$

and suppose that

$$v = \varepsilon^{1/2} v_0 + \varepsilon v_1 + \varepsilon^{3/2} v_2 + \cdots, \tag{5.4}$$

where

$$v_0 = A(y, s)e^{ik_c x}\Psi_{k_c} + c.c.,$$

and $y = \varepsilon^{1/2}x$ and $s = \varepsilon t$. Here we suppose that the time dependence of dynamics of v is expressed by only rescaled time variable s, not original t. It is well known that the dynamics of A is given by the Ginzburg–Landau equation

$$\tau A_s = \alpha A_{yy} + \beta A - \gamma |A|^2 A, \tag{5.5}$$

where α, β, γ, and τ are some constants determined by the subsequent calculation. We are interested in τ, though we can compute other constants α, β, and γ. We now show that $sgn\tau = sgnI_c$. Noting $\partial_x \to \partial_x + \varepsilon^{1/2}\partial_y$ and $\mu = \mu_c + \varepsilon\mu_c$, the linear operator $L = L_\mu(\partial_x)$ in the left-hand side of (5.3) can be expanded in ε as follows:

$$L = L_0 + \varepsilon^{1/2} L_1 + \varepsilon L_2, \tag{5.6}$$

where $L_0 = L_{\mu_c}(\partial_x) = L_c$, $L_1 = L_1(\partial_x, \partial_y)$, and $L_2 = L_2(\partial_x, \partial_y)$. Here we do not need the concrete expressions of L_1 and L_2. On the other hand, since the time variable of v is only s, applying $\partial_t \to \varepsilon\partial_s$, the linear operator K in the right-hand side of (5.3) can be expanded in ε as follows:

$$K = \varepsilon T\partial_s. \tag{5.7}$$

Substituting (5.4), (5.6), and (5.7) into (5.3) and comparing each coefficient of powers in ε, we have

$$\begin{aligned} &L_c v_0 = 0, \\ &L_c v_1 = -L_1 v_0 - a_2 v_0^2, \\ &L_c v_2 = -L_1 v_1 - L_2 v_0 + T\partial_s v_0 - 2a_2 v_0 v_1 - a_3 v_0^3. \end{aligned} \tag{5.8}$$

As a consequence, the dynamics of A is determined by the solvability condition for v_2 in the third equation of (5.8). Since τ is determined by $T\partial_s v_0$ in the right-hand side of the third equation of (5.8), applying the solvability condition, we find that $sgn\tau = sgn\mathrm{Re}\langle T\Psi_{k_c}, \Psi^*_{k_c}\rangle$, where $\Psi^*_{k_c} \in \mathbf{C}^n$ corresponds to the kernel of L_c^*, the adjoint operator of L_c, i.e.,

$$(k_c^2\, D^T - \, B^T)\Psi^*_{k_c} = 0.$$

As was seen in (3.11), under the gradient/skew-gradient structure, we have $\Psi^*_{k_c} = Q\Psi_{k_c}$. Hence it follows from (5.1) that $sgn\tau = sgn\mathrm{Re}\langle T\Psi_{k_c}, Q\Psi_{k_c}\rangle = sgnI_c$.

Thus we see that I_c determines the sign of time constant coefficient of amplitude equation (5.5), so that it affects the well posedness of (5.5). Next, we reconsider properties of I_c from another viewpoint when $n = 2$.

Let us recall a system of linear equations (3.4). When $n = 2$, the determinant of coefficient matrix of (3.4) becomes a quadratic equation in λ:

$$\alpha_0\lambda^2 + \alpha_1\lambda + \alpha_2 = 0, \tag{5.9}$$

where

$$\begin{aligned} &\alpha_0 = \tau_1\tau_2 > 0, \\ &\alpha_1 = \tau_2(k^2 d_{11} - b_{11}) + \tau_1(k^2 d_{22} - b_{22}), \end{aligned}$$

$T = \mathrm{diag}(\tau_1, \tau_2)$, $B = (b_{ij})$, and $D = (d_{ij})$.

We now consider a relation between $sgn\alpha_1$ and $sgnI_c$ at the bifurcation point (k_c, μ_c). Let $A = k_c^2 D - B$ for $\mu = \mu_c$. Then, we have

$$\alpha_1 = \tau_2 a_{11} + \tau_1 a_{22},$$

where $A = (a_{ij})$. On the other hand, we calculate $\langle T\Psi_{k_c}, Q\Psi_{k_c}\rangle$ explicitly. In the following calculation, we consider a case when $Q = \mathrm{diag}(1, -1)$ because other cases can be similarly treated. Since it follows from (5.2) that

$$\begin{aligned} a_{11}p + a_{12}q = 0, \\ a_{21}p + a_{22}q = 0, \end{aligned}$$

where $(p, q)^T = \Psi_{k_c}$, we have

$$\langle T\Psi_{k_c}, Q\Psi_{k_c}\rangle = \tau_1|p|^2 - \tau_2|q|^2 = \tau_1|p|^2 - \tau_2|p|^2\,\frac{a_{11}}{a_{12}}\frac{a_{21}}{a_{22}} = \tau_1|q|^2\,\frac{a_{12}}{a_{11}}\frac{a_{22}}{a_{21}} - \tau_2|q|^2.$$

Noting that $a_{12} = -a_{21}$ because $A^TQ = QA$ by virtue of (2.3), (2.4), and $B = f_u(\bar{u})$, we have

$$\langle T\Psi_{k_c}, Q\Psi_{k_c}\rangle = |p|^2\,\frac{\tau_2 a_{11} + \tau_1 a_{22}}{a_{22}} = -|q|^2\,\frac{\tau_2 a_{11} + \tau_1 a_{22}}{a_{11}}.$$

Therefore, it follows from (5.1) that $sgn\,\alpha_1 = sgn\,I_c \cdot sgn\,a_{22} = -sgn\,I_c \cdot sgn\,a_{11}$.

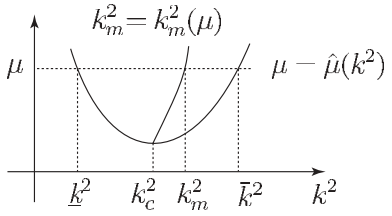


FIG. 2. *For $\mu > \hat{\mu}(k^2)$ near the Turing bifurcation point, there exists a unique extreme point of $E(k, \mu)$ given by $k_m = k_m(\mu)$.*

On the other hand, the quadratic equation (5.9) must have roots with $\mathrm{Re}\lambda < 0$ for $\mu < \mu_c$ and arbitrary k as in (3.8) because the homogeneous steady state $\bar{u}$ must be stable for $\mu < \mu_c$. Therefore, $\alpha_1 > 0$ must be true for $\mu < \mu_c$ near the bifurcation point (k_c, μ_c), so that $\alpha_1 > 0$ holds at the bifurcation point. This implies that the assumption (3.8) cannot be satisfied by changing the sign of I_c. Thus we see that the sign of I_c affects the instability of an underlying homogeneous steady state to be equipped with spatially periodic structure.

According to the justification theory of validity of amplitude equation (Collet and Eckmann [6], Eckhaus [11], Harten [16], Mielke and Schneider [23]), under certain conditions, the dynamics of (2.1) in a sufficiently small neighborhood of the bifurcation point (k_c, μ_c) is well approximated by (5.5) as long as (5.5) is well posed. Thus, recalling the definition of gradient systems, the above arguments lead us to the following observation.

OBSERVATION 5.3. *The sign of a time constant coefficient of amplitude equation is determined by I_c, the sign of $I(k, \mu)$ at the bifurcation point (k_c, μ_c). It is closely related to the well posedness of the amplitude equation and the instability of the underlying homogeneous steady state to be equipped with spatially periodic structure. Moreover, it is necessary that $I_c > 0$ so that the skew-gradient system can admit gradient structure in a sufficiently small neighborhood of the bifurcation point (k_c, μ_c).*

6. Properties of $E(k, \mu)$. In this section, we consider properties of (skew) free energy $E(k, \mu)$ defined by (4.11) for the Turing patterns near the bifurcation point and establish the uniqueness of the extremum of $E(k, \mu)$. First, we introduce a nondegeneracy condition at the bifurcation point (k_c, μ_c).

HYPOTHESIS 6.1. $E_c := \mathrm{Re}\,\partial_k \langle D\Psi_k, Q\Psi_k \rangle|_{k=k_c} \neq 0$, *where Ψ_k is defined by* (3.10).

We call this hypothesis the nondegeneracy condition at the bifurcation point (k_c, μ_c). The following result shows uniqueness of the extremum of (skew) free energy.

THEOREM 6.2. *Let us assume that the above hypothesis is met and $k_c \neq 0$.*

(1) *For any fixed $\mu > \mu_c$ near the Turing bifurcation point, there exists a unique extremum of $E(k) = E(k, \mu)$ in $(\underline{k}, \bar{k})$, where $\underline{k}$ and $\bar{k}$ are defined by $\hat{\mu}(\underline{k}^2) = \hat{\mu}(\bar{k}^2) = \mu$ with $\hat{\mu}$ given by* (3.6).

(2) *The wavenumber $k_m = k_m(\mu)$ corresponding to the extremum is approximately obtained by $\mathrm{Re}\langle D\Psi_k, Q\Psi_k \rangle = 0$ near the Turing bifurcation point* (*Figure* 2).

Remark 6.3. (1) Theorem 6.2 implies (cf. section 3) that if the Turing bifurcation parameter μ is in reaction term f (not in D), then $k_m(\mu)$ is approximately given by k_c (independent of μ). In this case, we denote $k_m(\mu) \approx k_c$.

(2) As seen in the next section, $k_c \leq k_m(\mu)$ holds for two component gradient/skew-gradient dissipative systems.

(3) The reader may expect that the above theorem can be proved by the theories in [6, 11, 16, 23] because the amplitude equation describes the dynamics in a sufficiently small neighborhood of the bifurcation point (k_c, μ_c). However, Theorem 6.2 is obtained by applying Theorem 4.1 and Lemma 4.3, which are valid away from the bifurcation point. It is essentially different from the results obtained by using the amplitude equation derived at the bifurcation point.

When the above assumption is satisfied, QD is not definite because of Theorem 3.2. It is necessary for existence of k satisfying $\mathrm{Re}\langle D\Psi_k, Q\Psi_k\rangle = 0$. Moreover, we notice that the above result does not require any other assumptions, and that the unique extremum is the (global) minimum when $E_c > 0$. In order to prove the above theorem, we prepare the following lemma.

LEMMA 6.4. *Under the assumption of Theorem* 6.2*, the following hold near the Turing bifurcation point* (k_c, μ_c):

(i) *When* $E_c > 0$, $d^2E/dk^2 > 0$ *if* $dE/dk = 0$.

(ii) *When* $E_c < 0$, $d^2E/dk^2 < 0$ *if* $dE/dk = 0$.

Proof. We consider case (i) because (ii) can be similarly treated. We suppose that $0 < \underline{k} < k_c < \bar{k}$ without loss of generality. In the following calculation, we neglect higher order terms of (3.9). By (3.9), we have

$$\phi_x = ikae^{ikx}\Psi_k + c.c.,$$

which yields

$$\begin{aligned}(6.1)\qquad \int_0^l \langle D\phi_x, Q\phi_x\rangle\, dx &= \int_0^l \langle\, D(ikae^{ikx}\Psi_k + c.c.),\, Q(ikae^{ikx}\Psi_k + c.c.)\,\rangle\, dx \\ &= 2lk^2a^2\mathrm{Re}\langle D\Psi_k,\, Q\Psi_k\rangle.\end{aligned}$$

Hence it follows from Lemma 4.3 that $\mathrm{Re}\langle D\Psi_k, Q\Psi_k\rangle = 0$ if $dE/dk = 0$. On the other hand, since

$$\begin{aligned}\phi_{xx} &= -k^2ae^{ikx}\Psi_k + c.c.,\\ \phi_{xk} &= iae^{ikx}\Psi_k + ik(\partial_k a)e^{ikx}\Psi_k - kxae^{ikx}\Psi_k + ikae^{ikx}(\partial_k\Psi_k) + c.c.,\\ \phi_k &= (\partial_k a)e^{ikx}\Psi_k + ixae^{ikx}\Psi_k + ae^{ikx}(\partial_k\Psi_k) + c.c.,\end{aligned}$$

we have

$$\begin{aligned}&\langle D\phi_x, Q\phi_{xk}\rangle\\ &= \langle D(ikae^{ikx}\Psi_k + c.c.),\ Q(iae^{ikx}\Psi_k + ik(\partial_k a)e^{ikx}\Psi_k\\ &\qquad\qquad - kxae^{ikx}\Psi_k + ikae^{ikx}(\partial_k\Psi_k) + c.c.)\rangle\end{aligned}$$

and

$$\begin{aligned}&\langle D\phi_{xx}, Q\phi_k\rangle\\ &= \langle\, D(-k^2ae^{ikx}\Psi_k + c.c.),\ Q(\,(\partial_k a)e^{ikx}\Psi_k + ixae^{ikx}\Psi_k + ae^{ikx}(\partial_k\Psi_k) + c.c.)\rangle.\end{aligned}$$

Therefore, when $\mathrm{Re}\langle D\Psi_k, Q\Psi_k\rangle = 0$, we have

$$\langle D\phi_x, Q\phi_{xk}\rangle - \langle D\phi_{xx}, Q\phi_k\rangle = 4k^2a^2\mathrm{Re}\langle D\Psi_k, Q\partial_k\Psi_k\rangle.$$

On the other hand, since $Q = Q^T$, QD is a symmetric matrix by (2.3) and $\partial_k(QD)|_{k=k_c} = 0$ by (3.7), we have

$$2\mathrm{Re}\langle D\Psi_k,\, Q\partial_k\Psi_k\rangle|_{k=k_c} = \mathrm{Re}\,\partial_k\langle QD\Psi_k,\, \Psi_k\rangle|_{k=k_c} = \mathrm{Re}\,\partial_k\langle D\Psi_k,\, Q\Psi_k\rangle|_{k=k_c}.$$

Thus, by (6.1) and Lemma 4.3, we see that near the Turing bifurcation point, $d^2E/dk^2 > 0$ holds if $dE/dk = 0$ when $E_c > 0$. □

Proof of Theorem 6.2. Since $\phi(x;\underline{k},\mu) = \phi(x;\bar{k},\mu) \equiv \bar{u}$, we have $E(\underline{k}) = E(\bar{k})$. Hence there exists $k_0 \in (\underline{k},\bar{k})$ such that $dE(k_0)/dk = 0$. Moreover, by Lemmas 4.3 and 6.4, we see that such a k_0 must be unique and approximately calculated by $\mathrm{Re}\langle D\Psi_k,\, Q\Psi_k\rangle = 0$ because of (6.1). □

7. Application. In this section, we apply our results to typical examples. They are helpful to understand the usefulness and meaning of our results.

First, we consider the real Ginzburg–Landau equation,

$$u_t = u_{xx} + u(\mu - u^2 - v^2), \quad v_t = v_{xx} + v(\mu - u^2 - v^2), \tag{7.1}$$

which can be rewritten as (2.1) with $T = D = Q = I_2$ and $F = \mu(u^2+v^2)/2 - (u^2+v^2)^2/4$. By using Theorem 3.2, we see that (7.1) does not have the Turing patterns (characterized by $k_c \neq 0$) because $QD = I_2$ is positive. It is well known, however, that (7.1) has a family of spatially periodic stationary solutions

$$\phi(x;k,\mu) = \sqrt{\mu - k^2}\,(\cos kx,\, \sin kx).$$

They are trivial patterns because the critical wavenumber k_c is equal to zero at the bifurcation point $\mu_c = 0$. On the other hand, recalling Definition 2.1, (7.1) is a gradient system because $QT = I_2$ is positive. Hence we apply Corollary 4.2 to the above family $\phi(x;k,\mu)$ for any fixed μ. It follows from direct calculation that

$$E(k) = E(k,\mu) = -\frac{1}{4}(\mu - k^2)^2,$$

which yields

$$E''(k) = \partial_k^2 E(k,\mu) = \mu - 3k^2.$$

Thus, an instability criterion is given by $\mu = \mu_E(k) = 3k^2$, so that $\phi(x;k,\mu)$ is unstable if $\mu < 3k^2$. This result is well known as the Eckhaus instability. For other examples of applications of Theorem 4.1 and Corollary 4.2; see [20].

Next, we apply Theorem 6.2 to the following model system with skew-gradient structure, which is studied in Ben-Jacob et al. [3]:

$$\tau_1 u_t = d_1 u_{xx} + \alpha(1-u^2)u - \beta v, \quad \tau_2 v_t = d_2 v_{xx} - \gamma(1+v^2)v + \beta u, \tag{7.2}$$

where $\alpha, \beta, \gamma, d_1, d_2, \tau_1$, and τ_2 are positive constants. This equation is rewritten as (2.1) with

$$T = \begin{pmatrix} \tau_1 & 0 \\ 0 & \tau_2 \end{pmatrix}, \quad D = \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

$$F = F(u,v) = \frac{\alpha}{4}(1-u^2)^2 - \beta uv + \frac{\gamma}{4}(1+v^2)^2.$$

As usual, we consider the linearized eigenvalue problem at the trivial steady state $\bar{u} = 0$ and obtain a system of linear equations

$$(\lambda T + k^2 D - B)\Psi_k = 0, \tag{7.3}$$

where $\langle \Psi_k, \Psi_k \rangle = 1$, and

$$B = \begin{pmatrix} \alpha & -\beta \\ \beta & -\gamma \end{pmatrix}.$$

It follows from $\det(k^2 D - B) = 0$ that

$$\beta^2 - (\alpha - k^2 d_1)(\gamma + k^2 d_2) = 0, \tag{7.4}$$

so that (7.3) has nontrivial solution $\Psi_k = (p_k, q_k)^T$ for $\lambda = 0$, where

$$p_k^2 = \frac{\beta^2}{\beta^2 + (\alpha - k^2 d_1)^2} := g_1(k^2),$$
$$q_k^2 = \frac{(\alpha - k^2 d_1)^2}{\beta^2 + (\alpha - k^2 d_1)^2} := g_2(k^2).$$

Hence, we have

$$\langle D\Psi_k, Q\Psi_k \rangle = d_1 p_k^2 - d_2 q_k^2 = d_1 g_1(k^2) - d_2 g_2(k^2), \tag{7.5}$$

$$\partial_k \langle D\Psi_k, Q\Psi_k \rangle = 2d_1 k g_1'(k^2) - 2d_2 k g_2'(k^2), \tag{7.6}$$

$$\langle T\Psi_k, Q\Psi_k \rangle = \tau_1 p_k^2 - \tau_2 q_k^2 = \tau_1 g_1(k^2) - \tau_2 g_2(k^2). \tag{7.7}$$

First, we choose d_2 as a Turing bifurcation parameter. By using (7.4), we have

$$d_2 = \hat{d}_2(k^2) = \frac{\beta^2}{k^2(\alpha - k^2 d_1)} - \frac{\gamma}{k^2}$$

for $0 < k^2 < \alpha/d_1$. By using $\hat{d}_2\,' = 0$, we have

$$k_c^2 = \frac{\alpha\sqrt{\beta^2 - \alpha\gamma}}{d_1(\beta + \sqrt{\beta^2 - \alpha\gamma})} > 0 \tag{7.8}$$

and

$$d_2^c = \hat{d}_2(k_c^2) = \frac{d_1(\beta + \sqrt{\beta^2 - \alpha\gamma})^2}{\alpha^2}. \tag{7.9}$$

As was explained in section 3, we can construct the Turing patterns $\phi(x; k, d_2)$ for $d_2 > \hat{d}_2(k^2)$ near the bifurcation point $(k_c^2, \hat{d}_2(k_c^2))$ provided

$$0 < \alpha < \gamma \quad \text{and} \quad \beta^2 - \alpha\gamma > 0. \tag{7.10}$$

By using (7.6), we can compute the value of $\partial_k \langle D\Psi_k, Q\Psi_k \rangle$ at $k = k_c$, which determines E_c. In fact, direct calculation yields

$$\partial_k \langle D\Psi_k, Q\Psi_k \rangle = \frac{4\beta^2 d_1(d_1 + d_2)k(\alpha - k^2 d_1)}{(\beta^2 + (\alpha - k^2 d_1)^2)^2} > 0,$$

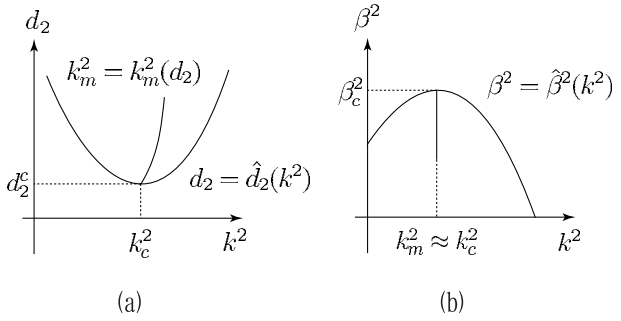


FIG. 3. *The minimizer of (skew) free energy for the Turing patterns of model equation* (7.2): (a) *The Turing bifurcation parameter* $\mu = d_2$; (b) $\mu = \beta^2$.

which implies $E_c > 0$. Thus we find that the Turing bifurcation point is nondegenerate and $E_c > 0$ when $n = 2$. Moreover, it follows from (7.5) and $\langle D\Psi_k, Q\Psi_k \rangle = 0$ that the minimizer of (skew) free energy per unit length is given by

$$k_m^2 = k_m^2(d_2) = \frac{\alpha}{d_1} - \frac{\beta}{\sqrt{d_1 d_2}}. \tag{7.11}$$

Furthermore, by using (7.9), we can easily verify that

$$k_c^2 < k_m^2(d_2)$$

holds provided $d_2 > d_2^c$. Since the above calculation can be used in general two component systems with gradient/skew-gradient structure, the above results are universal features of two component systems with gradient/skew-gradient structure. These results are presented in Figure 3(a).

On the other hand, by using (7.7), we can compute the value of $\langle T\Psi_k, Q\Psi_k \rangle$ at $k = k_c$, which determines the sign of I_c. In fact, it follows from (7.7) and (7.8) that

$$sgn\, I_c = sgn\, (2\beta\tau_1(\beta + \sqrt{\beta^2 - \alpha\gamma}) - \alpha(\gamma\tau_1 + \alpha\tau_2)).$$

Next, we choose β^2 as a Turing bifurcation parameter, which is a case treated in [3]. As was seen in section 3, we see that the Turing patterns appear for

$$\beta^2 < \hat{\beta}^2(k^2) = (\alpha - k^2 d_1)(\gamma + k^2 d_2)$$

near the bifurcation point (k_c^2, β_c^2), where

$$k_c^2 = \frac{\alpha d_2 - \gamma d_1}{2 d_1 d_2} > 0 \tag{7.12}$$

and

$$\beta_c^2 = \hat{\beta}^2(k_c^2) = \frac{(\alpha d_2 + \gamma d_1)^2}{4 d_1 d_2}. \tag{7.13}$$

In this case, while Turing patterns appear for $\beta^2 < \hat{\beta}^2(k^2)$, our results can also be applied. In a manner similar to the previous case, we obtain

$$k_m^2 = k_m^2(\beta^2) = \frac{\alpha d_2 - \gamma d_1}{2 d_1 d_2}.$$

Thus, as is shown in Figure 3(b), we verified that $k_m \approx k_c$ holds when the Turing bifurcation parameter is in the reaction term (not in D).

As we have observed in the arguments so far, the Turing bifurcation gives rise to a family of spatially periodic patterns with wavenumber k close to the critical wavenumber k_c. The (skew) free energy on these patterns has an extreme value only at the pattern with a uniquely determined wavenumber k_m. When Turing bifurcations occur in actual systems, it is expected that the wavenumber k_m plays an important role in determining which pattern, or, which wavenumber, is to be selected. In the next section, we investigate this problem for several concrete model systems by employing numerical experiments.

8. Pattern selection. In this section, combining numerical experiments with the analytical results obtained in the previous section, we investigate what pattern is to be uniquely selected among many near the Turing bifurcation point.

Let us explain our problem more specifically. We have discussed basic properties of the Turing patterns in an *infinite* domain. However, in real (numerical) experiments, they are observed in a *finite* domain. Thus we consider the gradient/skew-gradient systems (2.1) with (2.2) on a finite interval under the periodic boundary conditions

$$\begin{aligned} &Tu_t = Du_{xx} + f(u), \quad 0 < x < L, \\ &u(x,0) = \varepsilon u_0(x), \end{aligned} \tag{8.1}$$

where ε and $|D|/L$ are sufficiently small, $|D| = \max |d_{ij}|$, and $u_0(x)$ is bounded. For each fixed μ near the Turing bifurcation point μ_c, we numerically solve (8.1) by using the pseudospectral method and the discrete FFT. Then we study a spatial profile and the Fourier power spectrum of $u(x, T_1)$ for sufficiently large T_1. Notice that numerical computations sufficiently close to a (Turing) bifurcation point for sufficiently small $|D|/L$ are very delicate tasks.

According to [11], when μ is sufficiently close to μ_c, the power spectrum of $u(x, T_1)$ for sufficiently large T_1 has a (bell) shape such as Figure 5(e) centered at $k = k_c$ if the power spectrum of an initial data $u_0(x)$ takes the maximal peak only at $k = k_c$. It should be noted that this is valid for general systems regardless of whether they have gradient/skew-gradient dissipative structures.

Here we are interested in a spatial profile and the Fourier power spectrum of $u(x, T_1)$ for sufficiently large T_1 for a uniform distribution $u_0(x) \in (-1/2, 1/2)$ generated by pseudorandom numbers. Noting $k_m(\mu_c) = k_c$ and the above mentioned fact, we expect that the distribution of power spectrum of $u(x, T_1)$ is concentrated in a cluster centered at $k = k_c$ near the Turing bifurcation point $\mu = \mu_c$. We investigate the peak of distribution of power spectrum of $u(x, T_1)$ in the cluster.

Let us first consider the Swift–Hohenberg equation

$$u_t = \mu u - (1 + \partial_{xx})^2 u - u^3. \tag{8.2}$$

As seen in section 2, (8.2) is a gradient system that can be rewritten as (2.1). It is easy to see that (8.2) has a Turing bifurcation point $\mu_c = 0$ with a critical wavenumber $k_c = 1$. In fact, it is well known [5] that (8.2) has a family of spatially periodic stationary solutions

$$\phi(x; k, \mu) = \frac{2a}{\sqrt{3}} \cos(kx) + O(a^3),$$

where $a = \sqrt{\mu - (1 - k^2)^2}$. Applying the results in the previous section, we find that $k_m(\mu) \approx k_c$ holds. This suggests that the power spectrum of $u(x, T_1)$ for sufficiently

large T_1 takes the maximal peak at $k_c = 1$. We set $L = 200\pi$ so that L/λ_c is an integer, where $\lambda_c = 2\pi$ is wavelength of the Turing pattern with wavenumber k_c. In this case, the discrete FFT can capture a variation of wavenumber of the Turing pattern in the accuracy $\delta k = 2\pi/L = 0.01$. Notice that a relative variation of wavenumber to the basic pattern associated with $k_c = 1$ is given by $\delta k/k_c = 0.01$. For $\mu = 0.01$ and $\varepsilon = 0.0001$, Figures 4 and 5 show snapshots of spatial profiles and the Fourier power spectra of a numerical solution of (8.2). After sufficiently large time, the spatial profile in Figure 4(e) looks like a (meta)stable stationary pattern with spatially periodic structure. In what follows, T_1 is determined by the time when a solution of a system reaches a (meta)stable stationary pattern as in Figure 4(e). Notice that T_1 depends on an initial condition and all parameters in a system. In this case, we set $T_1 = 2000$. We regard a spatial pattern of $u(x, T_1)$ as the Turing pattern in our numerical experiments. Furthermore, we say that the Turing pattern with wavenumber 1.00 is selected because the Fourier power spectrum in Figure 5(e) has the maximal peak at $k = 1.00$.

In order to study what Turing pattern is to be uniquely selected among many, we numerically solve (8.2) for some other initial data. In general, the Fourier power spectrum of $u(x, T_1)$ does not necessarily have a clear (bell) shape (for example, see Figure 6). Therefore, we compute the power spectra of $u(x, T_1)$ for 30 random initial data and take their average. The result shows that the wavenumber $k = 1.00$ is selected with the highest probability. We call this wavenumber *the selected wavenumber* denoted by $k_s(\mu)$. In this case, $k_s(0.01) = 1.00$. The selected wavenumbers for various μ are shown in Table 1.

Thus we see that $k_s(\mu) \approx k_m(\mu) \approx k_c$ holds near the Turing bifurcation point in this example.

Next, we treat again the model system (7.2) from the previous section:

$$\tau_1 u_t = d_1 u_{xx} + \alpha(1 - u^2)u - \beta v, \quad \tau_2 v_t = d_2 v_{xx} - \gamma(1 + v^2)v + \beta u.$$

This system was used in Ben-Jacob et al. [3] for a study of patterns generated from small initial data with *compact support* in various dissipative systems. They studied the problem of front propagation of local Turing patterns generated by a small local perturbation into a linearly unstable steady state. For recent progress of this topic, we refer to Ebert and Saarloos [9] and Saarloos [34].

First, we consider a case $\mu = \beta^2$. As was seen in the previous section, in this case, $k_m(\mu) \approx k_c$ holds, and we may expect that $k_s(\mu) \approx k_m(\mu)$. In a manner similar to the previous example, we set $L = 200\pi$ for $k_c = 1$. We choose $\alpha = 1.0$, $\gamma = 2.0$, $d_1 = 0.25$, and $d_2 = 1.0$ satisfying (7.12). Then, it follows from (7.13) that $\beta_c = 1.5$. Moreover, noting (7.10), we choose $\beta = 1.48$, so that the distance from the bifurcation point is given by $|\mu - \mu_c|/\mu_c = (\beta_c^2 - \beta^2)/\beta_c^2 \approx 0.0264889$.

For $\tau_1 = \tau_2 = 1.0$, $\varepsilon = 0.0001$, and these parameter values, numerically solving (7.2), we have a spatial profile and the Fourier power spectrum of $u(x, T_1)$ for $T_1 = 3000$ as in Figure 6. In a manner similar to the previous example, computing the power spectra of $u(x, T_1)$ for 30 random initial data, and taking their average, we determine that $k_s((1.48)^2) = 0.97$. Moreover, the selected wavenumber $k_s(\mu)$ for various μ are shown in Table 2.

Contrary to our expectation, this statistical result shows that $k_s(\mu)$ does not necessarily coincide with $k_m(\mu)$. More precisely, $k_s(\mu) \approx k_m(\mu) \approx k_c$ holds in a sufficiently small neighborhood of the Turing bifurcation point, while $k_s(\mu) < k_m(\mu)$ holds slightly away from the bifurcation point. Similar results, as in Table 2, can also be obtained for other values of $\alpha, \gamma, d_1, d_2, \tau_1$, and τ_2.

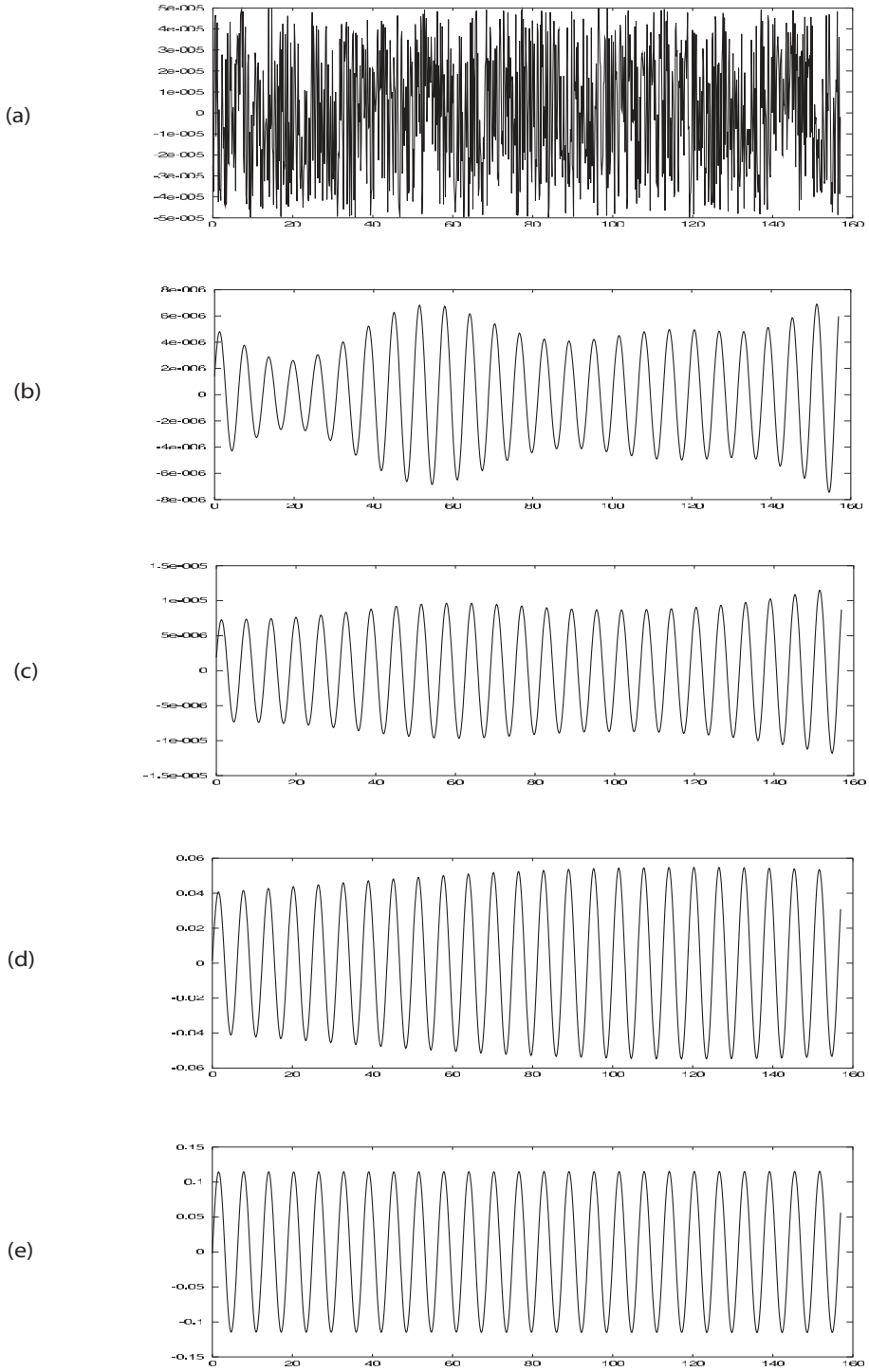


FIG. 4. *Snapshots of spatial profiles for numerical solutions of* (8.2) *in a window* $0 \leq x \leq 50\pi$*;* (a) $t = 0$, (b) $t = 50$, (c) $t = 100$, (d) $t = 1000$, (e) $t = 2000$.

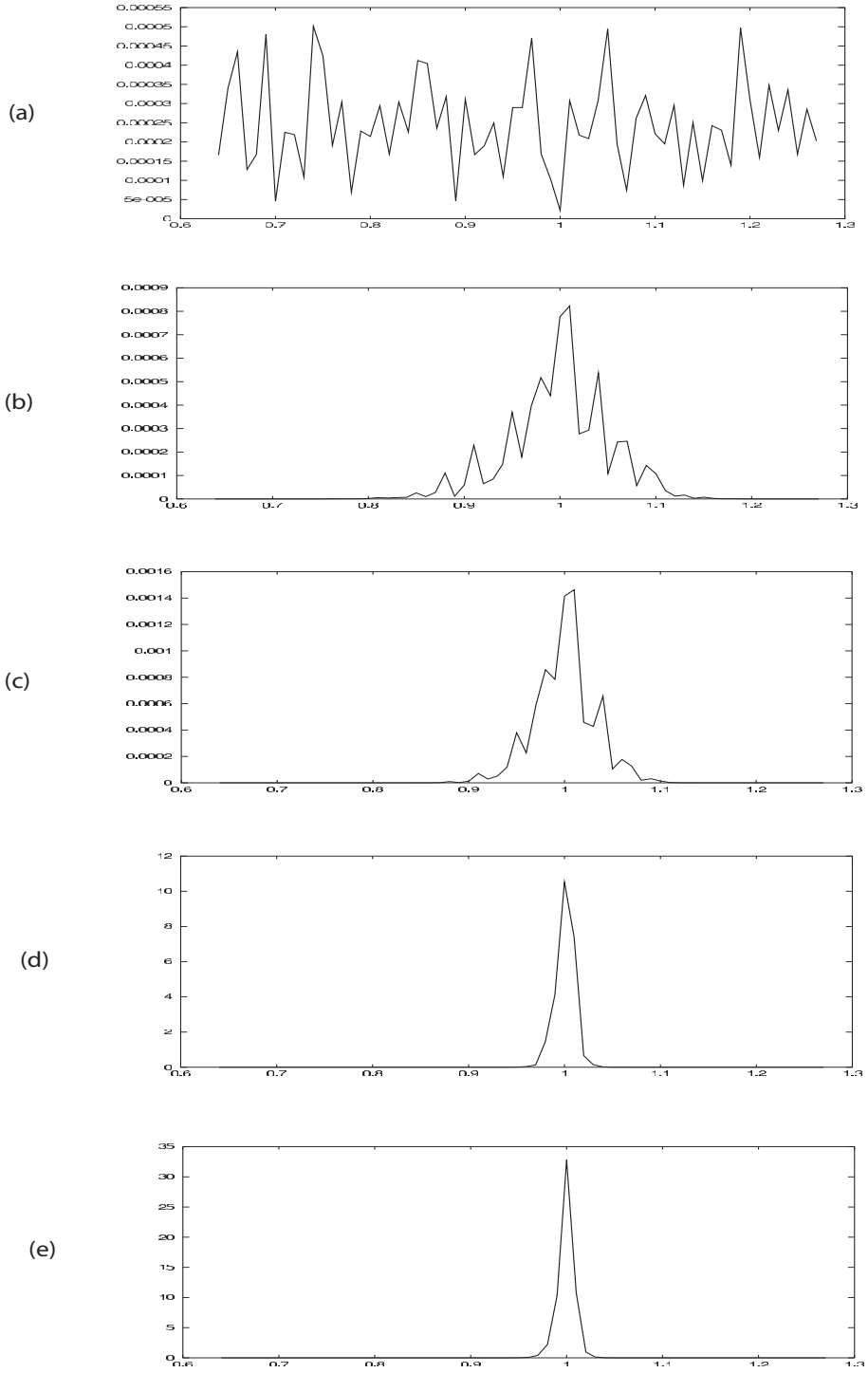


FIG. 5. *Snapshots of the Fourier power spectra for numerical solutions of* (8.2) *in a window* $0.64 \leq k \leq 1.28$*;* (a) $t = 0$, (b) $t = 50$, (c) $t = 100$, (d) $t = 1000$, (e) $t = 2000$.

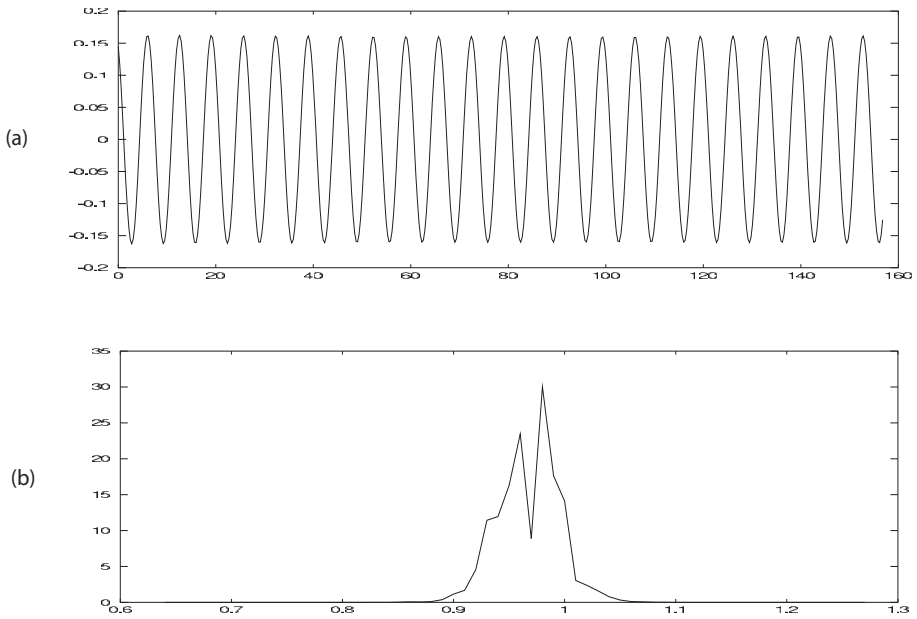


FIG. 6. *A snapshot for numerical solutions of* (7.2) *at* $t = 3000$*;* (a) *spatial profile in a window* $0 \le x \le 50\pi$*,* (b) *the Fourier power spectrum in a window* $0.64 \le k \le 1.28$*.*

TABLE 1
The selected wavenumbers $k_s(\mu)$ *for various* μ*.*

μ	T_1	$k_s(\mu)$
0.01	2000	1.00
0.02	1500	1.00
0.05	1000	1.00
0.1	500	1.00

TABLE 2
The selected wavenumbers $k_s(\mu)$ *for various* μ*.*

μ	$\lvert\mu - \mu_c\rvert/\mu_c$	$k_c(\approx k_m(\mu))$	$k_s(\mu)$
$(1.49)^2$	0.0132889	1.00	0.99
$(1.485)^2$	0.0199	1.00	0.97
$(1.48)^2$	0.0264889	1.00	0.97
$(1.46)^2$	0.0526222	1.00	0.94

Next, we consider the case $\mu = d_2$. As was seen in section 7, in this case, $k_m = k_m(\mu)$ does not coincide k_c, and $k_m > k_c$ holds for $\mu > \mu_c$. In the following numerical experiments, we set $\tau_1 = \tau_2 = 1.0$, $\varepsilon = 0.0001$, and $L = 200\pi$ for $k_c = 1$ as in the previous case. Noting (7.10), we choose $\alpha = 1.0$, $\beta = 1.5$, and $\gamma = 2.0$. Then it follows from (7.8) and (7.9) that $d_1 = 0.25$ and $d_2^c = 1.0$. In order to compare k_c and k_m, by using (7.11), we choose $\mu = d_2$ such that $k_m(\mu)$ can be captured by the discrete FFT. In a manner similar to the previous case, we can obtain (see Table 3) the selected wavenumbers $k_s(\mu)$ for various $\mu = d_2$.

TABLE 3
$k_s(\mu)$ and $k_m(\mu)$ for various μ when $\alpha = 1.0$, $\beta = 1.5$, and $\gamma = 2.0$.

μ	$\lvert\mu - \mu_c\rvert/\mu_c$	k_c	$k_s(\mu)$	$k_m(\mu)$
1.01354	0.0135359	1.00	1.00	1.01
1.02749	0.0274873	1.00	1.00	1.02
1.072	0.0720021	1.00	1.01	1.05
1.1562	0.156203	1.00	1.01	1.10

TABLE 4
$k_s(\mu)$ and $k_m(\mu)$ for various μ when $\alpha = 1.0$, $\beta = 1.43$, $\gamma = 2.0$.

μ	$\lvert\mu - \mu_c\rvert/\mu_c$	k_c	$k_s(\mu)$	$k_m(\mu)$
0.352115	0.0120813	1.00	1.01	1.02
0.358726	0.0310831	1.00	1.03	1.05
0.370618	0.0652656	1.00	1.05	1.10
0.383708	0.102891	1.00	1.08	1.15

Similarly, Table 4 shows a case in which $\alpha = 1.0$, $\beta = 1.43$, $\gamma = 2.0$, $d_1 \approx 0.129056$, and $d_2^c \approx 0.347912$.

In this example, the above statistical results show that $k_s(\mu) \approx k_m(\mu) \approx k_c$ holds in a sufficiently small neighborhood of the Turing bifurcation point, while $k_s(\mu) \leq k_m(\mu)$ holds (slightly) away from the bifurcation point.

As expected, these numerical results show that $k_s(\mu) \approx k_c$ holds in a sufficiently small neighborhood of the Turing bifurcation point μ_c. This is certainly true for any general system. On the other hand, the above results clearly show that $k_s(\mu) \approx k_c$ is not always true, even if μ is rather close to the bifurcation point μ_c. Thus, the statement $k_s(\mu) \approx k_c$, which has been widely accepted as a working principle[1] in many previous references, has to be modified so that it is valid in a realistic parameter region near the bifurcation point. As for gradient/skew-gradient dissipative systems, keeping Remark 6.3(2) in mind, we propose the following conjecture as a general (modified version of) pattern selection principle which is consistent with the above experimental results.

CONJECTURE 8.1. *For gradient/skew-gradient dissipative systems, the inequality*

$$k_s(\mu) \leq k_m(\mu) \tag{8.3}$$

holds near the Turing bifurcation point.

The reader may wonder about the validity of the above conjecture because $k_s(\mu)$ has no precise definition; how can we predict $k_s(\mu)$? A partial answer for this problem is reported in a separate paper [21]. In fact, for two component gradient/skew-gradient systems, we can explain a mechanism which determines $k_s(\mu)$ and confirm the validity of (8.3) near the Turing bifurcation point. As for n-component systems, however, we do not know whether or not (8.3) is true because we have not yet performed analytical calculations and numerical experiments as in sections 7 and 8. In general,

[1] This principle and the definition of the pattern to be (uniquely) selected have not been explicitly expressed.

some nondegeneracy conditions such as $E_c > 0$ may be required, where E_c is defined in Hypothesis 6.1.

The inequality (8.3) turns out to be very important in selection problems for roll patterns in two-dimensional problems. In [21], we verify the validity of the marginal stability hypothesis [30, 31, 32], which asserts that the selected roll pattern is determined by $D_\perp = 0$ corresponding to the zigzag instability criterion. As seen in [20], $D_\perp = 0$ is given by $k_m(\mu)$ in gradient/skew-gradient dissipative systems. For roll patterns in two-dimensional problems, we can introduce $k_s(\mu)$ in a manner similar to the one-dimensional case, and we find that (8.3) plays a key role in studying marginal stability hypotheses for roll patterns in two-dimensional problems under the gradient/skew-gradient structure.

9. Summary. In this paper, we consider some fundamental properties of Turing patterns in gradient/skew-gradient dissipative systems. The results are summarized as follows:

(1) To prove that QD is definite, we cannot produce the Turing patterns (Theorem 3.2). This characterizes Turing patterns a little more precisely than commonly accepted explanations based on two types of components with opposing kinetics and different diffusion rates.

(2) To give an instability criterion for spatially periodic steady states (Theorem 4.1), it is represented in terms of $I(k)$ and $E(k)$. $E(k)$ is (skew) free energy per unit length of spatially periodic steady states and convexity of $E(k)$ in wavenumber k determines the instability. This is an extension of the Eckhaus instability criterion. Notice that this criterion is valid far from a bifurcation point.

For any fixed bifurcation parameter μ, we define $I(k, \mu)$ and $E(k, \mu)$ for spatially periodic patterns $\phi(x; k, \mu)$ with wavenumber k near a bifurcation point.

(3) The sign of $I(k, \mu)$ at a bifurcation point determines the sign of time constant coefficient of the amplitude equation, which describes dynamics in a sufficiently small neighborhood of the bifurcation point. Moreover, $I(k, \mu)$ affects the instability of the underlying homogeneous steady state to be equipped with spatially periodic structure (see Observation 5.3).

(4) Uniqueness of extremum of $E(k, \mu)$ in k is established under the nondegeneracy condition (Theorem 6.2). This condition can be easily verified at the Turing bifurcation point.

(5) The wavenumber corresponding to the unique extremum of $E(k, \mu)$ gives an upper bound for the wavenumber of the Turing pattern to be uniquely selected among many near the Turing bifurcation point (Conjecture 8.1). The intrinsic symmetric property in the gradient/skew-gradient structure gives a prohibition law in selection of the Turing patterns near a bifurcation point.

As seen in section 7, from a practical viewpoint we propose the following calculation procedure to obtain the Turing pattern with minimum (skew) free energy per unit length:

(1) To verify that QD is not definite.

(2) To determine Turing bifurcation point, choose a bifurcation parameter μ, seek the relation $\mu = \hat{\mu}(k^2)$ by solving $\det(k^2 D - f_u(\bar{u})) = 0$, and determine Turing bifurcation point μ_c and critical wavenumber k_c by $\partial_\theta \hat{\mu}(\theta)|_{\theta=k_c^2} = 0$ and $\mu_c = \hat{\mu}(k_c^2)$.

(3) To compute zero eigenvector Ψ_k by $(k^2 D - f_u(\bar{u}))\Psi_k = 0$ for $\mu = \hat{\mu}(k^2)$.

(4) To check the hypothesis

$$I_c := \mathrm{Re}\langle T\Psi_k, Q\Psi_k\rangle|_{k=k_c} > 0 \text{ and } E_c := \mathrm{Re}\,\partial_k\langle D\Psi_k, Q\Psi_k\rangle|_{k=k_c} > 0.$$

(5) To compute the wavenumber $k_m = k_m(\mu)$ corresponding to minimizer of (skew) free energy by $Re\langle D\Psi_k, Q\Psi_k\rangle = 0$.

In addition to the above summary, we mention some remarks. First, we notice that if the bifurcation parameter μ is in reaction term f (not in D), $k_m(\mu) \approx k_c$ holds (Remark 6.3(1)). In many practical problems, μ is often an external parameter such as temparature, which is in the reaction term. Thus, recalling Conjecture 8.1, this suggests that a wavenumber to be selected is likely to deviate from k_c to the left in one-dimensional problems.

Moreover, we notice that it is necessary that $I_c > 0$ so that skew-gradient dissipative systems can admit gradient structure near a bifurcation point. On the other hand, when we change the sign of I_c, the underlying homogeneous steady state to be equipped with spatial structure becomes unstable, so that the spatial patterns lose their stability.

Furthermore, from the analytical results of section 7, we find that in two component gradient/skew-gradient dissipative systems, the Turing bifurcation point is nondegenerate and $E_c > 0$. This is a universal feature of two component systems. The degenerate case $E_c = 0$ can be observed for $n \geq 3$, and the complex spatial patterns near the bifurcation point in this case will be reported in a separate paper.

Finally, we mention further problems that should be studied in the framework of our theory. They seem to be significant steps to understand pattern selection principle in various fields of science. First, we must clarify what mechanism determines the selected wavenumber $k_s(\mu)$ and verify Conjecture 8.1. In particular, we must clarify conditions to guarantee the validity of inequality (8.3) in general n-component gradient/skew-gradient systems. This study is now in progress and is treated in a forthcoming paper [21].

In section 8, we consider what Turing pattern is to be selected among many for small random initial data. We are also interested in small initial data with compact support which is natural from a practical viewpoint. In this case, a local Turing pattern generated by a small local perturbation propagates into a linearly unstable uniform steady state [9, 34]. According to [9, 34], the wavenumber of spatially periodic structure of propagating patterns is determined by the (linear) marginal stability criterion, which yields a different one from $k_c, k_m(\mu)$, and $k_s(\mu)$ introduced here. Comparisons of our theory and theirs should be studied in the future.

In this paper, we consider pattern selection problems in a uniform environment. It is an ideal situation in which standard analysis such as asymptotic expansion can be easily performed. We are also interested in various cases that we meet in realistic problems. For example, the Turing parameter μ is dependent on spatial variable x, i.e., $\mu = \mu(x)$. This is known as a ramp problem [7, 19] which describes a bifurcation phenomena in a spatially nonuniform environment. In addition, a case in which μ is dependent on time variable t should be studied to compare our theory with real experiments.

Although pattern selection problems attracted the attention of many researchers, convincing mathematical results are very few. In fact, they are essentially to classify basins of attraction for stable steady states, and it turns out to be quite difficult in many cases. Our strategy to study pattern selection problems is not to determine the basins of attraction for steady states. The aim of our theory is rather to order steady states by using an appropriate (energy) functional which provides useful and practical information to determine what steady state is to be selected. We believe that our approach is a first step in understanding pattern selection mechanisms in various dissipative systems.

Acknowledgments. The author expresses his sincere gratitude to Professors Yasumasa Nishiura, Eiji Yanagida, and Kunimochi Sakamoto for their helpful advice. Moreover, the author would like to give appreciation to the referees who carefully read the manuscript.

REFERENCES

[1] H. Ando, Y. Sawada, H. Shimizu, and T. Sugiyama, *Pattern formation in hydra tissue without developmental gradients*, Developmental Biology, 133 (1989), pp. 405–414.
[2] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, 2nd ed., Springer-Verlag, New York, 1989.
[3] E. Ben-Jacob, H. Brand, G. Dee, L. Kramer, and J. S. Langer, *Pattern propagation in nonlinear dissipative systems*, Phys. D, 14 (1985), pp. 348–364.
[4] T. J. Bridges, *Multi-symplectic structures and wave propagation*, Math. Proc. Cambridge Philos. Soc., 121 (1997), pp. 147–190.
[5] P. Collet and J. P. Eckmann, *Instabilities and Fronts in Extended Systems*, Princeton University Press, Princeton, NJ, 1990.
[6] P. Collet and J. P. Eckmann, *The time dependent amplitude equation for the Swift–Hohenberg problem*, Comm. Math. Phys., 132 (1990), pp. 139–153.
[7] M. C. Cross and P. C. Hohenberg, *Pattern formation outside of equilibrium*, Rev. Mod. Phys., 65 (1993), pp. 851–1112.
[8] P. De Kepper, J. J. Perraud, B. Rudovics, and E. Dulos, *Experimental study of stationary Turing patterns and their interaction with travelling waves in a chemical system*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 4 (1994), pp. 1215–1231.
[9] U. Ebert and W. V. Saarloos, *Front propagation into unstable states*: *Universal algebraic convergence towards uniformly translating pulled fronts*, Phys. D, 146 (2000), pp. 1–99.
[10] W. Eckhaus, *Studies in non-linear stability theory*, Springer-Verlag, New York, 1965.
[11] W. Eckhaus, *The Ginzburg-Landau manifold is an attractor*, J. Nonlinear Sci., 3 (1993), pp. 329–348.
[12] A. Gierer and H. Meinhardt, *A theory of biological pattern formation*, Kybernetik, 12 (1972), pp. 30–39.
[13] H. Goldstein, *Classical Mechanics*, 2nd ed., Addison–Wesley, Reading, MA, 1980.
[14] M. Grillakis, J. Shatah, and W. Strauss, *Stability theory of solitary waves in the presence of symmetry* I, J. Funct. Anal., 74 (1987), pp. 160–197.
[15] H. Haken, *Synergetics*, 3rd ed., Springer-Verlag, Berlin, 1983.
[16] A. V. Harten, *On the validity of the Ginzburg-Landau equation*, J. Nonlinear Sci., 1 (1991), pp. 397–422.
[17] G. Iooss and D. D. Joseph, *Elementary stability and bifurcation theory*, Springer-Verlag, New York, 1980.
[18] S. Kondo and R. Asai, *A reaction-diffusion wave on the skin of the marine angel fish Pomacanthus*, Nature, 376 (1995), p. 765.
[19] L. Kramer and H. Riecke, *Wavelength selection in Rayleigh-Bénard convection*, Z. Phys. B Condensed Matter, 59 (1985), pp. 245–251.
[20] M. Kuwamura and E. Yanagida, *The Eckhaus and zigzag instability criteria in gradient/skew-gradient dissipative systems*, Phys. D, 175 (2003), pp. 185–195.
[21] M. Kuwamura, *A verification for marginal stability hypothesis for roll patterns*, in preparation.
[22] P. Manneville, *Dissipative Structures and Weak Turbulence*, Academic Press, Boston, 1990.
[23] A. Mielke and G. Schneider, *Attractors for modulation equations on unbounded domains—existence and comparison*, Nonlinearity, 8 (1995), pp. 743–768.
[24] H. Mori and Y. Kuramoto, *Dissipative Structures and Chaos*, Springer-Verlag, Berlin, 1998.
[25] J. D. Murray, *Mathematical Biology*, Springer-Verlag, Berlin, 1989.
[26] A. Newell and J. Whitehead, *Finite bandwidth, finite amplitude convection*, J. Fluid Mech., 38 (1969), pp. 279–303.
[27] G. Nicolis and I. Prigogine, *Self-Organization in Nonequilibrium Systems*, John Wiley and Sons, New York, 1977.
[28] Y. Nishiura, *Far-From-Equilibrium Dynamics*, Transl. Math. Monogr. 209, AMS, Providence, RI, 2002.
[29] Q. Ouyang and H. L. Swinney, *Transition from uniform state to hexagonal and striped Turing patterns*, Nature, 352 (1991), pp. 610–612.
[30] Y. Pomeau, S. Zaleski, and P. Manneville, *Dislocation motion in cellular structures*, Phys. Rev. A, 27 (1983), pp. 2710–2726.

[31] Y. Pomeau and P. Manneville, *Wavelength selection in cellular flows*, Phys. Lett. A, 75 (1980), pp. 296–299.
[32] A. Pocheau and V. Croquette, *Dislocation motion: A wavenumber selection mechanism in Rayleigh-Bénard convection*, J. Physique, 45 (1984), pp. 35–48.
[33] N. Rashevsky, *Mathematical Biophysics*, University of Chicago Press, Chicago, 1938.
[34] W. van Saarloos, *Front propagation into unstable states*, Phys. Rep., 386 (2003), pp. 29–222.
[35] A. M. Turing, *The chemical basis of morphogenesis*, Phil. Roy. Soc. B, 237 (1952), pp. 37–72.
[36] T. Yamaguchi, *Turing structure and pattern formation in reaction-diffusion systems*, J. National Institute of Material and Chemical Research, 5 (1997), pp. 151–164.
[37] L. Wolpert, *Positional information and the spatial pattern of cellular differentiation*, J. Theor. Biol., 25 (1969), pp. 1–47.

SIAM J. APPL. MATH.
Vol. 65, No. 2, pp. 644–666

STABILITY AND BIFURCATIONS IN NEURAL FIELDS WITH FINITE PROPAGATION SPEED AND GENERAL CONNECTIVITY*

FATIHCAN M. ATAY† AND AXEL HUTT‡

Abstract. A stability analysis is presented for neural field equations in the presence of finite propagation speed along axons and for a general class of connectivity kernels and synaptic properties. Sufficient conditions are given for the stability of equilibrium solutions. It is shown that the propagation delays play a significant role in nonstationary bifurcations of equilibria, whereas the stationary bifurcations depend only on the connectivity kernel. In the case of nonstationary bifurcations, bounds are determined on the frequencies of the resulting oscillatory solutions. A perturbative scheme is used to calculate the types of bifurcations leading to spatial patterns, oscillations, and traveling waves. For high propagation speeds a simple method is derived that allows the determination of the bifurcation type by visual inspection of the Fourier transforms of the kernel and its first moment. Results are numerically illustrated on a class of neurologically plausible systems with combinations of Gaussian excitatory and inhibitory connections.

Key words. synaptic networks, nonlocal interaction, delay, bifurcations, spatiotemporal patterns, traveling waves

AMS subject classifications. 92C20, 34K99, 37N25, 37G10

DOI. 10.1137/S0036139903430884

1. Introduction.

In recent years, there has been growing interest in the mechanisms of spatiotemporal activity in neural tissue. In this field, applications of various experimental techniques [37, 21, 39, 41] revealed formations of different spatial patterns, traveling waves, and pulses [28, 43, 48], standing pulses (e.g., [18]), or irregular spatial patterns [2, 40]. Since neural tissue exhibits multiscale properties in space and time, the analysis of such activity represents a challenging task. However, reduced biological models at fixed scales in time and space simplify the analysis and allow for analytical treatments (see, e.g., [5, 12, 42] for review). In this context, a well-known approach is to focus on neuronal ensembles [46, 47, 29], which allows for the successful reconstruction of empirical data measured on a macroscopic scale [24, 30, 35, 34, 25].

On a small spatial level ($\sim 50\mu m$), model neurons may consist of two compartments: synapses, which convert incoming action potentials to postsynaptic potentials, and a trigger zone, where these potentials sum up and are reconverted to outgoing action potentials. Due to the large spatial density of neurons ($\sim 10^4$ neurons/mm^3), one might consider ensemble activity on a larger spatial scale (> 1 mm), obtaining a coarse-grained description in space and time [46]. Consequently, macroscopic state variables of neuronal ensembles are mean pulse rates $P(x,t)$ and mean postsynaptic potentials $V(x,t)$, with x and t denoting the space and time coordinates, respectively. In the following, all quantities are meant to represent means of microscopic quantities.

Since the link between the microscopic description and the level of neural ensembles has been established in several previous works (e.g., [46, 38, 5, 4]), we only briefly

*Received by the editors June 30, 2003; accepted for publication (in revised form) April 6, 2004; published electronically January 5, 2005.
http://www.siam.org/journals/siap/65-2/43088.html
†Max Planck Institute for Mathematics in the Sciences, Inselstr. 22, Leipzig 04103, Germany (atay@member.ams.org).
‡Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, Berlin 10117, Germany (hutt@wias-berlin.de). The work of this author was supported by the DFG research center "Mathematics for key technologies" (FZT86), Berlin, Germany.

outline the basic mechanisms of activity conversion in neuronal fields. At chemical passive synapses, incoming pulse activity $J(x,t)$ is converted to postsynaptic potentials by convolution with an impulse response function $h(t)$, yielding

$$V(x,t) = \int_{-\infty}^{t} h(t-\tau)J(x,\tau)d\tau.$$

Since neuronal fields exhibit nonlocal interactions via axonal connections between synapses, incoming pulse activity obeys

$$J(x,t) = \beta \int_{\Omega} K(x,y)P(y,t-\Delta(x,y))dy \; + \; E(x,t),$$

where Ω is an appropriate spatial domain, K is the connectivity kernel, $\beta > 0$ is a scaling factor, and E is an additional external input. In the case of undamped axonal pulse propagation with finite velocity v and no additional constant delay, we get $\Delta(x,y) = |x-y|/v$. Essentially, the chain of activity conversion closes by the conversion of postsynaptic potentials to pulse rates

$$P(x,t) = S[V(x,t)],$$

where S is called the transfer function, which is taken to have a sigmoidal shape in most works. Considering all conversions, we obtain the integral equation

$$V(x,t) = \beta \int_{-\infty}^{t} \int_{\Omega} \left(h(t-\tau)K(x,y)S[V(y,\tau-|x-y|/v)] + E(x,\tau)\right)\, dy\, d\tau.$$

We then recast the impulse response function as a Green's function and thus stipulate

$$Lh(t) = \delta(t),$$

introducing a temporal differential operator L. Finally, we assume a homogeneous field where the connectivity $K(x,y)$ depends only on the distance $|x-y|$, and so we replace $K(x,y)$ by an even function $K(x-y)$. Hence, the final equation has the form

$$L(\partial/\partial t)V(x,t) = \beta \int_{\Omega} K(x-y)S(V(y,t-|x-y|/v))\, dy + E(x,t), \tag{1.1}$$

where L is a polynomial and $L(\partial/\partial t)$ denotes a temporal differentiation operator with constant coefficients. We shall refer to (1.1) as an nth-order system, where $n \geq 1$ is the order of L.

The model (1.1) has been treated in the literature in several contexts and with different choices for L. In most studies the effect of transmission speed has been neglected by letting $v = \infty$ in the model, the justification being that the signal propagation is sufficiently fast or the spatial scales of the problem are small [32]. Some recent works [5, 15, 19, 31, 23, 8] have addressed the case of finite v by numerical investigations for particular choices of the kernel K. Our aim is to give an analytical treatment of the effects of finite transmission speeds for general K and L, in relation to the stability and bifurcation of the equilibrium solutions.

In the next section we consider the equilibrium solutions of (1.1) and give a sufficient condition for their stability. Section 3 discusses the types of dynamics that may emerge when stability is lost. Here it is shown that the transmission speed needs to

be smaller than a certain threshold in order to have oscillatory bifurcations in first- and second-order systems. Furthermore, bounds are calculated for the frequencies of the oscillatory solutions. In section 4, a perturbative analysis is used to compute the bifurcating solutions, and a graphical method is given to determine the possible bifurcations for a given kernel. Applications to kernels derived from Gaussian distributions are presented in section 5, and the paper concludes with a discussion of the results.

2. Stability of equilibrium solutions. For the rest of the paper we make the following assumptions regarding (1.1). For the spatial domain we assume $\Omega = \mathbf{R}$, although the results remain valid virtually without modification when Ω is a circle. The polynomial L is stable; i.e., all its roots have negative real parts. The kernel $K : \mathbf{R} \to \mathbf{R}$ is continuous, integrable, and even; that is, $K(-z) = K(z)$ for all $z \in \mathbf{R}$. Finally, the transfer function $S : \mathbf{R} \to \mathbf{R}$ is differentiable and monotone increasing.

It is often convenient to normalize the time and space in (1.1). For instance, if l and τ are some characteristic length and time of the physical problem, then one can define $\bar{t} = t/\tau$, $\bar{x} = x/l$, $\bar{V}(\bar{x}, \bar{t}) = V(l\bar{x}, \tau\bar{t})$, $\bar{E}(\bar{x}, \bar{t}) = E(l\bar{x}, \tau\bar{t})$, $\bar{L}(\partial/\partial\bar{t}) = \tau^n L(\tau^{-1}\partial/\partial\bar{t})$, $\bar{K}(\bar{z}) = K(l\bar{z})$, and $\bar{v} = \tau v/l$ so that (1.1) becomes

$$\bar{L}(\partial/\partial\bar{t})\bar{V}(\bar{x}, \bar{t}) = l\tau^n\beta \int_\Omega \bar{K}(\bar{x} - \bar{y}) S(\bar{V}(\bar{y}, \bar{t} - |\bar{x} - \bar{y}|/\bar{v}))\, d\bar{y} + \bar{E}(\bar{x}, \bar{t}),$$

which has the same form as (1.1). A common choice for characteristic time is $\tau^n = 1/L(0)$, in which case $\bar{L}(0) = 1$. Thus, without loss of generality we consider (1.1) with the assumption that $L(0) = 1$. Most studies of neuronal fields assume first- or second-order time derivatives in (1.1). To address these models in a unified manner, we shall often refer to the following specific form:

$$L(\lambda) = \eta\lambda^2 + \gamma\lambda + 1, \quad \eta = 0 \text{ or } 1,\ \gamma > 0, \tag{2.1}$$

although certain results will be stated for arbitrary order stable polynomials L.

For a constant input $E(x, t) \equiv E^*$, an equilibrium solution $V(x, t) \equiv V^*$ satisfies

$$V^* = \beta \int_{-\infty}^{\infty} K(x - y) S(V^*)\, dy + E^*. \tag{2.2}$$

Let

$$\kappa = \int_{-\infty}^{\infty} K(z)\, dz = 2\int_0^{\infty} K(z)\, dz. \tag{2.3}$$

Then (2.2) can be written as

$$f(V^*) \stackrel{\text{def}}{=} V^* - \kappa\beta S(V^*) = E^*. \tag{2.4}$$

If S is bounded, then $f : \mathbf{R} \to \mathbf{R}$ is surjective; thus (2.4) has a solution V^* for any $E^* \in \mathbf{R}$. The uniqueness of V^* depends on the sign of κ and the shape of S. If S is positive and increasing on $\mathbf{R}$, such as a sigmoid function, and if $\kappa \le 0$, then f is increasing and hence also injective, in which case the solution V^* is unique. On the other hand, if $\kappa > 0$, then there may be multiple equilibria, as (2.4) can have more than one solution V^* for a given E^*.

The stability of the equilibrium solution V^* is determined by the linear variational equation

$$L(\partial/\partial t)u(x,t) = \alpha \int_{-\infty}^{\infty} K(x-y)u(y,t-|x-y|/v)\,dy, \tag{2.5}$$

where $u(x,t) = V(x,t) - V^*$ and $\alpha = \beta S'(V^*) \geq 0$. We shall use α as a bifurcation parameter in the following sections. Using the ansatz $u(x,t) = e^{\lambda t}\varphi(x)$ in (2.5) one obtains

$$L(\lambda)\varphi(x) = \alpha \int_{-\infty}^{\infty} K(x-y)\exp(-\lambda|x-y|/v)\varphi(y)\,dy. \tag{2.6}$$

Thus φ is an eigenfunction of an integral operator. Due to the difference kernel the eigenfunctions have the form $\varphi(x) = e^{ikx}$ for some $k \in \mathbf{R}$, and substituting into (2.6) followed by a change of variables $z = x - y$ in the integral gives

$$L(\lambda) = \alpha \int_{-\infty}^{\infty} K(z)\exp(-\lambda|z|/v)\exp(-ikz)\,dz. \tag{2.7}$$

The integral above is the Fourier transform of the function $K_\lambda(z) = K(z)\exp(-\lambda|z|/v)$ (up to a multiplicative factor, depending on which definition one uses), which is also equal to its cosine transform since $K_\lambda(z)$ is an even function of z. The dispersion relation (2.7) between the temporal and spatial modes λ and k is in general difficult to solve explicitly. A notable exception is the case of instantaneous information transmission, since when $v = \infty$, the right-hand side of (2.7) is independent of λ. In this paper we are interested in the effects of finite transmission speeds.

The solutions (λ, k) of (2.7) correspond to the perturbations $u(x,t) = e^{\lambda t}e^{ikx}$ about the equilibrium solution, which grow or decay in time depending on whether $\operatorname{Re}\lambda$ is positive or negative, respectively, thus determining the stability of V^*. We give sufficient conditions for asymptotic stability.

THEOREM 2.1. *Let $c = \alpha\int_{-\infty}^{\infty}|K(z)|\,dz$. If*

$$c < \min_{\omega\in\mathbf{R}} |L(i\omega)|, \tag{2.8}$$

then V^* is asymptotically stable. In particular, if $L(\lambda) = \lambda + 1$, then the condition

$$c < 1 \tag{2.9}$$

is sufficient for the asymptotic stability of V^*. If $L(\gamma) = \lambda^2 + \gamma\lambda + 1$ with $\gamma > 0$, then V^* is asymptotically stable provided that the condition

$$\frac{\gamma^2}{2} > 1 - \sqrt{1-c^2} \tag{2.10}$$

holds, in addition to (2.9).

The following lemma will be useful in the proof of the theorem.

LEMMA 2.2. *Let $L(\lambda)$ be a polynomial whose roots have nonpositive real parts. Then*

$$|L(\sigma + i\omega)| \geq |L(i\omega)|$$

for all $\sigma \geq 0$ and $\omega \in \mathbf{R}$.

Proof. If λ_k denote the roots of L, then $L(\lambda) = (\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_n)$, where n is the order of L. Thus

$$\begin{aligned}|L(\sigma + i\omega)| &= \prod_{k=1}^{n} |\sigma + i\omega - \lambda_k| \\ &= \prod_{k=1}^{n} \left((\sigma - \mathrm{Re}[\lambda_k])^2 + (\omega - \mathrm{Im}[\lambda_k])^2\right)^{1/2}.\end{aligned}$$

By assumption, $\sigma \geq 0$ and $\mathrm{Re}[\lambda_k] \leq 0$ for all k, so

$$\begin{aligned}|L(\sigma + i\omega)| &\geq \prod_{k=1}^{n} \left((-\mathrm{Re}[\lambda_k])^2 + (\omega - \mathrm{Im}[\lambda_k])^2\right)^{1/2} \\ &= \prod_{k=1}^{n} |i\omega - \lambda_k| \\ &= |L(i\omega)|. \qquad \square\end{aligned}$$

Proof of Theorem 2.1. In the ansatz $u(x,t) = e^{\lambda t}e^{ikx}$, let $\lambda = \sigma + \mathrm{i}\omega$, where σ and ω are real numbers. We will prove that $\sigma < 0$ if (2.8) holds. Suppose by way of contradiction that (2.8) holds but $\sigma \geq 0$. From the dispersion relation (2.7) it follows that

$$\begin{aligned}|L(\sigma + i\omega)| &= \alpha \left| \int_{-\infty}^{\infty} K(z) \exp(-(\sigma + i\omega)|z|/v) \exp(-ikz)\, dz \right| \\ &\leq \alpha \int_{-\infty}^{\infty} |K(z)|\, |\exp(-(\sigma + i\omega)|z|/v)|\, dz \\ &\leq \alpha \int_{-\infty}^{\infty} |K(z)| dz = c. \qquad (2.11)\end{aligned}$$

On the other hand, by Lemma 2.2,

$$|L(i\omega)| \leq |L(\sigma + i\omega)|,$$

which together with (2.11) implies

$$|L(i\omega)| \leq c$$

for some $\omega \in \mathbf{R}$. This, however, contradicts (2.8). Thus $\sigma < 0$, and the equilibrium solution is asymptotically stable. This proves the first statement of the theorem. In the specific case when L is given by $L(\lambda) = \lambda + 1$, one has $|L(i\omega)|^2 = 1 + \omega^2$. Hence if (2.9) is satisfied, then

$$c^2 < 1 \leq 1 + \omega^2 = |L(i\omega)|^2 \quad \text{for all } \omega \in \mathbf{R},$$

which is a sufficient condition for stability by (2.8). Similarly, suppose that L has the form $L(\lambda) = \lambda^2 + \gamma\lambda + 1$ and that (2.9) and (2.10) are satisfied. Then

$$|L(i\omega)|^2 = (1 - \omega^2)^2 + (\gamma\omega)^2.$$

Now consider the function

$$\begin{aligned}g(\omega) &\stackrel{\mathrm{def}}{=} |L(i\omega)|^2 - c^2 \\ &= \omega^4 + (\gamma^2 - 2)\omega^2 + (1 - c^2). \qquad (2.12)\end{aligned}$$

If $\gamma^2 \geq 2$, then $g(\omega)$ is positive for all ω by (2.9). On the other hand, if $\gamma^2 < 2$, then by (2.10)

$$0 < 2 - \gamma^2 < 2\sqrt{1 - c^2},$$

implying that the discriminant $(\gamma^2 - 2)^2 - 4(1 - c^2)$ is negative; thus g has no real roots. Thus, in either case, $g(\omega)$ is positive, or, equivalently, $c < |L(i\omega)|$ for all ω, and stability again follows by the first statement of the theorem. □

3. Bifurcations. When $\alpha = 0$, the eigenvalues λ are simply given by the roots of L, so that $\operatorname{Re}\lambda < 0$ by the assumption that L is a stable polynomial, and the equilibrium point is asymptotically stable. As α is increased, stability can be lost if an eigenvalue λ crosses the imaginary axis. At the critical transition there is an eigenvalue of the form $\lambda = i\omega$, $\omega \in \mathbf{R}$, and the dispersion relation (2.7) has the form

$$L(i\omega) = \alpha \int_{-\infty}^{\infty} K(z) \exp(-i(kz + \omega|z|/v))\, dz. \tag{3.1}$$

The possibilities for the resulting behavior when α is near such a critical value can then be qualitatively classified as follows:

I. Stationary bifurcations
   a. $\omega = 0$ and $k = 0$: bifurcation to a spatially and temporally constant solution.
   b. $\omega = 0$ and $k \neq 0$: bifurcation to a spatially periodic solution which is constant in time, leading to spatial patterns (Turing modes).

II. Nonstationary bifurcations
   a. $\omega \neq 0$ and $k = 0$: Hopf bifurcation to periodic oscillations of a spatially uniform solution.
   b. $\omega \neq 0$ and $k \neq 0$: bifurcation to traveling waves, with wave speed equal to ω/k.

The conditions for stationary bifurcations are easily characterized by the relation (3.1), recalling the assumption that $L(0) = 1$. Thus for case Ia one has

$$1 = \alpha \int_{-\infty}^{\infty} K(z)\, dz = \alpha\kappa \tag{3.2}$$

with κ as defined in (2.3). This is only possible if $\kappa > 0$ and is the mechanism for appearance of multiple equilibrium solutions of (2.2). Similarly, the condition (2.7) for case Ib is

$$1 = \alpha \int_{-\infty}^{\infty} K(z) \exp(ikz)\, dz = \alpha \hat{K}(k), \quad k \neq 0, \tag{3.3}$$

where $\hat{K}$ denotes the Fourier transform of K. As α is increased from zero, the first mode that becomes unstable in the linearized equation is expected to give an indication of what would be observed in the full nonlinear system (1.1). Hence, spatial patterns are typically observed as bifurcations from equilibria if a nonzero k is the first mode that loses stability. From (3.3) it follows that a necessary condition for this is that the maximum value of the Fourier transform of K is positive and occurs at a nonzero frequency k.

It is clear from (3.2) and (3.3) that stationary bifurcations are independent of the order of the temporal differentiation operator L or the transmission speed v. Their

analysis only involves the properties of the Fourier transform of the kernel function. On the other hand, L and v turn out to be crucial in nonstationary bifurcations. Indeed, our next result shows that a sufficiently small transmission speed is actually a necessary condition for nonstationary bifurcations in first- and second-order systems.

THEOREM 3.1. *Suppose $L(\lambda) = \eta\lambda^2 + \gamma\lambda + 1$, where η may possibly be zero. If*

$$v > \frac{\alpha}{|\gamma|} \int_{-\infty}^{\infty} |zK(z)|\, dz, \tag{3.4}$$

then (2.5) *has no solutions of the form $u(x,t) = \exp i(\omega t + kx)$ with ω real and nonzero.*

Proof. From the dispersion relation (2.7),

$$\begin{aligned} L(\lambda) &= \alpha \int_{-\infty}^{\infty} K(z) \exp(-\lambda|z|/v)(\cos kz - i \sin kz)\, dz \\ &= \alpha \int_{-\infty}^{\infty} K(z) \exp(-\lambda|z|/v) \cos kz\, dz \end{aligned}$$

since the function $K(z)\exp(-\lambda|z|/v)$ is even in z. Separating the real and imaginary parts of the above expression at the bifurcation value $\lambda = i\omega$ gives

$$\operatorname{Re} L(i\omega) = \alpha \int_{-\infty}^{\infty} K(z) \cos(\omega z/v) \cos(kz)\, dz, \tag{3.5}$$

$$\operatorname{Im} L(i\omega) = -\alpha \int_{-\infty}^{\infty} K(z) \sin(\omega|z|/v) \cos(kz)\, dz. \tag{3.6}$$

Suppose $L(\lambda) = \eta\lambda^2 + \gamma\lambda + 1$. Then $\operatorname{Im} L(i\omega) = \gamma\omega$, and (3.6) implies

$$\begin{aligned} |\gamma\omega| &= \alpha \left| \int_{-\infty}^{\infty} K(z) \sin(\omega|z|/v) \cos(kz)\, dz \right| \\ &\le \alpha \int_{-\infty}^{\infty} |K(z) \sin(\omega z/v)|\, dz \\ &\le \alpha \int_{-\infty}^{\infty} |K(z)\omega z/v|\, dz, \end{aligned}$$

where we have used the estimate $|\sin(x)| \le |x|$ for all $x \in \mathbf{R}$. If $\omega \neq 0$, then $|\omega|$ may be cancelled in the last inequality to yield

$$|\gamma| \le \frac{\alpha}{v} \int_{-\infty}^{\infty} |zK(z)|\, dz.$$

This, however, contradicts the assumption (3.4). Hence $\omega = 0$, which proves the theorem. ∎

We note that the above result is valid for first- and second-order systems; in higher order systems, bifurcation values $\lambda = i\omega \neq 0$ may occur even with $v = \infty$ [12].

For bifurcating oscillatory solutions, it is possible to put a priori bounds on the possible values of the frequencies ω in terms of the kernel function and the operator L, as given by the next result.

THEOREM 3.2. *Let c be as defined in Theorem* 2.1. *Then there exists $B > 0$, depending only on L and c, such that*

$$|\omega| \le B \tag{3.7}$$

whenever $u(x,t) = \exp i(\omega t + kx)$, $\omega, k \in \mathbf{R}$, *is a solution of* (2.5). *Furthermore, if* $c < 1$, *then there exists* $A > 0$, *depending only on* L *and* c, *such that*

$$0 < A \leq |\omega|. \tag{3.8}$$

In particular, if $L(\lambda) = \lambda + 1$, *then*

$$\omega^2 \leq c^2 - 1, \tag{3.9}$$

and if $L(\lambda) = \lambda^2 + \gamma\lambda + 1$, *then*

$$\begin{array}{cc} (1 - \frac{1}{2}\gamma^2) - \delta \leq \omega^2 \leq (1 - \frac{1}{2}\gamma^2) + \delta & \text{if} \quad 0 \leq c < 1, \\ 0 \leq \omega^2 \leq (1 - \frac{1}{2}\gamma^2) + \delta & \text{if} \quad c \geq 1, \end{array} \tag{3.10}$$

where $\delta = \sqrt{(1 - \frac{1}{2}\gamma^2)^2 - 1 + c^2}$.

Remark. The existence of a solution of the form $u(x,t) = \exp i(\omega t + kx)$ implies that the equilibrium point is not asymptotically stable. It is then a consequence of Theorem 2.1 that the right sides of the inequalities in (3.9) and (3.10) are nonnegative.

Proof of Theorem 3.2. If $\lambda = i\omega$ satisfies the dispersion relation (2.7) for some k, then

$$|L(i\omega)| \leq \alpha \int_{-\infty}^{\infty} |K(z)|\, dz = c. \tag{3.11}$$

Since $|L(i\omega)| \to \infty$ as $\omega \to \pm\infty$ for any nonconstant polynomial L, the above inequality implies an upper bound B on $|\omega|$, which proves (3.7). For the particular case when $L(\lambda) = \lambda + 1$, (3.11) gives

$$|L(i\omega)|^2 = \omega^2 + 1 \leq c^2,$$

proving (3.9). Similarly, for $L(\lambda) = \lambda^2 + \gamma\lambda + 1$, (3.11) yields

$$|L(i\omega)|^2 = \omega^4 + (\gamma^2 - 2)\omega^2 + 1 \leq c^2. \tag{3.12}$$

If we let $u = \omega^2$, then the inequality above is equivalent to saying that possible values of $u \geq 0$ are those which render the function

$$h(u) \stackrel{\text{def}}{=} u^2 + (\gamma^2 - 2)u + (1 - c^2)$$

negative or zero. This is only possible if h has at least one root in the interval $[0, \infty)$, implying that the discriminant $(\gamma^2 - 2)^2 - 4(1 - c^2)$ is nonnegative. Letting $\delta = \sqrt{(1 - \frac{1}{2}\gamma^2)^2 - 1 + c^2}$, the roots of h can be written as $(1 - \frac{1}{2}\gamma^2) \pm \delta$. Thus $h(\omega^2) \leq 0$ for ω^2 satisfying

$$(1 - \tfrac{1}{2}\gamma^2) - \delta \leq \omega^2 \leq (1 - \tfrac{1}{2}\gamma^2) + \delta. \tag{3.13}$$

It remains to ensure that the interval above is a subset of $[0, \infty)$. If $c < 1$, then it is easy to see that both roots of h are nonnegative. For if the smaller root is negative, we have

$$0 > (1 - \tfrac{1}{2}\gamma^2) - \delta > (1 - \tfrac{1}{2}\gamma^2) - |1 - \tfrac{1}{2}\gamma^2|,$$

so $(1-\frac{1}{2}\gamma^2)<0$. But then both the conditions (2.9) and (2.10) are satisfied, and by Theorem 2.1 $\lambda = i\omega$ cannot be a solution to (2.7). On the other hand, if $c \geq 1$, then

$$(1-\tfrac{1}{2}\gamma^2) - \delta \leq (1-\tfrac{1}{2}\gamma^2) - |1-\tfrac{1}{2}\gamma^2| \leq 0$$

and

$$(1-\tfrac{1}{2}\gamma^2) + \delta \geq (1-\tfrac{1}{2}\gamma^2) + |1-\tfrac{1}{2}\gamma^2| \geq 0.$$

So, in this case the lower bound on ω^2 in (3.13) can be replaced by zero. This establishes (3.10). Finally, to prove (3.8) for arbitrary L assume that $c<1$. Then $1 = L(0) > c$. By the continuity of L there exists $A>0$ such that $|L(i\omega)| > c$ whenever $|\omega| \leq A$. Since (3.11) is not satisfied, (2.5) does not have a solution of the form $\exp i(\omega t + kx)$ with $|\omega| \leq A$, which completes the proof. ∎

4. Perturbative analysis. In order to study the type of bifurcations that may arise in a given situation, the dispersion relation (3.1) needs to be solved for ω and k. However, explicit solutions are difficult to obtain for general kernel functions. The results of the previous sections imply that in the absence of delays, one has a simpler case, where nonstationary bifurcations do not exist in first- and second-order systems. Consequently, the role of delays can be systematically examined by following the changes in the bifurcation structure as the value of the transmission speed is decreased from infinity. Hence we introduce the parameter $\varepsilon = 1/v$ and consider the change in dynamics as ε is increased from zero. This leads to an approximation scheme that provides valuable insight into the effects of axonal delays in the dynamics of the system.

Consider the power series estimate

$$\exp(-\lambda|z|/v) = \sum_{m=0}^{N} \frac{(-\lambda|z|/v)^m}{m!} + \mathcal{O}(v^{-(N+1)}).$$

Substitution in the dispersion relation (2.7) at the bifurcation value $\lambda = i\omega$ gives a finite series in powers of $\varepsilon = 1/v$,

$$\begin{aligned} L(i\omega) &= \alpha \int_{-\infty}^{\infty} K(z)\exp(-ikz)\left[\sum_{m=0}^{N}\frac{(-i\varepsilon\omega|z|)^m}{m!} + \mathcal{O}(\varepsilon^{N+1})\right] dz \\ &= \alpha \sum_{m=0}^{N} \frac{(-i\varepsilon\omega)^m}{m!}\hat{K}_m(k) + \mathcal{O}(\varepsilon^{N+1}), \end{aligned} \tag{4.1}$$

where the $\hat{K}_m$ denote the transforms of the moments of K:

$$\hat{K}_m(k) = \int_{-\infty}^{\infty} |z|^m K(z)\exp(-ikz)\,dz = 2\int_0^{\infty} z^m K(z)\cos(kz)\,dz \tag{4.2}$$

and the integrals are assumed to exist. Separating the real and imaginary parts of (4.1) then yields

$$\alpha^{-1}\operatorname{Re} L(i\omega) = \hat{K}_0(k) - \frac{\varepsilon^2}{2}\omega^2\hat{K}_2(k) + \frac{\varepsilon^4}{24}\omega^4\hat{K}_4(k) - \cdots, \tag{4.3}$$

$$\alpha^{-1}\operatorname{Im} L(i\omega) = -\varepsilon\omega\hat{K}_1(k) + \frac{\varepsilon^3}{6}\omega^3\hat{K}_3(k) - \frac{\varepsilon^5}{120}\omega^5\hat{K}_5(k) + \cdots. \tag{4.4}$$

The number of terms needed for the above series to be useful depends on the value of ε as well as the shape of the kernel K. If K is highly concentrated near the origin, then a few terms are sufficient. To make this precise, suppose that K is of exponential order, which is a reasonable assumption in most practical situations. In other words, suppose there exist positive numbers κ_1 and κ_2 such that

$$|K(z)| \leq \kappa_1 \exp(-\kappa_2|z|) \quad \text{for all } z \in \mathbf{R}.$$

Then, by (4.2),

$$\begin{aligned}\left|\hat{K}_m(k)\right| &\leq \int_{-\infty}^{\infty} |z|^m \kappa_1 \exp(-\kappa_2|z|)\, dz = 2\kappa_1 \int_0^{\infty} z^m \exp(-\kappa_2 z)\, dz \\ &= 2\kappa_1 \kappa_2^{-(m+1)} \Gamma(m+1) = 2\kappa_1 \kappa_2^{-(m+1)} m!,\end{aligned}$$

so the mth term in the series (4.1) is bounded in absolute value by

$$2\frac{\kappa_1}{\kappa_2}\left(\frac{\varepsilon|\omega|}{\kappa_2}\right)^m \leq 2\frac{\kappa_1}{\kappa_2}\left(\frac{B}{\kappa_2}\varepsilon\right)^m,$$

where we have used Theorem 3.2 to bound the values of ω. Hence, in case of small ε (large transmission speed) or B (e.g., small α) or a large value of κ_2 (fast decay of K away from the origin), the finite series has increased accuracy. We assume that at least one of these conditions is satisfied so that a small number of terms suffices to determine the general behavior.

In order to observe the qualitative effects of finite transmission speed, we thus neglect third- and higher order terms in ε in the series (4.1). Then, for L given by

$$L(\lambda) = \eta\lambda^2 + \gamma\lambda + 1, \quad \eta = 0 \text{ or } 1,\ \gamma > 0,$$

(4.3)–(4.4) become

$$\alpha^{-1}(1 - \eta\omega^2) = \hat{K}(k) - \tfrac{1}{2}\varepsilon^2\omega^2\hat{K}_2(k), \tag{4.5}$$

$$\alpha^{-1}\gamma\omega = -\varepsilon\omega\hat{K}_1(k), \tag{4.6}$$

where we have substituted the more conventional notation $\hat{K}$ for the Fourier transform $\hat{K}_0$ of the kernel. For stationary bifurcations ($\omega = 0$), one obtains from the first equation that

$$\hat{K}(k) = 1/\alpha, \tag{4.7}$$

which is the same as the conditions (3.2)–(3.3) given by exact calculation. For a nonstationary bifurcation, $\omega \neq 0$, so (4.6) implies that

$$\hat{K}_1(k^*) = -\gamma/\varepsilon\alpha. \tag{4.8}$$

If $\hat{K}_1(k)$ assumes negative values, then it has a minimum since it is continuous and tends to zero as $k \to \pm\infty$; this minimum value corresponds to the first mode that loses stability as ε or α is increased. More precisely, if

$$k^* = \min_k \hat{K}_1(k) = \min_k \int_{-\infty}^{\infty} |z| K(z) \exp(-ikz)\, dz \tag{4.9}$$

exists and $\hat{K}_1(k^*) < 0$, then k^* is the sought solution of (4.8). Substituting k^* into (4.5) gives

$$\omega^2 = \frac{\alpha\hat{K}(k^*) - 1}{\frac{1}{2}\alpha\varepsilon^2\hat{K}_2(k^*) - \eta}, \tag{4.10}$$

which has a solution for ω whenever the right-hand side is nonnegative. This gives a simple procedure to calculate the pairs (ω, k) satisfying the dispersion relation and corresponding to the bifurcating solution $\exp(\omega t + kx)$.

It remains to determine what type of bifurcation actually occurs. This depends on the mode by which the equilibrium solution, which is stable for $\alpha = 0$, loses its stability as the bifurcation parameter α is increased. The procedure described in the above paragraph gives a simple graphical method. Thus if one plots the curves $\hat{K}(k)$ and $-\hat{K}_1(k)/\gamma v$ in the same graph and thinks of $1/\alpha$ as a horizontal line being lowered from $+\infty$, then the first intersection point specifies the bifurcation type. If the horizontal line touches the graph of $\hat{K}(k)$ first, then (4.7) is satisfied and a stationary bifurcation occurs. If, on the other hand, it touches $-\hat{K}_1(k)/\gamma v$ first, then (4.8) is satisfied and a nonstationary bifurcation occurs. Furthermore, the value of k at the intersection point being zero or nonzero specifies whether the bifurcating solution is spatially constant or not, respectively. It is worthwhile to note that the types of bifurcations that can occur depends only the extremal values of $\hat{K}$ and $\hat{K}_1$ and not on the exact shapes of their graphs. This observation has two important consequences. First, the bifurcation structure depends on some general qualities of the kernel and not on its precise shape. And second, although our analysis is based on an approximation scheme, the qualitative conclusions regarding the type of bifurcations are generally robust, except for some degenerate cases, such as when the maximum values of $\hat{K}(k)$ and $-\hat{K}_1(k)/\gamma v$ are equal.

An example for the investigation of possible bifurcations is illustrated in Figure 1 for some typical kernel functions representing the possibilities for different types of inhibitory and excitatory interaction within the field. For each kernel type in the first column of the figure, the corresponding graphs of $\hat{K}(k)$ and $-\hat{K}_1(k)/\gamma v$ are plotted in the second column. By the argument outlined above, the possible bifurcations for each type of kernel can be directly read off from the graphs in the second column. The actual graphs in the figure are calculated from Gaussian distributions; however, it is clear that small variations in the graphs do not change the bifurcation types. In this way, it is possible to draw some general conclusions concerning different interaction kernels.

The analysis presented in this section is useful for understanding the relationship between the interactions within the field and the resulting dynamics. The stationary bifurcations of equilibria are determined by the Fourier transform $\hat{K}$ of the connectivity kernel. The nonstationary bifurcations, on the other hand, are characterized by the transforms of the moments of the kernel. For first- and second-order systems this characterization can be reduced to the consideration of the single term $\hat{K}_1/\gamma v$, over the parameter ranges where the approximation scheme is justified. Outside of this range, e.g., for very low transmission speeds, more terms need to be considered in the series (4.1), together with a numerical solution of the system (4.3)–(4.4). Nevertheless, already in the term $\hat{K}_1/\gamma v$ one can see the ingredients responsible for nonstationary bifurcations: the operator L (through γ) representing the local temporal behavior, the kernel (through $\hat{K}_1$) representing spatial interaction, and the transmission speed v connecting the two aspects of the dynamics. Figure 1 gives a summary of the bifurcations resulting from the interplay of these elements. In the next section we

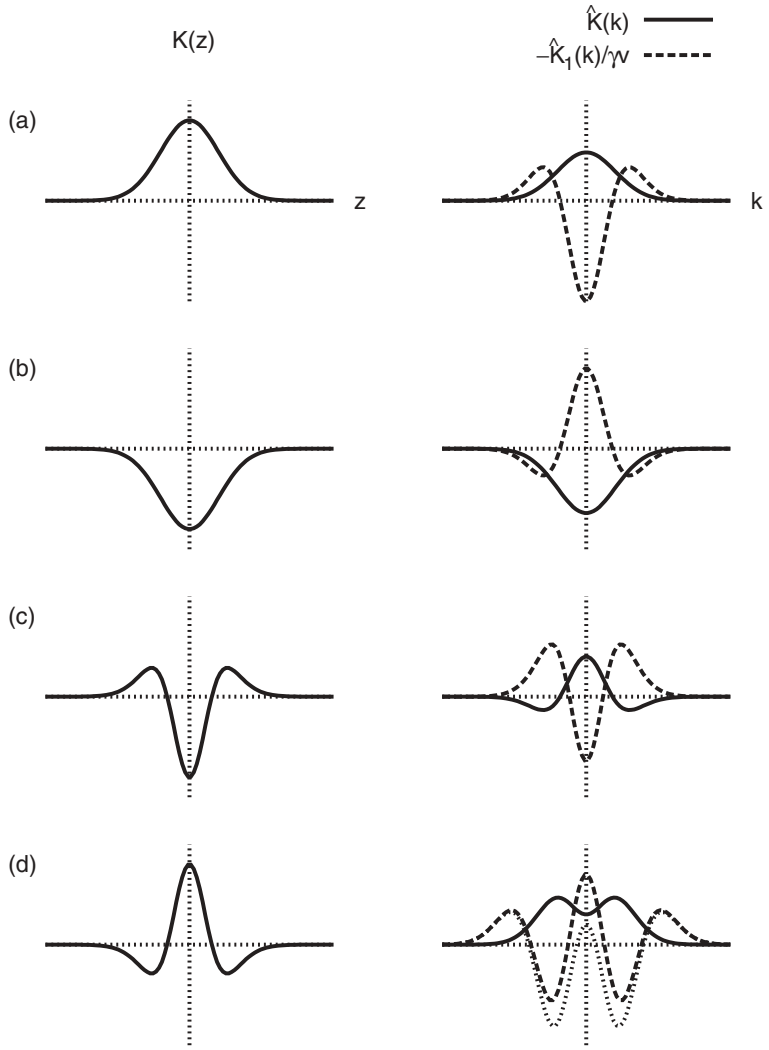


FIG. 1. *Typical interaction kernels and possible bifurcation types. The first column shows the kernels, with the corresponding Fourier transforms in the second column. The maxima of* $\hat{K}$ *and* $-\hat{K}_1/\gamma v$*, respectively, determine the stationary and oscillatory bifurcations, the largest peak giving the actual bifurcation taking place as* α *is increased. Hence, depending on the value of* γv*, some typical cases are* (a) *an excitatory field, possible bifurcations* Ia *and* IIb*;* (b) *an inhibitory field, possible bifurcation* IIa*;* (c) *local inhibition and lateral excitation, possible bifurcations* Ia *and* IIb*;* (d) *local excitation and lateral inhibition, possible bifurcations* Ib *and* IIa *or* IIb*. In the last subfigure, two distinct possibilities for* $-\hat{K}_1/\gamma v$ *are shown with dashed and dotted lines.*

present numerical simulations for the corresponding dynamical behavior in the nonlinear system (1.1), obtained on the basis of the foregoing analysis.

5. Applications. We now examine the previous results numerically for a particular model. To this end, we set the differential operator to

$$L\left(\frac{\partial}{\partial t}\right) = \frac{\partial^2}{\partial t^2} + \gamma\frac{\partial}{\partial t} + 1 \tag{5.1}$$

and further specify the connectivity kernel. Since a neuronal field might exhibit excitatory and inhibitory connections, the kernel K contains both excitatory and inhibitory distributions over space. In case of a homogeneous and isotropic neuronal field, a choice of K is

$$K(z) = \frac{1}{\sqrt{\pi}}(a_{\mathrm{e}} e^{-z^2} - a_{\mathrm{i}} r e^{-r^2 z^2}), \tag{5.2}$$

where $a_{\mathrm{e}}, a_{\mathrm{i}}$ denote excitatory and inhibitory synaptic weights and $r = \sigma_{\mathrm{e}}/\sigma_{\mathrm{i}}$ gives the relation of excitatory and inhibitory spatial connectivity ranges σ_{e} and σ_{i}. Since the present work treats dynamics on a mesoscopic spatial scale, it does not resolve single synapses and synaptic interaction is considered in terms of normalized distributions of excitatory and inhibitory connections as in (5.2). Thus a purely excitatory connection (Figure 1(a)) is obtained when $a_{\mathrm{i}} = 0$ and $a_{\mathrm{e}} > 0$, whereas the choice $a_{\mathrm{e}} = 0$ and $a_{\mathrm{i}} > 0$ gives an inhibitory connection (Figure 1(b)). Similarly, for $a_{\mathrm{e}} > a_{\mathrm{i}} > 0$, local inhibition and lateral excitation (Figure 1(c)) or local excitation and lateral inhibition (Figure 1(d)) can be obtained by choosing $r > a_{\mathrm{e}}/a_{\mathrm{i}}$ or $0 < r < 1$, respectively. We shall mostly focus on these last two cases. Finally, we take $\beta = 1$ and choose the transfer function in (1.1) as the sigmoid $S(y) = 1/(1 + \exp(-1.8(y - 3)))$ according to previous works [46, 36].

The subsequent temporal integration procedure applies a fourth-order Runge–Kutta algorithm, while the spatial integration algorithm discretizes the field into N intervals and applies

$$\int_0^l f(z)dz \approx \sum_{i=1}^{N} \frac{1}{2}(f(z_i) + f(z_{i+1}))\Delta x \tag{5.3}$$

for any function f, with l the field length and $\Delta x = l/N$. Further, for periodic boundary conditions, the integration obeys the circular rule

$$\int_{-\infty}^{\infty} K(|x - y|)f(y)dy \approx \int_0^l K(l/2 - |l/2 - |x - y||)f(y)dy. \tag{5.4}$$

5.1. Stability of V^*. The equilibria V^* are found from (2.4). Figure 2 shows solutions V^* of (2.4) with respect to the external input for various values of κ. In the case where $\kappa > 2.2$, there exist up to three solutions A, B, and C subject to the external input, whereas there is only a single solution for $\kappa \leq 2.2$. Theorem 2.1 gives a sufficient condition for the stability of these equilibria. Note that for $\gamma > \sqrt{2}$ the inequality (2.10) is automatically satisfied, so $c < 1$ is a sufficient condition for asymptotic stability by the theorem. From (5.2) we have

$$c = \alpha|2a_{\mathrm{e}}\Phi(x_0) - 2a_{\mathrm{i}}\Phi(x_0 r) - (a_{\mathrm{e}} - a_{\mathrm{i}})|, \quad x_0 = \sqrt{\frac{1}{1 - r^2}\ln\left(\frac{a_{\mathrm{e}}}{a_{\mathrm{i}} r}\right)}, \tag{5.5}$$

where Φ is the Gaussian error function and $0 < r < 1$ or $r > a_{\mathrm{e}}/a_{\mathrm{i}}$. The spatial distance x_0 marks the change of sign of the kernel function and thus separates inhibitory from excitatory connections. The external input E^* affects c through $\alpha = S'(V^*)$. In Figure 3, c is plotted with respect to the input E^* for $r = 0.5$ and various parameter values of $a_{\mathrm{e}}, a_{\mathrm{i}}$ at the different equilibria V^*. Stability is guaranteed by Theorem 2.1 at least in the region $c < 1$. In this line, Figure 4 shows a space-time plot of field

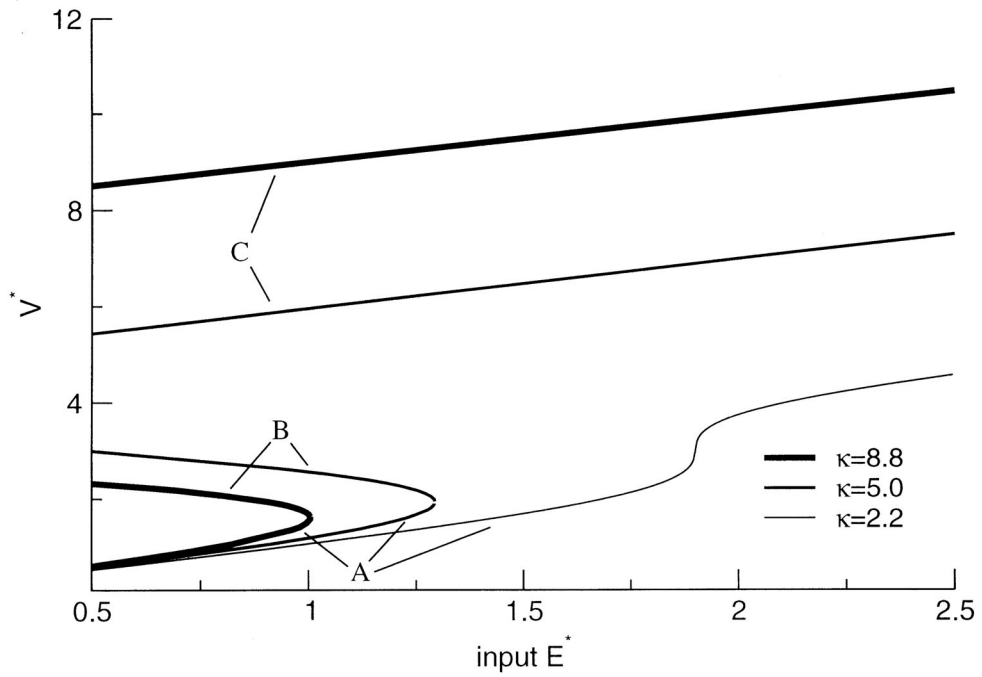


FIG. 2. *Stationary constant fields V^* plotted with respect to the external input E^* for various parameters κ. Up to three solutions A, B, and C may exist for a given input level.*

activity that relaxes to a lower solution A for $c = 0.85$ (cf. Figure 2). On the other hand, at sufficiently high values of α (and thus of c), stability is lost since when

$$\alpha = \frac{1}{\kappa}, \tag{5.6}$$

the condition (3.2) for a type Ia bifurcation, is satisfied. This bifurcation point is also indicated in Figure 3. Therefore, the constant solutions denoted B in Figures 2–3 are unstable. Interestingly, this general result shows accordance to findings in previous works for special connectivity kernels [36, 23]. Finally, in the region $c > 1$ and $\alpha < 1/\kappa$, additional bifurcations might occur, yielding loss of stability. These are discussed in the following section.

5.2. Bifurcations. Recall that the external input defines the set of constant fields V^*, which subsequently determine the value of α; hence α is an appropriate bifurcation parameter. For bifurcations to periodic patterns (Turing case Ib), the threshold condition from (3.3) reads

$$\alpha_{\text{thr}} = \frac{1}{a_{\text{e}} e^{-k_0^2/4} - a_{\text{i}} e^{-k_0^2/4r^2}}, \qquad k_0^2 = \frac{4r^2}{r^2 - 1} \ln \frac{a_{\text{e}} r^2}{a_{\text{i}}}, \tag{5.7}$$

where $k_0 = \arg\max_k \hat{K}(k)$. As $\alpha > 0$, we obtain directly from (5.7) that $r < 1$, i.e., there is no Turing instability for $r > 1$. Figure 5 displays thresholds α_{thr} with respect to parameters r, confirming this finding. Figure 6 displays a space-time plot of the corresponding Turing instability with $r = 0.5$.

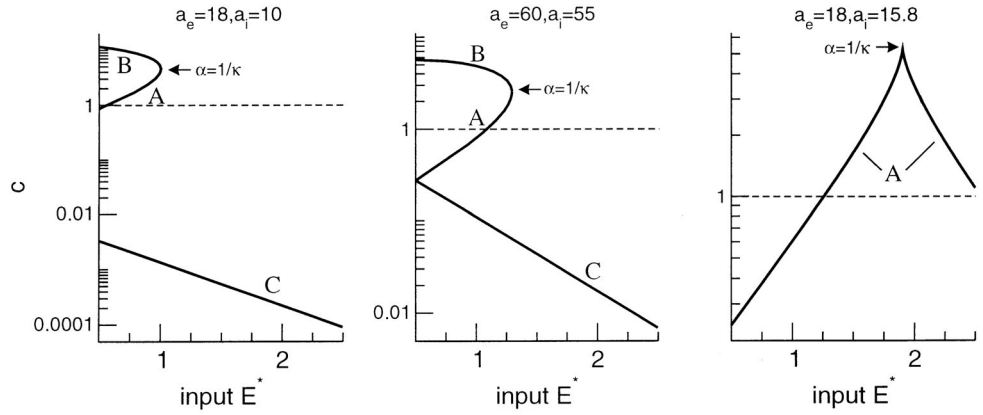


FIG. 3. *Parameter c from Theorem* 2.1 *plotted with respect to the external input* E^* *for various parameters* $a_\mathrm{e}, a_\mathrm{i}$ *and* $r = 0.5,\ \gamma > \sqrt{2}$. *The characters* $A, B,$ *and* C *denote stationary solutions (see Figure* 2*) and solutions in the region below the dashed line fulfill the sufficient condition of asymptotic stability. It turns out that stationary solutions* C *are asymptotically stable for all external inputs in case of* $a_\mathrm{e} = 18,\ a_\mathrm{i} = 10$ *and* $a_\mathrm{e} = 60,\ a_\mathrm{i} = 55$.

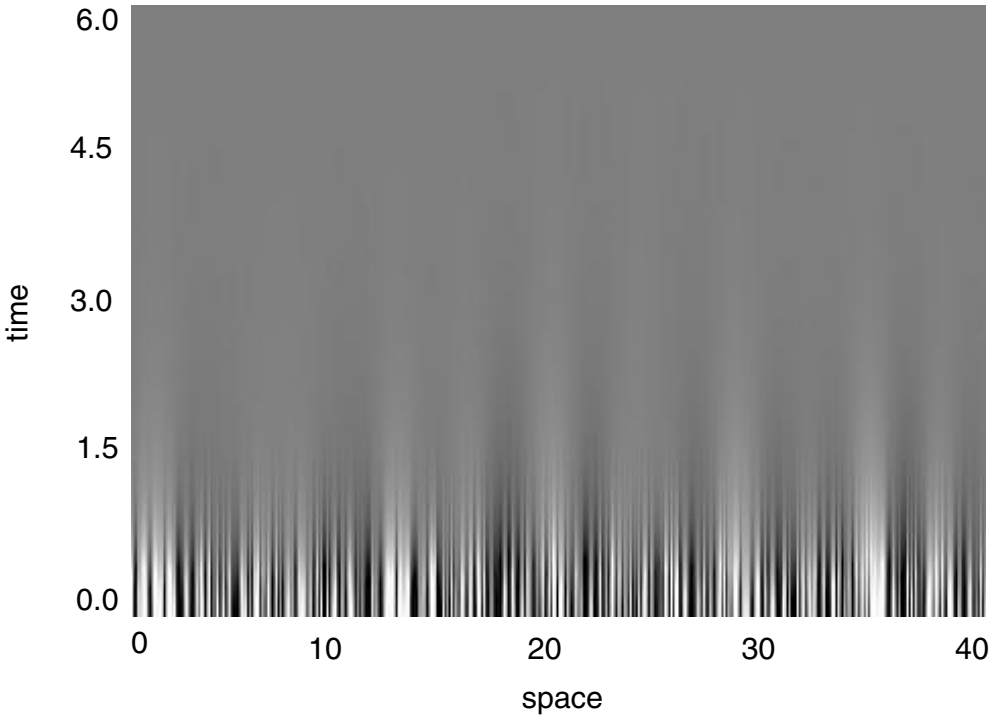


FIG. 4. *Space-time plot of an asymptotically stable field for the Gaussian connectivity kernel and parameters* $E^* = 0.5$, $r = 0.5$, $\gamma = 2$, $a_\mathrm{e} = 60$, $a_\mathrm{i} = 55$, $v = 100$, $\beta = 1$, $N = 400$. *Initial values* $V^0(x,t)$ *are chosen randomly from a uniform distribution on* $[V^* - 0.1, V^* + 0.1]$ *for* $t \in [-l/v, 0]$, *where* $l = 40$. *The gray scale encodes the deviations from the stationary solution.*

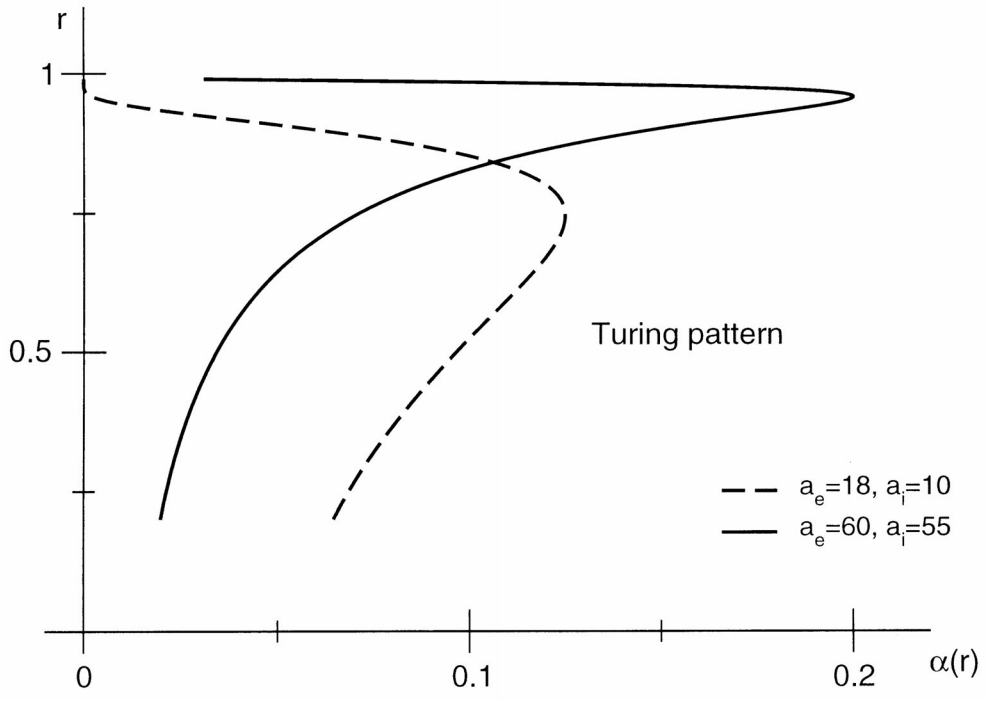


FIG. 5. *Thresholds of stationary Turing bifurcations* α_{thr} *plotted for two parameter sets* $a_{\mathrm{e}}, a_{\mathrm{i}}$. *The regime of the Turing instability obeys* $\alpha > \alpha_{\mathrm{thr}}$, *i.e., the right-hand side of each curve. The thresholds* α_{thr} *are independent from the synaptic parameter* γ *and the propagation speed* v, *while the external input* E^* *determines* α *implicitly.*

Next, we consider oscillatory phenomena. From Theorem 3.1, a necessary condition for oscillatory behavior is

$$v < v_{\mathrm{thr}} = \frac{\alpha}{|\gamma|\sqrt{\pi}} \left[\frac{a_{\mathrm{i}}}{r} - a_{\mathrm{e}} + 2 \left(a_{\mathrm{e}} e^{-x_0^2} - \frac{a_{\mathrm{i}}}{r} e^{-r^2 x_0^2} \right) \right], \tag{5.8}$$

with x_0 taken from (5.5). Figure 7 shows plots of thresholds v_{thr} with respect to the parameter r for two parameter couples of $a_{\mathrm{e}}, a_{\mathrm{i}}$. It turns out that condition (5.8) is fulfilled and oscillations are expected for a wide range of $r > 1$, whereas $r < 1$ (lateral inhibition) allows only for a small parameter regime. For appropriate parameters, the properties of the kernel and the temporal delays introduced by finite propagation speed interact in a way that destabilizes the stationary state and produces oscillations. Similar effects have also been found in previous works [6, 22].

Section 4 gives conditions for an oscillatory bifurcation in case of large propagation velocity. To obtain oscillating activity constant in space (case IIa), Figure 1(d) illustrates the conditions $r < 1$ and $-\hat{K}_1(0)/(\gamma v) > \max_k \hat{K}(k)$, implying

$$\frac{a_{\mathrm{e}} - a_{\mathrm{i}}/r}{\gamma v \sqrt{\pi}} > a_{\mathrm{e}} e^{-k_0^2/4} - a_{\mathrm{i}} e^{-k_0^2/4r^2}, \tag{5.9}$$

where k_0 is taken from (5.7). Figure 8 displays the corresponding spatiotemporal activity for appropriate parameters. From the figure, an oscillation frequency of about $\omega = 0.28$ can be observed. This agrees well with the theory, as from (3.10)

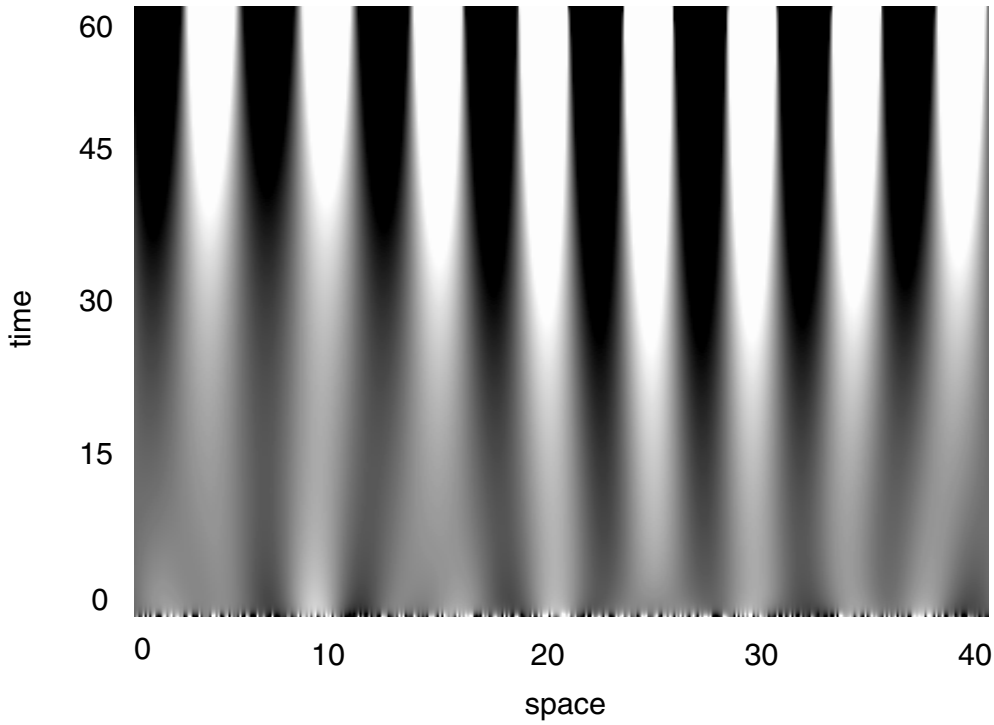


FIG. 6. *Space-time plot of the Turing instability, obtained for the Gaussian connectivity kernel and parameters* $E^* = 0.74$, $r = 0.5$, $\gamma = 2$, $a_e = 60$, $a_i = 55$, $v = 100$, $\beta = 1$, $N = 400$, *and* $1/\alpha = 0.727$. *Initial conditions* $V^0(x,t)$ *are chosen randomly from a uniform distribution on* $[V^* - 0.1, V^* + 0.1]$ *for* $t \in [-l/v, 0]$, *where* $l = 40$.

with $c \geq 1$ a frequency in the interval $[0, 0.26]$ is predicted at the bifurcation. The small discrepancy arises from choosing the simulation parameters somewhat beyond the bifurcation values in order to obtain reasonably high amplitude solutions for visualization.

On the other hand, for traveling waves (case IIb) $k \neq 0$ and Figure 1(c) gives the conditions $r > a_e/a_i$ and $\max_k -\hat{K}_1(k)/(\gamma v) > \hat{K}(0)$. A series expansion for $\hat{K}_1$ [20] yields a single implicit condition for parameters a_e, a_i, r and k. Figure 9 shows the corresponding space-time plot of the wave instability for appropriate parameters. Here $r > 1$, and the field activity is shifted from local to lateral spatial locations, facilitating traveling waves.

Finally we examine how the phase velocity of traveling waves depends on the propagation velocity v in the system. From (4.10) the phase velocity reads

$$v_{\mathrm{ph}} = \frac{\omega}{k^*} = \frac{v}{k^*}\sqrt{\frac{\alpha\hat{K}(k^*) - 1}{\frac{1}{2}\alpha\hat{K}_2(k^*) - v^2}}, \tag{5.10}$$

where k^* solves (4.8). In Figure 10, v_{ph} is plotted and exhibits a slightly nonlinear dependence on the propagation velocity for the applied parameters. We point to the small ratio of phase velocity to propagation velocity in accordance with previous findings [32, 23].

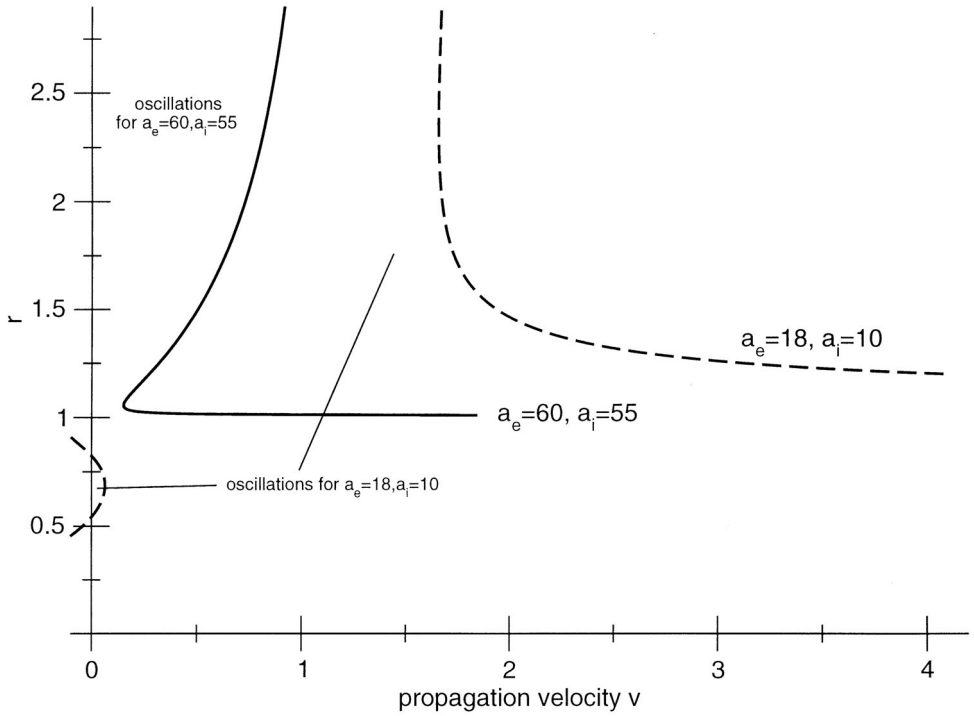


FIG. 7. *Thresholds of oscillatory phenomena v_{thr} for two parameter sets $a_{\mathrm{e}}, a_{\mathrm{i}}$. The sufficient condition for oscillations is fulfilled for $v < v_{\mathrm{thr}}$, i.e., the left-hand side of each curve.*

6. Conclusion. We have presented an analysis of the stability of equilibrium solutions for a general class of neural field equations on the real line. The details of bifurcations arising from loss of stability provide important information concerning a variety of dynamical behavior that is of neuroscientific interest, including spatial patterns and traveling waves. The stationary bifurcations and the resulting spatial patterns depend only on the connectivity kernel, and are completely determined by its Fourier transform $\hat{K}(k)$. On the other hand, the axonal delays due to finite propagation speed are shown to have significant effects on the nonstationary bifurcations. In fact, we have proved that in first- and second-order systems, nonstationary bifurcations of equilibria can occur only if the delays are sufficiently large, that is, when the transmission speed is sufficiently small. This behavior is different from that of higher order systems, where nonstationary bifurcations can occur even in the absence of delays. By a perturbation approach we have expressed the conditions for bifurcation in terms of the Fourier transforms of the moments of the kernel function. For high signal transmission speeds, only the first kernel moment needs to be considered to draw qualitative conclusions. For first- and second-order systems this leads to a simple method for determining the possible bifurcation types by comparing the Fourier transforms $\hat{K}(k)$ and $-\hat{K}_1(k)/v\gamma$. Furthermore, the bifurcations depend only on the extremal values of the transforms, rather than the precise shapes of the kernels.

The analysis presented here, being applicable to a broad range of connectivity and synaptic properties and transfer functions, suggests some general conclusions on the types of nonlinear dynamics that can be observed in a fairly wide class of systems.

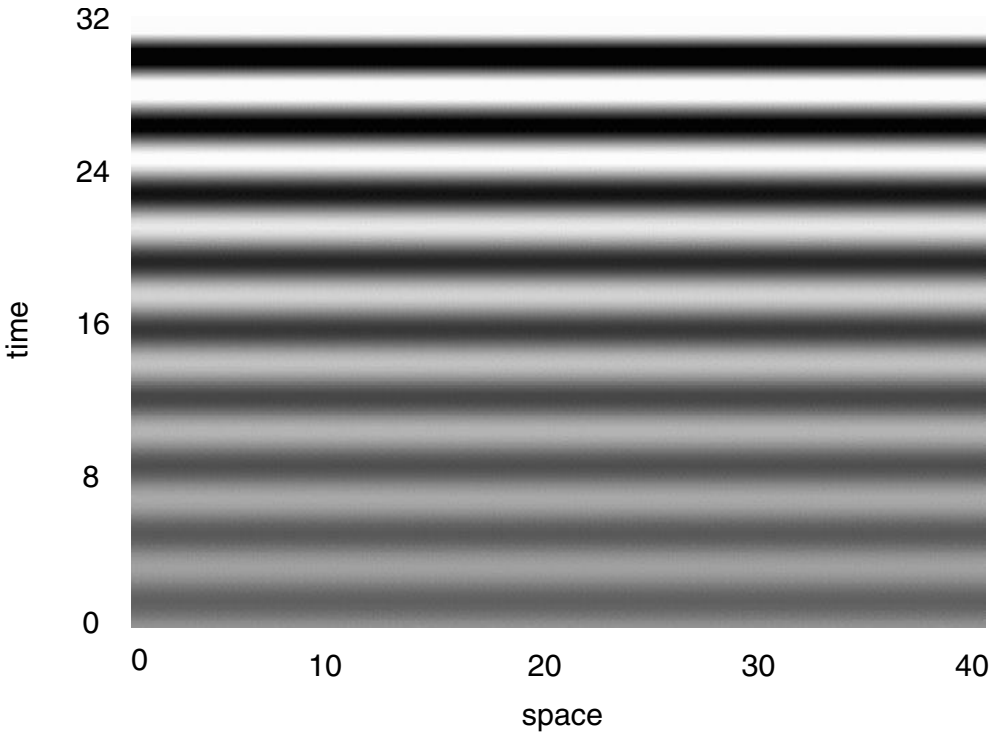


FIG. 8. *Space-time plot of the Hopf instability leading to periodic oscillations of a spatially constant solution (type* IIa *bifurcation), obtained for the Gaussian connectivity kernel and parameters* $E^* = 0.91$, $r = 0.2$, $\gamma = 0.5$, $a_\mathrm{e} = 18$, $a_\mathrm{i} = 10$, $v = 6$, $\beta = 1$, $N = 400$. *Initial conditions are* $V^0(x,t) = V^* + 0.02$ *for* $t \in [-l/v, 0]$, *with* $V^* = 0.11$ *and* $l = 40$.

For instance, one generally expects to see oscillatory behavior whenever the signal transmission speed is sufficiently small. In fact, for completely general kernels, the peaks of $\hat{K}(k)$ and $-\hat{K}_1(k)$ are more likely to occur at some nonzero k rather than at the precise value $k = 0$. This suggests that in first- and second-order systems the prevalent dynamics arising from bifurcations of equilibria will be either spatial patterns or traveling waves, depending on whether the transmission speed is large or small, respectively. Nevertheless, more specific kernel types may dictate different dynamical behavior depending on the application.

There are many studies of discrete networks which exhibit in-phase periodic behavior by increased constant delay (e.g., [49, 6]) and propagation delay (e.g., [17, 26, 22]). These studies consider specific network connectivities and obtain similar results with respect to the role of delays in oscillatory behavior. In this context, we would like to mention the recent work of Earl and Strogatz [9], who studied the stability of discrete, homogeneous oscillator networks with constant connection delays. They obtain a rather strong stability condition for global in-phase oscillations that is independent of the connectivity topology, provided each node has the same number of connections. The in-phase oscillations correspond to type IIa bifurcations in our study; however, we also have the possibility of other bifurcation types, indicating that the neural fields considered here can exhibit a richer range of dynamics. An important difference arises from the nature of the delays in the two models. In the discrete network with a constant connection delay, each unit knows the state of its neighbors

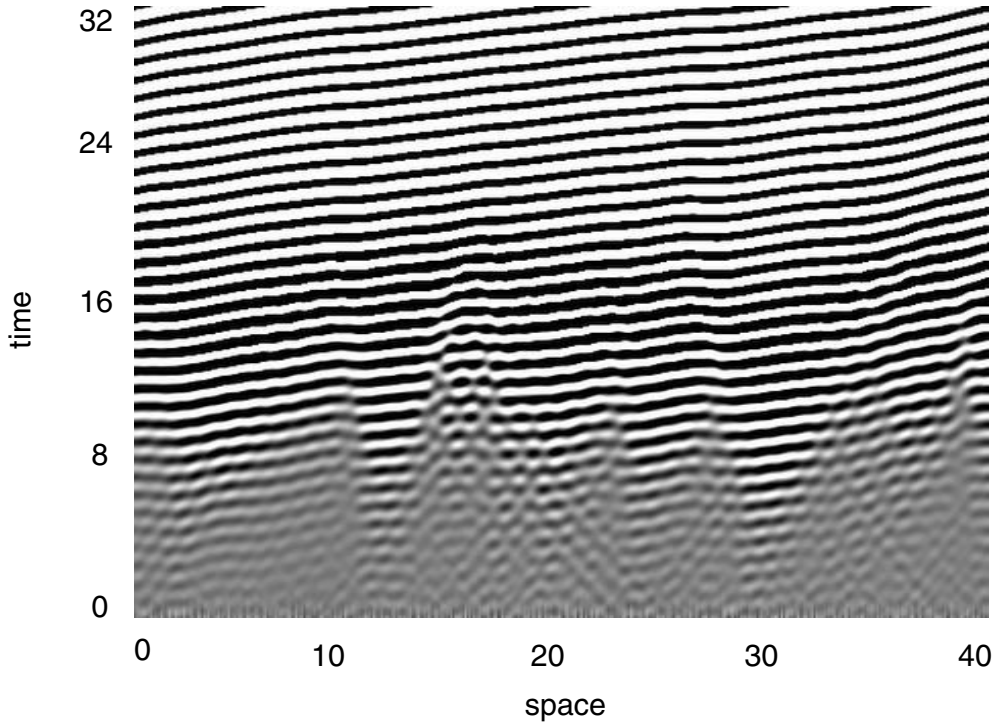


FIG. 9. *Space-time plot of the wave instability (type* IIb *bifurcation), obtained for the Gaussian connectivity kernel and parameters* $E^* = 1.29$, $r = 3$, $\gamma = 2$, $a_e = 60$, $a_i = 55$, $v = 1$, $\beta = 1$, $N = 400$. *Initial conditions* $V^0(x,t)$ *are chosen randomly from a uniform distribution on* $[V^* - 0.1, V^* + 0.1]$ *for* $t \in [-l/v, 0]$, *where* $l = 40$.

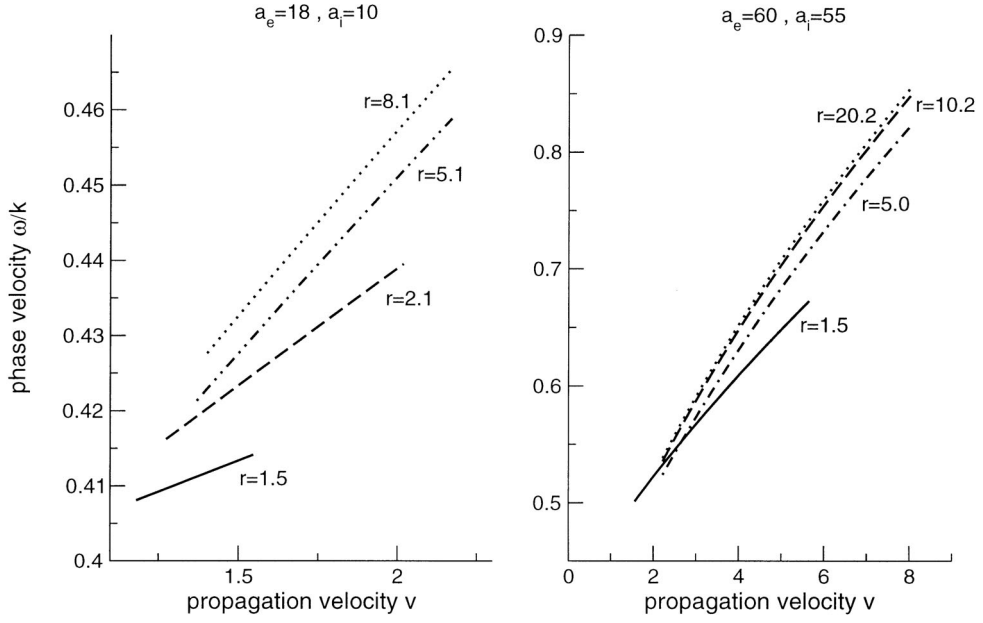


FIG. 10. *The wave velocity of traveling waves with respect to axonal propagation velocity. Parameters are* $\gamma = 0.2$, $0.03 \le \alpha \le 0.11$ *and* $\omega/v \approx 0.7$ *(left, for* $v \approx 2$*) and* $\omega/v \approx 0.3$ *(right, for* $v \approx 7$*), respectively.*

at the same time instant (although not at the present time), so it is plausible that this type of arrangement favors in-phase oscillations. On the other hand, our model involves distance-dependent delays, for which travelling waves may be the more natural type of oscillatory behavior. The details of this interesting connection will be given in a future paper.

Our work is mostly motivated by experimental findings (e.g., [29, 44, 16]). In this line, the presented study aims to generalize the analysis of synaptically coupled neuronal fields in order to gain a classification scheme for observed spatiotemporal patterns. Here, we would like to mention the important generalization of Amari [1] in lateral-inhibition-type fields without axonal delay. Since neurophysiological properties of observed neural tissue are not accessible precisely, a classification scheme might link model functionals with observed phenomena. For example, observed traveling waves necessitate an axonal propagation velocity below a certain threshold defined by connectivity kernel properties and synaptic response properties (Theorem 3.1), and furthermore, their frequencies are confined to a bounded band (Theorem 3.2). In addition, this classification might be important for estimating interaction parameters from multisite neuronal data (e.g., [14]). Due to the large number of different activity phenomena, further studies in this area could incorporate additional mechanisms like standing and traveling pulse fronts as in [32, 33], boundary effects in local neuronal areas (e.g., [7]), the influence of external inputs [45, 10] local in space and time, or the constant delayed feedback found experimentally in thalamocortical connections [41] and visual areas [27, 11] and which has been addressed in several theoretical studies (e.g., [13, 5, 3]). In particular, the mutual treatment of both constant delayed feedback and propagation delay proposes new insights into information processing between distant brain areas. Moreover, consideration of neural fields in higher space dimensions might yield further interesting results.

Acknowledgment. We thank Matthew P. James for a critical reading of the manuscript.

REFERENCES

[1] S. Amari, *Dynamics of pattern formation in lateral-inhibition type neural fields*, Biol. Cybernetics, 27 (1977), pp. 77–87.

[2] G. G. Blasdel and G. Salama, *Voltage-sensitive dyes reveal a modular organization in monkey striate cortex*, Nature, 321 (1986), pp. 579–585.

[3] P. C. Bressloff, *Synaptically generated wave propagation in excitable neural media*, Phys. Rev. Lett., 82 (1999), pp. 2979–2982.

[4] P. C. Bressloff, *Traveling waves and pulses in a one-dimensional network of excitable integrate-and-fire neurons*, J. Math. Biol., 40 (2000), pp. 169–198.

[5] P. C. Bressloff and S. Coombes, *Physics of the extended neuron*, Internat. J. Modern Phys. B, 11 (1997), pp. 2343–2392.

[6] N. Brunel and V. Hakim, *Fast global oscillations in networks of integrate-and-fire neurons with low firing rates*, Neural Comput., 11 (1999), pp. 1621–1671.

[7] C. Cavada and P. S. Goldman-Rakic, *Multiple visual areas in the posterior parietal cortex of primates*, Prog. Brain Res., 95 (1993), pp. 123–137.

[8] S. Coombes, G. J. Lord, and M. R. Owen, *Waves and bumps in neuronal networks with axo-dendritic synaptic interactions*, Phys. D, 178 (2003), pp. 219–241.

[9] M. G. Earl and S. H. Strogatz, *Synchronization in oscillator networks with delayed coupling: A stability criterion*, Phys. Rev. E, 67 (2003), p. 036204.

[10] M. Enculescu and M. Bestehorn, *Activity dynamics in nonlocal interacting neural fields*, Phys. Rev. E, 67 (2003), p. 041904.

[11] A. K. Engel, P. Koenig, A. K. Kreiter, and W. Singer, *Interhemispheric synchronization of oscillatory neuronal responses in cat visual cortex*, Science, 252 (1991), pp. 1177–1179.

[12] B. Ermentrout, *Neural networks as spatio-temporal pattern-forming systems*, Rep. Progr. Phys., 61 (1998), pp. 353–430.
[13] U. Ernst, K. Pawelzik, and T. Geisel, *Delay-induced multi-stable synchronization of biological oscillators*, Phys. Rev. E, 57 (1998), pp. 2150–2162.
[14] O. François, C. Larota, J. Horikawa, and T. Hervé, *Diffusion and innovation rates for multidimensional neuronal data with large spatial covariances*, Network: Comput. Neural Syst., 11 (2000), pp. 211–220.
[15] W. J. Freeman, *Characteristics of the synchronization of brain activity imposed by finite conduction velocities of axons*, Int. J. Bif. Chaos, 10 (2000), pp. 2307–2322.
[16] W. J. Freeman, *Neurodynamics: An Exploration in Mesoscopic Brain Dynamics*, Perspectives in Neural Computing, Springer-Verlag, Berlin, 2000.
[17] W. Gerstner, *Rapid phase locking in systems of pulse-coupled oscillators with delays*, Phys. Rev. Lett., 76 (1996), pp. 1755–1758.
[18] P. S. Goldman-Rakic, *Cellular basis of working memory*, Neuron, 14 (1995), pp. 477–485.
[19] D. Golomb and G. B. Ermentrout, *Effects of delay on the type and velocity of travelling pulses in neuronal networks with spatially decaying connectivity*, Network: Comput. Neural Syst., 11 (2000), pp. 221–246.
[20] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, San Diego, 2000.
[21] A. Grindvald, L. Anglister, J. A. Freeman, R. Hildesheim, and A. Manker, *Real-time optical imaging of naturally evoked electrical activity in intact frog brain*, Nature, 308 (1984), pp. 848–850.
[22] H. Haken, *Effect of delay on phase locking in a pulse coupled neural network*, Eur. Phys. J. B, 18 (2000), pp. 545–550.
[23] A. Hutt, M. Bestehorn, and T. Wennekers, *Pattern formation in intracortical neuronal fields*, Network: Comput. Neural Syst., 14 (2003), pp. 351–368.
[24] V. K. Jirsa and H. Haken, *A derivation of a macroscopic field theory of the brain from the quasi-microscopic neural dynamics*, Phys. D, 99 (1997), pp. 503–526.
[25] V. K. Jirsa, K. J. Jantzen, A. Fuchs, and J. A. S. Kelso, *Spatiotemporal forward solution of the EEG and MEG using network modelling*, IEEE Trans. Medical Imaging, 21 (2002), pp. 493–504.
[26] W. M. Kistler, R. Seitz, and J. L. van Hemmen, *Modeling collective excitations in cortical tissue*, Phys. D, 114 (1998), pp. 273–295.
[27] L. A. Krubitzer and J. H. Kaas, *Cortical integration of parallel pathways in the visual system of primates*, Brain Res., 478 (1987), pp. 161–165.
[28] J. W. Lance, *Current concepts of migraine pathogenesis*, Neurology, 43 (1993), pp. S11–S15.
[29] P. L. Nunez, *Neocortical dynamics and human EEG rhythms*, Oxford University Press, New York, Oxford, 1995.
[30] P. L. Nunez, *Toward a quantitative description of large-scale neocortical dynamic function and EEG*, Behav. Brain Sci., 23 (2000), pp. 371–437.
[31] R. Osan and G. B. Ermentrout, *The evolution of synaptically generated waves in one- and two-dimensional domains*, Phys. D, 163 (2002), pp. 217–235.
[32] D. J. Pinto and G. B. Ermentrout, *Spatially structured activity in synaptically coupled neuronal networks:* I. *Travelling fronts and pulses*, SIAM J. Appl. Math., 62 (2001), pp. 206–225.
[33] D. J. Pinto and G. B. Ermentrout, *Spatially structured activity in synaptically coupled neuronal networks:* II. *Lateral inhibition and standing pulses*, SIAM J. Appl. Math., 62 (2001), pp. 226–243.
[34] C. J. Rennie, P. A. Robinson, and J. J. Wright, *Unified neurophysical model of EEG spectra and evoked potentials*, Biol. Cybernetics, 86 (2002), pp. 457–471.
[35] P. A. Robinson, P. N. Loxley, S. C. O'Connor, and C. J. Rennie, *Modal analysis of corticothalamic dynamics, electroencephalographic spectra and evoked potentials*, Phys. Rev. E, 63 (2001), p. 041909.
[36] P. A. Robinson, C. J. Rennie, and J. J. Wright, *Propagation and stability of waves of electrical activity in the cerebral cortex*, Phys. Rev. E, 56 (1997), pp. 826–840.
[37] B. M. Salzberg, H. V. Davila, and L. B. Cohen, *Optical recording of impulses in individual neurons of an invertebrate central nervous system*, Nature, 246 (1973), pp. 508–509.
[38] H. G. Schuster and P. Wagner, *A model for neuronal oscillations in the visual cortex*: 1. *Mean-field theory and derivation of the phase equations*, Biol. Cybern., 64 (1990), pp. 77–82.
[39] W. Singer and C. M. Gray, *Visual feature integration and the temporal correlation hypothesis*, Ann. Rev. Neurosci., 18 (1995), pp. 555–586.

[40] H. Spors and A. Grindvald, *Spatio-temporal dynamics of odor representations in the mammalian olfactory bulb*, Neuron, 34 (2002), pp. 301–315.
[41] M. Steriade, E. G. Jones, and R. R. Llinas, *Thalamic Oscillations and Signalling*, Wiley, New York, 1990.
[42] A. K. Sturm and P. König, *Mechanisms to synchronize neuronal activity*, Biol. Cybernet., 84 (2001), pp. 153–172.
[43] P. Tass, *Oscillatory cortical activity during visual hallucinations*, J. Biol. Phys., 23 (1997), pp. 21–66.
[44] C. Uhl, ed., *Analysis of Neurophysiological Brain Functioning*, Springer-Verlag, Berlin, 1999.
[45] T. Wennekers, *Dynamic approximation of spatio-temporal receptive fields in nonlinear neural field models*, Neural Comput., 14 (2002), pp. 1801–1825.
[46] H. R. Wilson and J. D. Cowan, *A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue*, Kybernetik, 13 (1973), pp. 55–80.
[47] J. J. Wright and R. R. Kydd, *The electroencephalogram and cortical neural networks*, Network, 3 (1992), pp. 341–362.
[48] J. Y. Wu, L. Guan, and Y. Tsau, *Propagating activation during oscillations and evoked responses in neocortical slices*, J. Neurosci., 19 (1999), pp. 5005–5015.
[49] M. K. S. Yeung and S. H. Strogatz, *Time delay in the Kuramoto model of coupled oscillators*, Phys. Rev. Lett., 82 (1999), pp. 648–651.

SIAM J. APPL. MATH.
Vol. 65, No. 2, pp. 667–683

RESOLVENT ESTIMATES FOR PLANE COUETTE FLOW*

PABLO BRAZ E SILVA†

Abstract. We discuss the problem of deriving estimates for the resolvent of the linear operator associated with three-dimensional perturbations of plane Couette flow and determining its dependence on the Reynolds number R. Depending on the values of the parameters involved, we derive estimates analytically. For the remaining values of the parameters, we prove that deriving estimates for the resolvent can be reduced to estimating the solutions of a 4th-order linear homogeneous ordinary differential equation with nonhomogeneous boundary conditions. We study these boundary value problems numerically. Our results indicate the L_2 norm of the resolvent to be proportional to R^2.

Key words. Couette flow, resolvent estimates

AMS subject classifications. 76E05, 47A10, 35Q30, 76D05

DOI. 10.1137/S0036139903426940

1. Introduction.

It is well known that plane Couette flow is stable for infinitesimal perturbations for all values of the Reynolds number R [12]. In laboratory experiments, though, transition to turbulence is observed for Reynolds numbers as low as approximately 350 [4, 16]. This discrepancy may be caused by a small domain of attraction of the Couette flow. Therefore, it is of great interest to understand how this domain of attraction scales with the Reynolds number R.

The so-called resolvent technique for nonlinear differential equations allows one to derive nonlinear stability results from linear stability. To this end, one uses estimates for the resolvent of a linear operator. One of its advantages is the quantification of stability; that is, when successfully applied, the method gives information about the domain of attraction of a stable solution [7, 8].

For plane Couette flow, recent works use the resolvent technique to derive a threshold amplitude for perturbations of the base flow, that is, to give a lower bound on the size of perturbations that can lead to turbulence [9, 6, 2]. In this case, successful application of the method requires estimates for the resolvent $(s\mathcal{I} - \mathcal{L}_R)^{-1}$ of the linear operator $\mathcal{L}_R$ associated with perturbations of the base flow, for the parameter s belonging to the unstable half-plane $\mathrm{Re}(s) \geq 0$. These estimates should show exactly how the norm of the resolvent depends on R. Our aim is to study this dependence.

For large enough values of $|s|$, depending on the Reynolds number, analytical estimates for the L_2 norm of the resolvent have already been proved [1, 10]. To derive an estimate valid for the whole unstable half-plane, direct numerical computations have been used indicating the L_2 norm of the resolvent to be proportional to R^2 [6, 17]. In [10], R-dependent weighted norms are used. Direct numerical computations indicate that in one of the norms considered, the resolvent is proportional to R.

We study the three-dimensional case, with periodic boundary conditions in two of the directions. Our results indicate the L_2 norm of the resolvent to be proportional to R^2, agreeing with the computations in [6, 17]. Our main result is a theorem showing

*Received by the editors May 1, 2003; accepted for publication (in revised form) June 7, 2004; published electronically January 27, 2005. This work was supported by a postdoctoral fellowship from FAPESP/Brazil: 02/13270-1.

http://www.siam.org/journals/siap/65-2/42694.html

†Departamento de Matemática, Universidade Federal de Pernambuco, CEP 50740-540, Recife, PE, Brazil (pablo@dmat.ufpe.br).

that the problem of proving the resolvent estimates can be reduced to estimating the solutions of a 4th-order homogeneous linear ordinary differential equation with nonhomogenous boundary conditions. Numerical computations, which are simple and reliable in this case, are used only to study the norms of the solutions of those boundary value problems. The analysis carried out here has other advantages. First of all, it clarifies the reasons for the R^2 growth of the L_2 norm of the resolvent, since it shows exactly where the extra factor of R comes into the game. It also gives some physical insight about the problem, showing that different components of perturbations of the base flow have different scales with respect to R. We also discuss the reasons for the better dependence of the resolvent on R when the weighted norm from [10] is used.

2. The problem. We first give some notations that will be used throughout this work.

In general, elements of $\mathbb{R}^3$ will be represented by boldface letters. The same letter may be used for one of the coordinates of the vector. For example, when convenient, we write $\mathbf{x} = (x, y, z) \in \mathbb{R}^3$. We denote by Ω the set

$$\Omega := [0, 2\pi] \times [0, 2\pi] \times [0, 1].$$

The Euclidean inner product in $\mathbb{R}^3$ is denoted by $\cdot$; that is, for $\mathbf{x} = (x_1, x_2, x_3)$, $\mathbf{y} = (y_1, y_2, y_3)$, we have

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{3} x_i y_i.$$

The L_2 inner product and norm over Ω are denoted, respectively, by

$$\langle \mathbf{u}_1, \mathbf{u}_2 \rangle = \int_\Omega \overline{\mathbf{u}}_1 \cdot \mathbf{u}_2 \, d\mathbf{x}, \quad \|\mathbf{u}\|^2 = \langle \mathbf{u}, \mathbf{u} \rangle.$$

In our choice of coordinates, the Couette flow is the vector field $\mathbf{U} = (0, z, 0)$, which is a steady solution of

$$\begin{aligned} &\mathbf{U}_t + (\mathbf{U} \cdot \nabla)\mathbf{U} + \nabla P = \frac{1}{R}\Delta \mathbf{U}, \\ &\nabla \cdot \mathbf{U} = 0, \\ &\mathbf{U}(x, y, 0, t) = (0, 0, 0), \\ &\mathbf{U}(x, y, 1, t) = (0, 1, 0), \\ &\mathbf{U}(x, y, z, t) = \mathbf{U}(x + 2\pi, y, z, t), \\ &\mathbf{U}(x, y, z, t) = \mathbf{U}(x, y + 2\pi, z, t) \end{aligned} \tag{2.1}$$

for P a constant. The positive parameter R is the Reynolds number. We consider $R \geq 1$, since this is the physically interesting case. We also note that there are no technical reasons for this assumption, only a slight simplification of the presentation. Problem (2.1) describes the flow of an incompressible fluid between the two parallel planes $z = 0$ and $z = 1$, the plane $z = 0$ at rest and the plane $z = 1$ moving in the y direction with constant velocity 1.

We want to analyze the resolvent of the linear operator associated with perturbations of the Couette flow $\mathbf{U}$. Therefore, we consider the initial boundary value

problem

$$
\begin{aligned}
&\mathbf{u}_t + (\mathbf{u}\cdot\nabla)\mathbf{U} + (\mathbf{U}\cdot\nabla)\mathbf{u} + \nabla p = \frac{1}{R}\Delta\mathbf{u} + \mathbf{F},\\
&\nabla\cdot\mathbf{u} = 0,\\
&\mathbf{u}(x,y,0,t) = \mathbf{u}(x,y,1,t) = (0,0,0),\\
&\mathbf{u}(x,y,z,t) = \mathbf{u}(x+2\pi,y,z,t),\\
&\mathbf{u}(x,y,z,t) = \mathbf{u}(x,y+2\pi,z,t),\\
&\mathbf{u}(x,y,z,0) = (0,0,0),
\end{aligned}
\tag{2.2}
$$

which is the linearization of the equations governing three-dimensional perturbations $\mathbf{u}(\mathbf{x},t) = (u(\mathbf{x},t), v(\mathbf{x},t), w(\mathbf{x},t))$ of $\mathbf{U}$. The forcing $\mathbf{F}(\mathbf{x},t) = (F(\mathbf{x},t), G(\mathbf{x},t), H(\mathbf{x},t))$ is a given C^∞ function, satisfying

$$
\int_0^\infty \|\mathbf{F}(\cdot,t)\|^2\,dt < \infty, \quad \nabla\cdot\mathbf{F} = 0. \tag{2.3}
$$

The pressure term $p(x,y,z,t)$ in (2.2) is determined up to a constant in terms of $\mathbf{u}$ by the linear elliptic problem

$$
\begin{aligned}
\Delta p = -\nabla\cdot((\mathbf{u}\cdot\nabla)\mathbf{U}) - \nabla\cdot((\mathbf{U}\cdot\nabla)\mathbf{u}) = -2w_y,\\
p_z(x,y,0,t) = \frac{1}{R}w_{zz}(x,y,0,t),\\
p_z(x,y,1,t) = \frac{1}{R}w_{zz}(x,y,1,t).
\end{aligned}
\tag{2.4}
$$

Moreover, if p is given by the problem above, the solution $\mathbf{u}$ of (2.2) remains divergence-free. Therefore, we drop the continuity equation and write (2.2) as the linear evolution equation

$$
\begin{aligned}
\mathbf{u}_t = \mathcal{L}_R\mathbf{u} + \mathbf{F},\\
\mathbf{u}(\mathbf{x},0) = (0,0,0),
\end{aligned}
\tag{2.5}
$$

where the linear operator $\mathcal{L}_R$ is defined by

$$
\mathcal{L}_R\mathbf{u} := \frac{1}{R}\Delta\mathbf{u} - (\mathbf{u}\cdot\nabla)\mathbf{U} - (\mathbf{U}\cdot\nabla)\mathbf{u} - \nabla p, \tag{2.6}
$$

with p given in terms of $\mathbf{u}$ by (2.4).

It was proven in [12] that all the eigenvalues of $\mathcal{L}_R$ have negative real part for all values of R and that the eigenvalue with the greatest real part is at least at a distance proportional to $\frac{1}{R}$ from the imaginary axis. Our aim is to estimate the L_2 norm of the resolvent $(s\mathcal{I}-\mathcal{L}_R)^{-1}$ of $\mathcal{L}_R$ on the unstable half-plane $\mathrm{Re}(s) \geq 0$, and to determine its dependence on R. Our results indicate the resolvent constant $\sup_{\mathrm{Re}(s)\geq 0}\|(s\mathcal{I}-\mathcal{L}_R)^{-1}\|$ to be proportional to R^2, which agrees with the direct numerical computations of [6, 17]. Our analysis clarifies the role played by each component of the function $\mathbf{u}$, and it allows us to determine the origin of the R^2 growth of the resolvent constant.

3. Estimates for the resolvent. For large $|s|$, estimates were already proved [1, 10]. We state Theorem 1 from [1].

THEOREM 3.1. *If* $\mathrm{Re}(s) \geq 0$, $|s| \geq 2\sqrt{2}(1+\sqrt{R})$, *then*

$$
\|(s\mathcal{I}-\mathcal{L}_R)^{-1}\|^2 \leq \frac{8}{|s|^2}(1+\sqrt{R})^2 \leq 1.
$$

Using these estimates and the maximum modulus theorem for holomorphic mappings in Banach spaces [3], one can prove (see [1]) that

$$\sup_{\mathrm{Re}(s)\geq 0} \|(s\mathcal{I}-\mathcal{L}_R)^{-1}\| = \sup_{\xi\in\mathbb{R}} \|(i\xi\mathcal{I}-\mathcal{L}_R)^{-1}\|. \tag{3.1}$$

Therefore, for our purposes, it is sufficient to consider $s = i\xi$ purely imaginary. Using this result one can easily prove (see [1]) the following corollary.

COROLLARY 3.2. *Let $s = i\xi$, $\xi \in \mathbb{R}$. If $|\xi| \geq 2(1+\sqrt{R})$, then*

$$\|(s\mathcal{I}-\mathcal{L}_R)^{-1}\|^2 \leq \frac{8}{|s|^2}(1+\sqrt{R})^2 \leq 1.$$

Hence, our aim is to estimate the resolvent $(s\mathcal{I}-\mathcal{L}_R)^{-1}$ for $s = i\xi$, $0 \leq |\xi| < 2(1+\sqrt{R})$. We write the problem (2.2) componentwise:

$$\begin{aligned} u_t + zu_y + p_x &= \frac{1}{R}\Delta u + F, \\ v_t + zv_y + w + p_y &= \frac{1}{R}\Delta v + G, \\ w_t + zw_y + p_z &= \frac{1}{R}\Delta w + H, \\ u_x + v_y + w_z &= 0, \\ u(\mathbf{x},0) = v(\mathbf{x},0) &= w(\mathbf{x},0) = 0, \end{aligned} \tag{3.2}$$

with u, v, w vanishing at $z = 0$, $z = 1$ and 2π periodic in both x and y directions. Taking the Laplace transform with respect to t of the equation in (2.5), we get the resolvent equation

$$s\widetilde{\mathbf{u}} = \mathcal{L}_R\widetilde{\mathbf{u}} + \widetilde{\mathbf{F}}. \tag{3.3}$$

Componentwise, the transformed problem is

$$\begin{aligned} s\widetilde{u} + z\widetilde{u}_y + \widetilde{p}_x &= \frac{1}{R}\Delta\widetilde{u} + \widetilde{F}, \\ s\widetilde{v} + z\widetilde{v}_y + \widetilde{w} + \widetilde{p}_y &= \frac{1}{R}\Delta\widetilde{v} + \widetilde{G}, \\ s\widetilde{w} + z\widetilde{w}_y + \widetilde{p}_z &= \frac{1}{R}\Delta\widetilde{w} + \widetilde{H}, \\ \widetilde{u}_x + \widetilde{v}_y + \widetilde{w}_z &= 0. \end{aligned} \tag{3.4}$$

Our aim is to get an estimate of the form

$$\|\widetilde{\mathbf{u}}(\cdot,s)\|^2 \leq CR^\gamma \|\widetilde{\mathbf{F}}(\cdot,s)\|^2 \tag{3.5}$$

for $\mathrm{Re}(s) \geq 0$, where C is an absolute constant. Since the most important part of the argument is to determine the exponent γ, we keep the notation simple by representing by C any absolute constant appearing in different parts of this work, possibly with different numerical values. We obtain $\gamma = 4$, which implies the norm of the resolvent to be proportional to R^2. Actually, our analysis shows that different components of the velocity have different dependence on R. We get

$$\begin{aligned} \|\widetilde{u}\|^2 &\leq CR^4\|\widetilde{\mathbf{F}}\|^2, \\ \|\widetilde{v}\|^2 &\leq CR^4\|\widetilde{\mathbf{F}}\|^2, \\ \|\widetilde{w}\|^2 &\leq CR^2\|\widetilde{\mathbf{F}}\|^2. \end{aligned} \tag{3.6}$$

The inequalities above provide some physical insight about the problem. For a given forcing, components of the perturbations which are parallel to the planes may grow as R^2, while the worst growth for the normal component is R.

To derive the estimates, we use the well-known equivalent formulation of the problem in terms of the normal velocity and the normal vorticity [13, 10]. The vorticity is defined by

$$\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3) := \mathbf{curl}\, \mathbf{u}. \tag{3.7}$$

The transformed normal component of the velocity $\widetilde{w}$ is the solution of

$$\begin{gathered} \left(s + z\frac{\partial}{\partial y}\right)\Delta\widetilde{w} = \frac{1}{R}\Delta^2\widetilde{w} + \Delta\widetilde{H}, \\ \widetilde{w}(x,y,0,s) = \widetilde{w}(x,y,1,s) = 0, \\ \widetilde{w}_z(x,y,0,s) = \widetilde{w}_z(x,y,1,s) = 0. \end{gathered} \tag{3.8}$$

The transformed normal component of the vorticity $\widetilde{\eta}_3$ satisfies

$$\begin{gathered} \left(s + z\frac{\partial}{\partial y}\right)\widetilde{\eta}_3 + \widetilde{w}_x = \frac{1}{R}\Delta\widetilde{\eta}_3 + \widetilde{G}_x - \widetilde{F}_y, \\ \widetilde{\eta}_3(x,y,0,s) = \widetilde{\eta}_3(x,y,1,s) = 0. \end{gathered} \tag{3.9}$$

Expand in a Fourier series in the x and y directions. We represent by k_1 and k_2 the respective parameters. Let $k^2 := k_1^2 + k_2^2$. The transformed functions $\widehat{w}$, $\widehat{\eta}_3$ are the solutions of the system

$$\begin{gathered} \frac{1}{R}\widehat{w}'''' - \left(s + \frac{2k^2}{R} + ik_2z\right)\widehat{w}'' + \left(sk^2 + \frac{k^4}{R} + ik_2k^2z\right)\widehat{w} = k^2\widehat{H} - \widehat{H}'', \\ \widehat{w}(k_1,k_2,0,s) = \widehat{w}(k_1,k_2,1,s) = \widehat{w}'(k_1,k_2,0,s) = \widehat{w}'(k_1,k_2,1,s) = 0, \end{gathered} \tag{3.10}$$

and

$$\begin{gathered} \frac{1}{R}\widehat{\eta}_3'' - \left(s + \frac{k^2}{R} + ik_2z\right)\widehat{\eta}_3 = ik_1\widehat{w} + ik_2\widehat{F} - ik_1\widehat{G}, \\ \widehat{\eta}_3(k_1,k_2,0,s) = \widehat{\eta}_3(k_1,k_2,1,s) = 0. \end{gathered} \tag{3.11}$$

In the problems above, $'$ denotes the derivative with respect to z. The equations in (3.10) and (3.11) are, respectively, the classical Orr–Sommerfeld and Squire equations [13, 11, 14, 15]. The transformed normal velocity $\widehat{w}$, solution of (3.10), acts as a forcing term in the equation of the transformed normal vorticity (3.11). To simplify the notation, we define the differential operators T, T_0 by

$$\begin{gathered} T := \frac{1}{R}\mathcal{D}^2 - \left(s + \frac{k^2}{R} + ik_2z\right), \\ T_0 := \mathcal{D}^2 - k^2, \end{gathered} \tag{3.12}$$

where $\mathcal{D}$ denotes the derivative with respect to z. Then, the differential equation in (3.10) is written as

$$TT_0\widehat{w} = N := k^2\widehat{H} - \widehat{H}''. \tag{3.13}$$

The equation for the transformed normal vorticity is

$$T\widehat{\eta}_3 = ik_1\widehat{w} + ik_2\widehat{F} - ik_1\widehat{G}. \tag{3.14}$$

Lemma 3.3 follows directly from Parseval's identity.

LEMMA 3.3. *If*

$$\|\widehat{\mathbf{u}}(k_1,k_2,\cdot,s)\|^2 \le CR^\gamma \|\widehat{\mathbf{F}}(k_1,k_2,\cdot,s)\|^2 \tag{3.15}$$

for all $(k_1,k_2)\in\mathbb{Z}\times\mathbb{Z}$ *and for all* $s\in\mathbb{C}$, $\mathrm{Re}(s)\ge 0$, *then*

$$\|(s\mathcal{I}-\mathcal{L}_R)^{-1}\|^2 \le CR^\gamma, \quad \forall s\in\mathbb{C},\ \mathrm{Re}(s)\ge 0. \tag{3.16}$$

Therefore, we aim for an estimate of the form (3.15). We begin by estimating the normal velocity.

3.1. Estimates for the normal velocity. We separate the analysis into three cases: $k^2 \ge \frac{R}{\sqrt{2}}$, $k=0$, and $0<k^2<\frac{R}{\sqrt{2}}$.

Case $k^2 \ge \frac{R}{\sqrt{2}}$. The transformed normal velocity is the solution of problem (3.10). By Corollary 3.2, we need only to consider $s=i\xi$, $\xi\in\mathbb{R}$, $0\le|\xi|<2(1+\sqrt{R})$. Therefore, (3.10) reads as

$$\begin{aligned} &\frac{1}{R}\widehat{w}'''' - \left(i\xi + \frac{2k^2}{R} + ik_2z\right)\widehat{w}'' + \left(i\xi k^2 + \frac{k^4}{R} + ik_2k^2z\right)\widehat{w} = N,\\ &\widehat{w}(k_1,k_2,0,\xi) = \widehat{w}(k_1,k_2,1,\xi) = \widehat{w}'(k_1,k_2,0,\xi) = \widehat{w}'(k_1,k_2,1,\xi) = 0, \end{aligned} \tag{3.17}$$

where $N = k^2\widehat{H} - \widehat{H}''$. We prove the following theorem.

THEOREM 3.4. *If* $k^2 \ge \frac{R}{\sqrt{2}}$, *then*

$$\begin{aligned} &\|\widehat{w}''(k_1,k_2,\cdot,s)\|^2 + (k^2+k_1^2)\|\widehat{w}'(k_1,k_2,\cdot,s)\|^2 + k^4\|\widehat{w}(k_1,k_2,\cdot,s)\|^2\\ &\quad\le CR^2\|\widehat{H}(k_1,k_2,\cdot,s)\|^2. \end{aligned}$$

Proof. Taking the inner product of the differential equation in (3.17) with $\widehat{w}$ and integrating by parts, one obtains

$$\begin{aligned} &\frac{1}{R}\|\widehat{w}''\|^2 + \left(\frac{2k^2}{R} + i\xi\right)\|\widehat{w}'\|^2 + \left(i\xi k^2 + \frac{k^4}{R}\right)\|\widehat{w}\|^2\\ &\qquad + ik_2\langle\widehat{w},\widehat{w}'\rangle + ik_2\langle\widehat{w}', z\widehat{w}'\rangle + ik_2k^2\langle\widehat{w}, z\widehat{w}\rangle = \langle\widehat{w}, N\rangle. \end{aligned} \tag{3.18}$$

As can be easily checked through integration by parts, $\langle\widehat{w},\widehat{w}'\rangle$ is purely imaginary, and $\langle\widehat{w}, z\widehat{w}\rangle$, $\langle\widehat{w}', z\widehat{w}'\rangle$ are both real. Hence, taking the real part of (3.18) and using the triangle inequality, one gets

$$\frac{1}{R}\|\widehat{w}''\|^2 + \frac{2k^2}{R}\|\widehat{w}'\|^2 + \frac{k^4}{R}\|\widehat{w}\|^2 - |k_2|\,|\langle\widehat{w},\widehat{w}'\rangle| \le |\langle\widehat{w}, N\rangle|. \tag{3.19}$$

We note that (3.19) is valid for all values of the parameters.

If $k_2\neq 0$, use the inequality

$$|\langle\widehat{w},\widehat{w}'\rangle| \le \frac{R}{4|k_2|}\|\widehat{w}\|^2 + \frac{|k_2|}{R}\|\widehat{w}'\|^2$$

to get

$$\frac{1}{R}\|\widehat{w}''\|^2 + \left(\frac{2k^2}{R} - \frac{k_2^2}{R}\right)\|\widehat{w}'\|^2 + \left(\frac{k^4}{R} - \frac{R}{4}\right)\|\widehat{w}\|^2 \le |\langle \widehat{w}, N\rangle|. \tag{3.20}$$

Since $k^2 \ge \frac{R}{\sqrt{2}}$ implies $\frac{k^4}{R} - \frac{R}{4} \ge \frac{k^4}{2R}$, inequality (3.20) gives

$$\frac{1}{R}\|\widehat{w}''\|^2 + \frac{k^2 + k_1^2}{R}\|\widehat{w}'\|^2 + \frac{k^4}{2R}\|\widehat{w}\|^2 \le |\langle \widehat{w}, N\rangle|. \tag{3.21}$$

The desired estimates follow from this inequality. To derive them, we first note that the differential equation in (3.17) is linear. Therefore, if $\widehat{w}_1$, $\widehat{w}_2$ are the solutions of

$$\begin{aligned} TT_0\widehat{w}_1 &= k^2\widehat{H}, \\ TT_0\widehat{w}_2 &= -\widehat{H}'', \end{aligned}$$

both satisfying the same boundary conditions as $\widehat{w}$, then $\widehat{w} = \widehat{w}_1 + \widehat{w}_2$. We prove estimates for $\widehat{w}_1$ and $\widehat{w}_2$.

Using inequality (3.21) for $\widehat{w}_1$, and the Cauchy–Schwarz inequality, one gets

$$\frac{1}{R}\|\widehat{w}_1''\|^2 + \frac{k^2 + k_1^2}{R}\|\widehat{w}_1'\|^2 + \frac{k^4}{2R}\|\widehat{w}_1\|^2 \le k^2\|\widehat{w}_1\|\|\widehat{H}\|. \tag{3.22}$$

This inequality implies

$$\|\widehat{w}_1''\|^2 + (k^2 + k_1^2)\|\widehat{w}_1'\|^2 + k^4\|\widehat{w}_1\|^2 \le CR^2\|\widehat{H}\|^2. \tag{3.23}$$

For $\widehat{w}_2$, first note that

$$\langle \widehat{w}_2, \widehat{H}''\rangle = \langle \widehat{w}_2'', \widehat{H}\rangle, \tag{3.24}$$

since the boundary conditions satisfied by $\widehat{w}_2$ imply that the boundary terms after integration by parts vanish. Therefore, using (3.21), and the Cauchy–Schwarz inequality, we get

$$\frac{1}{R}\|\widehat{w}_2''\|^2 + \frac{k^2 + k_1^2}{R}\|\widehat{w}_2'\|^2 + \frac{k^4}{2R}\|\widehat{w}_2\|^2 \le \|\widehat{w}_2''\|\|\widehat{H}\|. \tag{3.25}$$

This inequality implies

$$\|\widehat{w}_2''\|^2 + (k^2 + k_1^2)\|\widehat{w}_2'\|^2 + k^4\|\widehat{w}_2\|^2 \le CR^2\|\widehat{H}\|^2. \tag{3.26}$$

Since $\widehat{w} = \widehat{w}_1 + \widehat{w}_2$, inequalities (3.23) and (3.26) imply

$$\|\widehat{w}''\|^2 + (k^2 + k_1^2)\|\widehat{w}'\|^2 + k^4\|\widehat{w}\|^2 \le CR^2\|\widehat{H}\|^2. \tag{3.27}$$

If $k_2 = 0$, then $k_1 \ne 0$ and inequality (3.19) is

$$\frac{1}{R}\|\widehat{w}''\|^2 + \frac{2k_1^2}{R}\|\widehat{w}'\|^2 + \frac{k_1^4}{R}\|\widehat{w}\|^2 \le |\langle \widehat{w}, N\rangle|.$$

From this inequality, estimates follow by the same argument as above, with no restriction on k_1. □

Case $k = 0$. In this case, we prove the following.

THEOREM 3.5. *If $k = 0$, we have*

$$\|\widehat{w}''(0,0,\cdot,s)\|^2 + \|\widehat{w}'(0,0,\cdot,s)\|^2 + \|\widehat{w}(0,0,\cdot,s)\|^2 \le CR^2\|\widehat{H}(0,0,\cdot,s)\|^2. \tag{3.28}$$

Proof. For this case, problem (3.17) is

$$\begin{gathered} \frac{1}{R}\widehat{w}'''' - i\xi\widehat{w}'' = -\widehat{H}'', \\ \widehat{w}(0,0,0,s) = \widehat{w}(0,0,1,s) = \widehat{w}'(0,0,0,s) = \widehat{w}'(0,0,1,s) = 0, \end{gathered} \tag{3.29}$$

where $s = i\xi$. Taking the inner product of the equation with $\widehat{w}$ and integrating by parts, one gets

$$\frac{1}{R}\|\widehat{w}''\|^2 + i\xi\|\widehat{w}'\|^2 = -\langle\widehat{w}'', \widehat{H}\rangle. \tag{3.30}$$

Taking the real part of this equation, and using the Cauchy–Schwarz inequality on its right-hand side, we obtain

$$\|\widehat{w}''\|^2 \le R^2\|\widehat{H}\|^2. \tag{3.31}$$

Application of Poincaré's inequality twice gives us the estimate

$$\|\widehat{w}''\|^2 + \|\widehat{w}'\|^2 + \|\widehat{w}\|^2 \le CR^2\|\widehat{H}\|^2. \qquad \square \tag{3.32}$$

Case $0 < k^2 < \frac{R}{\sqrt{2}}$. For this case, we show that the problem can be reduced to estimating the solutions of linear homogeneous ordinary differential equations with nonhomogeneous boundary conditions. The method used here is similar to the approach in [1] to estimate the stream function for the case of two space dimensions.

THEOREM 3.6. *If, for all $R \ge 1$, the solutions $\phi_1(k_1,k_2,z,s)$ and $\phi_2(k_1,k_2,z,s)$ of*

$$\begin{array}{ll} TT_0\phi_1 = 0, & TT_0\phi_2 = 0, \\ \phi_1(k_1,k_2,0,s) = 0, & \phi_2(k_1,k_2,0,s) = 0, \\ \phi_1(k_1,k_2,1,s) = 0, & \phi_2(k_1,k_2,1,s) = 0, \\ \phi_1'(k_1,k_2,0,s) = 1, & \phi_2'(k_1,k_2,0,s) = 0, \\ \phi_1'(k_1,k_2,1,s) = 0, & \phi_2'(k_1,k_2,1,s) = 1 \end{array} \tag{3.33}$$

satisfy

$$\begin{array}{ll} |k|\|\phi_1(k_1,k_2,\cdot,s)\|^2 \le C & |k|\|\phi_2(k_1,k_2,\cdot,s)\|^2 \le C \\ \|\phi_1'(k_1,k_2,\cdot,s)\|^2 \le C & \|\phi_2'(k_1,k_2,\cdot,s)\|^2 \le C \end{array} \tag{3.34}$$

for some absolute constant $C > 0$ and for all $0 < k^2 < \frac{R}{\sqrt{2}}$, $s = i\xi$, $\xi \in \mathbb{R}$, $0 \le |\xi| < 2(1+\sqrt{R})$, then

$$|k|\|\widehat{w}'(k_1,k_2,\cdot,s)\|^2 + k^2\|\widehat{w}(k_1,k_2,\cdot,s)\|^2 \le CR^2\|\widehat{\mathbf{F}}\|^2 \tag{3.35}$$

for all $R \ge 1$, $0 < k^2 < \frac{R}{\sqrt{2}}$, $s = i\xi$, $\xi \in \mathbb{R}$, $0 \le |\xi| < 2(1+\sqrt{R})$.

Proof. The transformed normal velocity $\widehat{w}$ is the solution of

$$\begin{gathered} TT_0\widehat{w} = N, \\ \widehat{w}(0) = \widehat{w}(1) = \widehat{w}'(0) = \widehat{w}'(1) = 0, \end{gathered} \tag{3.36}$$

where $N = k^2\widehat{H} - \widehat{H}''$. To simplify the notation, we do not write explicitly the dependence of $\widehat{w}$ on all the parameters.

Let g and h be solution of the system

$$\begin{aligned} Th &= \left(\frac{1}{R}\mathcal{D}^2 - \left(i\xi + \frac{k^2}{R} + ik_2 z\right)\right) h = N, \quad h(0) = h(1) = 0, \\ T_0 g &= (\mathcal{D}^2 - k^2)g = h, \quad g(0) = g(1) = 0. \end{aligned} \tag{3.37}$$

Taking the inner product of the first equation with h, and integrating by parts, one gets

$$-\frac{1}{R}\|h'\|^2 - \frac{k^2}{R}\|h\|^2 - i\xi\|h\|^2 - ik_2\langle h, zh\rangle = \langle h, N\rangle. \tag{3.38}$$

Taking the real part of the equation above, and noting that $\langle h, zh\rangle \in \mathbb{R}$, we get

$$\frac{1}{R}\|h'\|^2 + \frac{k^2}{R}\|h\|^2 \leq |\langle h, N\rangle|. \tag{3.39}$$

As done before, since the equation satisfied by h in (3.37) is linear, we study separately h_1, h_2, the solutions of

$$\begin{aligned} Th_1 &= k^2\widehat{H}, \quad h_1(0) = h_1(1) = 0, \\ Th_2 &= -\widehat{H}'', \quad h_2(0) = h_2(1) = 0. \end{aligned} \tag{3.40}$$

For h_1, inequality (3.39) is

$$\frac{1}{R}\|h_1'\|^2 + \frac{k^2}{R}\|h_1\|^2 \leq |\langle h_1, k^2\widehat{H}\rangle| \leq k^2\|h_1\|\|\widehat{H}\|, \tag{3.41}$$

which implies

$$\|h_1\|^2 \leq R^2\|\widehat{H}\|^2. \tag{3.42}$$

For h_2, using inequality (3.39) and integrating by parts once, we have

$$\frac{1}{R}\|h_2'\|^2 + \frac{k^2}{R}\|h_2\|^2 \leq |\langle h_2, -\widehat{H}''\rangle| = |\langle h_2', -\widehat{H}'\rangle| \leq \|h_2'\|\|\widehat{H}'\|. \tag{3.43}$$

Therefore,

$$\|h_2'\|^2 + k^2\|h_2\|^2 \leq CR^2\|\widehat{H}'\|^2. \tag{3.44}$$

From (2.3), we have

$$\widehat{H}' = -ik_1\widehat{F} - ik_2\widehat{G}. \tag{3.45}$$

Therefore, (3.44) gives

$$\|h_2'\|^2 + k^2\|h_2\|^2 \leq CR^2(k_1^2\|\widehat{F}\|^2 + k_2^2\|\widehat{G}\|^2) \leq Ck^2R^2(\|\widehat{F}\|^2 + \|\widehat{G}\|^2), \tag{3.46}$$

which implies

$$\|h_2\|^2 \leq CR^2(\|\widehat{F}\|^2 + \|\widehat{G}\|^2). \tag{3.47}$$

Using (3.42) and (3.47), we conclude that $h = h_1 + h_2$ satisfies

$$\|h\|^2 \leq CR^2(\|\widehat{F}\|^2 + \|\widehat{G}\|^2 + \|\widehat{H}\|^2) = CR^2\|\widehat{\mathbf{F}}\|^2. \tag{3.48}$$

For g, estimates follow in a similar and simpler way. Taking the inner product of the second equation in (3.37) with g and integrating by parts, one can prove that

$$k^2\|g'\|^2 + k^4\|g\|^2 \leq C\|h\|^2. \tag{3.49}$$

Using the differential equation for g in (3.37), one can also bound g''. Therefore, one gets

$$\|g''\|^2 + k^2\|g'\|^2 + k^4\|g\|^2 \leq C\|h\|^2. \tag{3.50}$$

Using (3.48) and (3.50), we conclude that

$$\|g''\|^2 + k^2\|g'\|^2 + k^4\|g\|^2 \leq CR^2\|\widehat{\mathbf{F}}\|^2. \tag{3.51}$$

It follows from the definition of g that it satisfies

$$\begin{aligned} TT_0 g &= k^2\widehat{H} - \widehat{H}'', \\ g(0) &= g(1) = 0. \end{aligned} \tag{3.52}$$

Therefore, g satisfies the same differential equation satisfied by $\widehat{w}$, but with different boundary conditions, since $g'(0)$ and $g'(1)$ do not necessarily vanish. But those values can be estimated. Indeed, using the one-dimensional Sobolev inequality $|g'|_\infty^2 \leq \|g'\|^2 + 2\|g'\|\|g''\|$, and (3.51), we have

$$\begin{aligned} |k||g'(0)|^2 \leq |k||g'|_\infty^2 \leq |k|\|g'\|^2 + 2|k|\|g'\|\|g''\| \leq CR^2\|\widehat{\mathbf{F}}\|^2, \\ |k||g'(1)|^2 \leq |k||g'|_\infty^2 \leq |k|\|g'\|^2 + 2|k|\|g'\|\|g''\| \leq CR^2\|\widehat{\mathbf{F}}\|^2. \end{aligned} \tag{3.53}$$

Now, let ϕ be the solution of

$$\begin{aligned} &TT_0\phi = 0, \\ &\phi(0) = \phi(1) = 0, \\ &\phi'(0) = g'(0), \\ &\phi'(1) = g'(1). \end{aligned} \tag{3.54}$$

Then, $\widehat{w} = g - \phi$, as can be easily checked. Since we already have estimates for g, estimates for ϕ will imply estimates for $\widehat{w}$. Now, note that if ϕ_1 and ϕ_2 are the solutions of

$$\begin{array}{cc} TT_0\phi_1 = 0, & TT_0\phi_2 = 0, \\ \phi_1(0) = \phi_1(1) = 0, & \phi_2(0) = \phi_2(1) = 0, \\ \phi_1'(0) = 1, & \phi_2'(0) = 0, \\ \phi_1'(1) = 0, & \phi_2'(1) = 1, \end{array} \tag{3.55}$$

then $\phi = g'(0)\phi_1 + g'(1)\phi_2$. Therefore, if for some absolute constant C we have

$$\begin{array}{cc} |k|\|\phi_1(k_1,k_2,\cdot,s)\|^2 \leq C, & |k|\|\phi_2(k_1,k_2,\cdot,s)\|^2 \leq C, \\ \|\phi_1'(k_1,k_2,\cdot,s)\|^2 \leq C, & \|\phi_2'(k_1,k_2,\cdot,s)\|^2 \leq C, \end{array} \tag{3.56}$$

then, using (3.53), we get

$$\begin{aligned} k^2\|\phi\|^2 &\le 2k^2|g'(0)|^2\|\phi_1\|^2 + 2k^2|g'(1)|^2\|\phi_2\|^2 \le CR^2\|\widehat{\mathbf{F}}\|^2, \\ |k|\|\phi'\|^2 &\le 2|k||g'(0)|^2\|\phi_1'\|^2 + 2|k||g'(1)|^2\|\phi_2'\|^2 \le CR^2\|\widehat{\mathbf{F}}\|^2. \end{aligned} \tag{3.57}$$

Since $\widehat{w} = g - \phi$, inequalities (3.51) and (3.57) imply

$$|k|\|\widehat{w}'\|^2 + k^2\|\widehat{w}\|^2 \le CR^2\|\widehat{\mathbf{F}}\|^2, \tag{3.58}$$

which proves the theorem. ∎

We study the solutions ϕ_1 and ϕ_2 of (3.55) numerically. These problems are suitable for a numerical approach for two main reasons: First, they are homogeneous problems, with fixed nonhomogeneous boundary conditions for all values of the parameters k_1, k_2, ξ, R. Second, they need to be studied only for bounded values of k_1, k_2, and s, namely for $0 < k^2 < \frac{R}{\sqrt{2}}$, and $s = i\xi$, $\xi \in \mathbb{R}$, $0 \le |\xi| < 2(1 + \sqrt{R})$. The results are shown in section 4, providing evidence for the bounds (3.56).

Therefore, from the three cases studied above, we conclude that for all values of the parameters k_1, k_2, and s, we have

$$\begin{aligned} |k|\|\widehat{w}'\|^2 + k^2\|\widehat{w}\|^2 \le CR^2\|\widehat{\mathbf{F}}\|^2, \quad k^2 \neq 0, \\ \|\widehat{w}'\|^2 + \|\widehat{w}\|^2 \le CR^2\|\widehat{\mathbf{F}}\|^2, \quad k = 0. \end{aligned} \tag{3.59}$$

Having bounds for the normal velocity $\widehat{w}$, we now derive the bounds for the normal vorticity and use them to estimate $\widehat{u}$, $\widehat{v}$, the remaining components of the velocity.

3.2. Estimates for the normal vorticity. We prove the following theorem.

THEOREM 3.7. *If the estimates* (3.59) *hold, then*

$$\|\widehat{\eta}_3'\|^2 + k^2\|\widehat{\eta}_3\|^2 \le CR^4\|\widehat{\mathbf{F}}\|^2. \tag{3.60}$$

Moreover, inequality (3.60) *implies*

$$\|\widehat{u}\|^2 + \|\widehat{v}\|^2 \le CR^4\|\widehat{\mathbf{F}}\|^2. \tag{3.61}$$

Proof. The function $\widehat{\eta}_3$ is the solution of

$$\begin{aligned} T\widehat{\eta}_3 = \frac{1}{R}\widehat{\eta}_3'' - \left(i\xi + \frac{k^2}{R} + ik_2 z\right)\widehat{\eta}_3 = ik_1\widehat{w} + ik_2\widehat{F} - ik_1\widehat{G}, \\ \widehat{\eta}_3(k_1, k_2, 0, \xi) = \widehat{\eta}_3(k_1, k_2, 1, \xi) = 0. \end{aligned} \tag{3.62}$$

Taking the inner product of the differential equation with $\widehat{\eta}_3$, and integrating by parts the first term of the resulting equation once, we get

$$\frac{1}{R}\|\widehat{\eta}_3'\|^2 + \left(i\xi + \frac{k^2}{R}\right)\|\widehat{\eta}_3\|^2 + ik_2\langle\widehat{\eta}_3, z\widehat{\eta}_3\rangle = -ik_1\langle\widehat{\eta}_3, \widehat{w}\rangle - ik_2\langle\widehat{\eta}_3, \widehat{F}\rangle + ik_1\langle\widehat{\eta}_3, \widehat{G}\rangle.$$

Since $\langle\widehat{\eta}_3, z\widehat{\eta}_3\rangle \in \mathbb{R}$, taking the real part of the equation above and using the Cauchy–Schwarz inequality, we have

$$\frac{1}{R}\|\widehat{\eta}_3'\|^2 + \frac{k^2}{R}\|\widehat{\eta}_3\|^2 \le |k_1|\|\widehat{\eta}_3\|\|\widehat{w}\| + |k_2|\|\widehat{\eta}_3\|\|\widehat{F}\| + |k_1|\|\widehat{\eta}_3\|\|\widehat{G}\|. \tag{3.63}$$

If $k^2 = 0$, the desired estimates follow directly. For $k^2 \neq 0$, (3.63) implies

$$\frac{|k|}{R}\|\widehat{\eta}_3\| \leq \frac{|k_1|}{|k|}\|\widehat{w}\| + \frac{|k_2|}{|k|}\|\widehat{F}\| + \frac{|k_1|}{|k|}\|\widehat{G}\| \leq \|\widehat{w}\| + \|\widehat{F}\| + \|\widehat{G}\| \leq CR\|\widehat{\mathbf{F}}\|,$$

where we used (3.59) to bound $\|\widehat{w}\|$. Therefore,

$$k^2\|\widehat{\eta}_3\|^2 \leq CR^4\|\widehat{\mathbf{F}}\|^2. \tag{3.64}$$

Using (3.63) and (3.64), we can bound $\widehat{\eta}_3'$ by

$$\|\widehat{\eta}_3'\|^2 \leq CR^4\|\widehat{\mathbf{F}}\|^2. \tag{3.65}$$

Inequalities (3.64) and (3.65) together give

$$\|\widehat{\eta}_3'\|^2 + k^2\|\widehat{\eta}_3\|^2 \leq CR^4\|\widehat{\mathbf{F}}\|^2, \tag{3.66}$$

which proves the first part of the theorem.

We now use (3.64) to bound $\widehat{u}$, $\widehat{v}$, components of the velocity. The velocity components u and v can be recovered once one knows the normal velocity w and normal vorticity η_3 by solving, with appropriate boundary conditions, the equations

$$-u_{xx} - u_{yy} = \eta_{3y} + w_{xz}, \tag{3.67}$$

$$-v_{xx} - v_{yy} = w_{yz} - \eta_{3x}. \tag{3.68}$$

For the transformed functions, the equations above are

$$k^2\widehat{u} = ik_2\widehat{\eta}_3 + ik_1\widehat{w}', \tag{3.69}$$

$$k^2\widehat{v} = ik_2\widehat{w}' - ik_1\widehat{\eta}_3. \tag{3.70}$$

Using (3.59) and (3.64), the estimates

$$k^2\|\widehat{u}\| \leq CR^2\|\widehat{\mathbf{F}}\|, \tag{3.71}$$

$$k^2\|\widehat{v}\| \leq CR^2\|\widehat{\mathbf{F}}\| \tag{3.72}$$

follow. □

Inequalities (3.59), (3.71), (3.72) and Corollary 3.2 together imply

$$\begin{aligned}\|\widehat{\mathbf{u}}(k_1,k_2,\cdot,s)\|^2 &= \|\widehat{u}(k_1,k_2,\cdot,s)\|^2 + \|\widehat{v}(k_1,k_2,\cdot,s)\|^2 + \|\widehat{w}(k_1,k_2,\cdot,s)\|^2 \\ &\leq CR^4\|\widehat{\mathbf{F}}(k_1,k_2,\cdot,s)\|^2 + CR^4\|\widehat{\mathbf{F}}(k_1,k_2,\cdot,s)\|^2 + CR^2\|\widehat{\mathbf{F}}(k_1,k_2,\cdot,s)\|^2 \\ &\leq CR^4\|\widehat{\mathbf{F}}(k_1,k_2,\cdot,s)\|^2\end{aligned}$$

for all $(k_1,k_2) \in \mathbb{Z}\times\mathbb{Z}$ and for all $s \in \mathbb{C}$, $\mathrm{Re}(s) \geq 0$. By Lemma 3.3, this implies the resolvent estimate

$$\|(s\mathcal{I} - \mathcal{L}_R)^{-1}\|^2 \leq CR^4 \quad \forall s \in \mathbb{C},\ \mathrm{Re}(s) \geq 0. \tag{3.73}$$

Remarks about weighted norms. In [10], the authors define a weighted norm $\|\cdot\|_3$, which is given in our coordinate system by

$$\|\widetilde{\mathbf{u}}\|_3^2 := \|\widetilde{u}\|^2 + \|\widetilde{v}\|^2 + R^2\|\widetilde{w}\|^2. \tag{3.74}$$

Via direct numerical computations, they conclude that

$$\|\widetilde{\mathbf{u}}\|_3^2 \leq CR^2\|\widetilde{\mathbf{F}}\|_3^2. \tag{3.75}$$

Our analysis shows that, if one gets estimates of the type (3.27) for all values of k_1, k_2, (that is, estimating the normal velocity by the normal component of the forcing only), inequality (3.75) follows. Indeed, in this case, the estimates for the normal vorticity would be

$$\|\widehat{\eta}_3'\|^2 + k^2\|\widehat{\eta}_3\|^2 \leq CR^2\|\widehat{F}\|^2 + CR^2\|\widehat{G}\|^2 + CR^4\|\widehat{H}\|^2.$$

Then, using (3.69) and (3.70),

$$\|\widehat{u}\|^2 \leq CR^2\|\widehat{F}\|^2 + CR^2\|\widehat{G}\|^2 + CR^4\|\widehat{H}\|^2,$$
$$\|\widehat{v}\|^2 \leq CR^2\|\widehat{F}\|^2 + CR^2\|\widehat{G}\|^2 + CR^4\|\widehat{H}\|^2.$$

Therefore,

$$\|\widehat{u}\|^2 + \|\widehat{v}\|^2 + R^2\|\widehat{w}\|^2 \leq CR^2\|\widehat{F}\|^2 + CR^2\|\widehat{G}\|^2 + CR^4\|\widehat{H}\|^2 = CR^2\|\widehat{\mathbf{F}}\|_3^2.$$

We believe this to be the case. In our argument though, we need to use all the components of the forcing $\widehat{\mathbf{F}}$ to bound $\widehat{w}$ for $0 < k^2 < \frac{R}{\sqrt{2}}$. We do not see how to overcome this technical difficulty at the moment. To get a better R growth for the perturbations using our estimates, we could define a weighted norm $\|\cdot\|_R$, scaling the components of $\widehat{\mathbf{u}}$ in the obvious way:

$$\|\widehat{\mathbf{u}}\|_R^2 := \frac{1}{R^2}\|\widehat{u}\|^2 + \frac{1}{R^2}\|\widehat{v}\|^2 + \|\widehat{w}\|^2. \tag{3.76}$$

For this norm, we have

$$\|\widehat{\mathbf{u}}\|_R^2 \leq CR^2\|\widehat{\mathbf{F}}\|^2, \tag{3.77}$$

but this does not imply a resolvent estimate, since we do not have the same norms on both sides of the inequality.

4. Numerical results. The only part of the argument that relies on numerical computations are the estimates for ϕ_1 and ϕ_2, solutions of

$$\frac{1}{R}\phi_1'''' - \left(i\xi + \frac{2k^2}{R} + ik_2z\right)\phi_1'' + \left(i\xi k^2 + \frac{k^4}{R} + ik_2k^2z\right)\phi_1 = 0,$$
$$\phi_1(k_1,k_2,0,\xi) = \phi_1(k_1,k_2,1,\xi) = 0,$$
$$\phi_1'(k_1,k_2,0,\xi) = 1,$$
$$\phi_1'(k_1,k_2,1,\xi) = 0,$$

and

$$\frac{1}{R}\phi_2'''' - \left(i\xi + \frac{2k^2}{R} + ik_2z\right)\phi_2'' + \left(i\xi k^2 + \frac{k^4}{R} + ik_2k^2z\right)\phi_2 = 0,$$
$$\phi_2(k_1,k_2,0,\xi) = \phi_2(k_1,k_2,1,\xi) = 0,$$
$$\phi_2'(k_1,k_2,0,\xi) = 0,$$
$$\phi_2'(k_1,k_2,1,\xi) = 1,$$

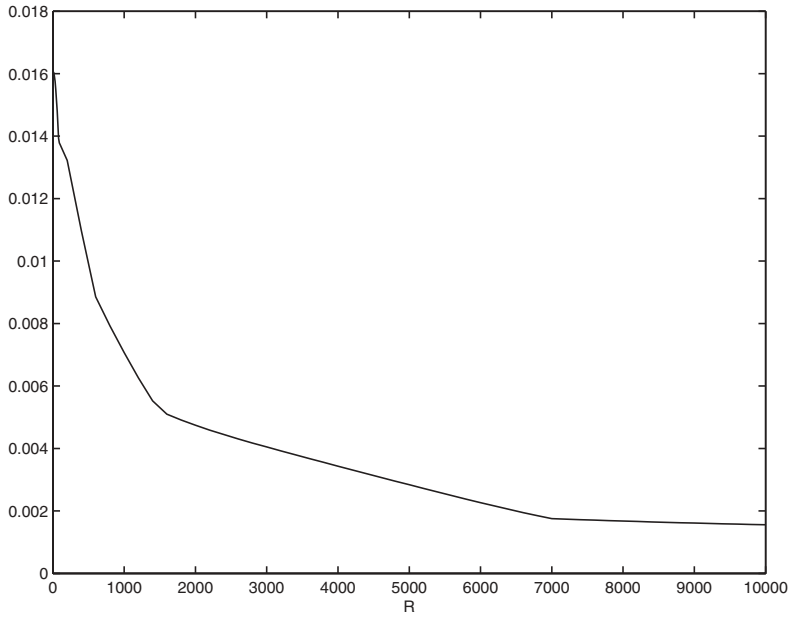


FIG. 1. $\max_{k_1,k_2,\xi} |k|\|\phi_1(k_1,k_2,\cdot,i\xi)\|^2$ *for* $0 < k_1^2 + k_2^2 < \frac{R}{\sqrt{2}}$, $0 \le |\xi| < 2(1+\sqrt{R})$.

for the parameter range

$$
\begin{aligned}
&(k_1,k_2) \in \mathbb{Z}, \quad 0 < k^2 = k_1^2 + k_2^2 < \frac{R}{\sqrt{2}},\\
&\xi \in \mathbb{R}, \quad 0 \le |\xi| < 2(1+\sqrt{R}).
\end{aligned}
\tag{4.1}
$$

We solved these problems using the MATLAB boundary value problem solver BVP4C, which makes use of a collocation method. A detailed description of the routine, and the methods used therein, can be found in [5]. For each value of R, we calculate the maximum of $|k|\|\phi_1(k_1,k_2,\cdot,\xi)\|^2$, $\|\phi_1'(k_1,k_2,\cdot,\xi)\|^2$, $\|\phi_2(k_1,k_2,\cdot,\xi)\|$, $\|\phi_2'(k_1,k_2,\cdot,\xi)\|$ for the parameter range (4.1). The results, for values of R up to 10000, are shown in Figures 1–4. The curves for ϕ_1 and ϕ_2 are very similar. Actually, for all considered values of R, the absolute value of the difference between the values in Figures 1 and 3 is of order 10^{-6}; the difference between the values in Figures 2 and 4 is of order 10^{-5}. These results are shown in Figures 5 and 6.

The numerical computations were performed with different absolute and relative tolerances, using continuation in the Reynolds number for the initial guess of the solution. The results were similar in all cases. Moreover, one just needs to ensure that the values of the norms above are bounded. Therefore, the results should be reliable. They indicate that, for all R,

$$
\begin{aligned}
|k|\|\phi_1(k_1,k_2,\cdot,s)\|^2 \le 1, \quad & |k|\|\phi_2(k_1,k_2,\cdot,s)\|^2 \le 1,\\
\|\phi_1'(k_1,k_2,\cdot,s)\|^2 \le 1, \quad & \|\phi_2'(k_1,k_2,\cdot,s)\|^2 \le 1
\end{aligned}
$$

for $0 < k^2 < \frac{R}{\sqrt{2}}$, $s = i\xi$, $\xi \in \mathbb{R}$, $0 \le |\xi| < 2(1+\sqrt{R})$.

5. Conclusions. The estimates derived here indicate the L_2 norm of the resolvent of the linear operator associated with three-dimensional perturbations of plane

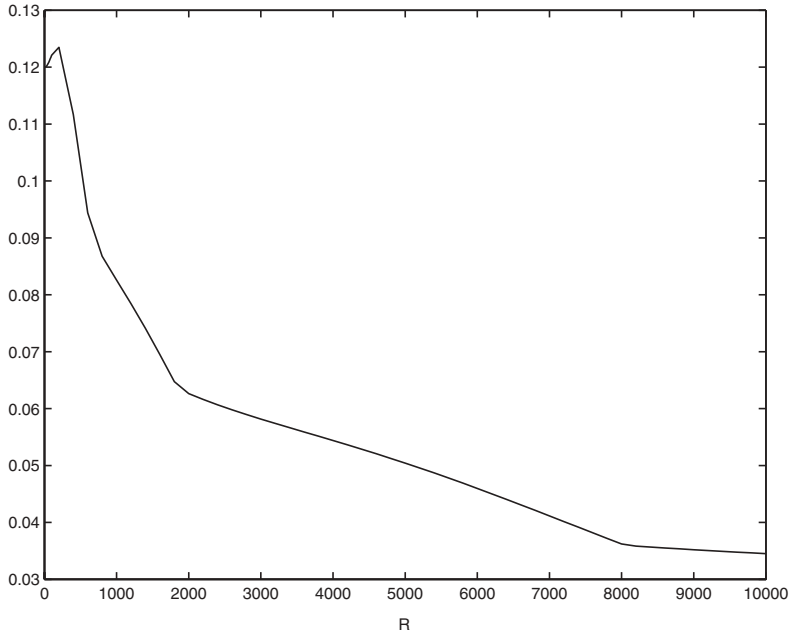


FIG. 2. $\max_{k_1,k_2,\xi} \|\phi_1'(k_1,k_2,\cdot,i\xi)\|^2$ *for* $0 < k_1^2 + k_2^2 < \frac{R}{\sqrt{2}}$, $0 \le |\xi| < 2(1+\sqrt{R})$.

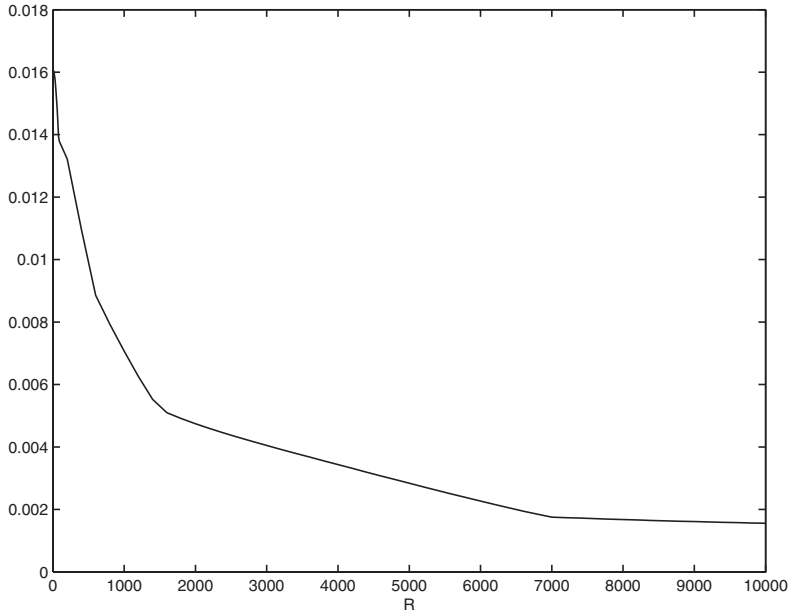


FIG. 3. $\max_{k_1,k_2,\xi} |k| \|\phi_2(k_1,k_2,\cdot,i\xi)\|^2$ *for* $0 < k_1^2 + k_2^2 < \frac{R}{\sqrt{2}}$, $0 \le |\xi| < 2(1+\sqrt{R})$.

Couette flow to be proportional to R^2 for the whole unstable half-plane $\mathrm{Re}(s) \ge 0$. They agree with previous numerical computations [6, 17]. In our argument though, numerical computations are used only to estimate the solutions of 4th-order

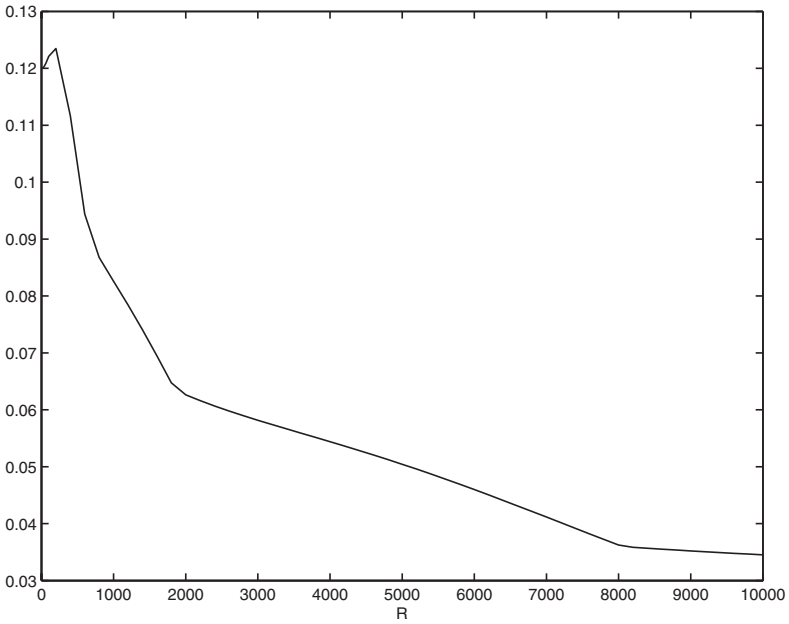


FIG. 4. $\max_{k_1,k_2,\xi} \|\phi_2'(k_1,k_2,\cdot,i\xi)\|^2$ *for* $0 < k_1^2 + k_2^2 < \frac{R}{\sqrt{2}}$, $0 \le |\xi| < 2(1+\sqrt{R})$.

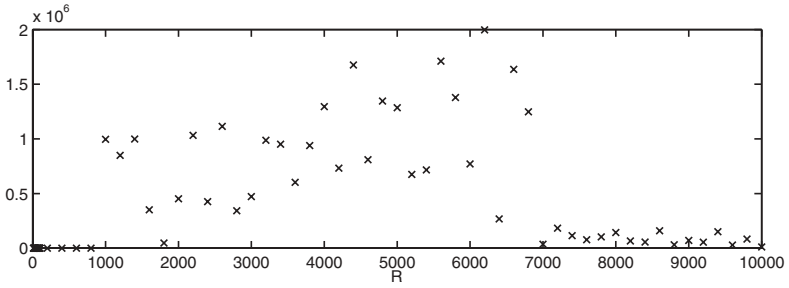


FIG. 5. $|\max_{k_1,k_2,\xi} |k| \|\phi_1(k_1,k_2,\cdot,i\xi)\|^2 - \max_{k_1,k_2,\xi} |k| \|\phi_2(k_1,k_2,\cdot,i\xi)\|^2|$.

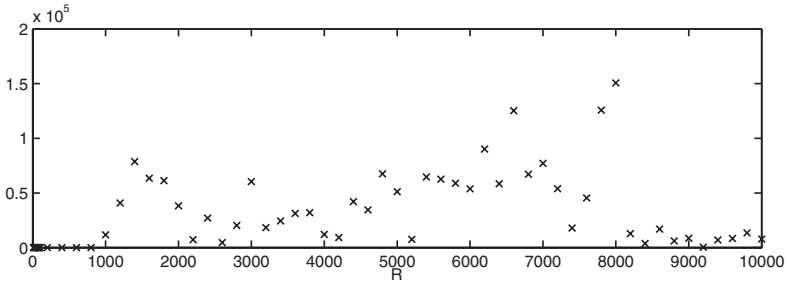


FIG. 6. $|\max_{k_1,k_2,\xi} \|\phi_1'(k_1,k_2,\cdot,i\xi)\|^2 - \max_{k_1,k_2,\xi} \|\phi_2'(k_1,k_2,\cdot,i\xi)\|^2|$.

homogeneous linear ordinary differential equations, with nonhomogeneous boundary conditions. Deriving the estimates analytically for the entire unstable half-plane is an

open problem, as far as we know. We believe that Theorem 3.6 may be useful towards a complete proof of the resolvent estimates. We hope to address this question in the future.

REFERENCES

[1] P. Braz e Silva, *Resolvent estimates for* 2*-dimensional perturbations of plane Couette flow*, Electron. J. Differential Equations, 2002 (2002), pp. 1–15.

[2] P. Braz e Silva, *Stability of Plane Couette Flow: The resolvent Method*, Ph.D. thesis, The University of New Mexico, Albuquerque, 2003.

[3] S. B. Chae, *Holomorphy and Calculus in Normed Spaces*, Monogr. Textbooks Pure Appl. Math. 92, Marcel Dekker, New York, 1985.

[4] F. Daviaud, J. Hagseth, and P. Bergé, *Subcritical transition to turbulence in plane Couette flow*, Phys. Rev. Lett., 69 (1992), pp. 2511–2514.

[5] J. Kierzenka and L. F. Shampine, *A BVP solver based on residual control and the MATLAB PSE*, ACM Trans. Math. Software, 27 (2001), pp. 299–316.

[6] G. Kreiss, A. Lundbladh, and D. S. Henningson, *Bounds for threshold amplitudes in subcritical shear flows*, J. Fluid Mech., 270 (1994), pp. 175–198.

[7] H.-O. Kreiss and J. Lorenz, *Stability for time dependent differential equations*, Acta Numer., 7 (1998), pp. 203–285.

[8] H.-O. Kreiss and J. Lorenz, *Resolvent estimates and quantification of nonlinear stability*, Acta Math. Sin. (Engl. Ser.), 16 (2000), pp. 1–20.

[9] M. Liefvendahl and G. Kreiss, *Bounds for threshold amplitudes in subcritical shear flows*, J. Nonlinear Math. Phys., 9 (2002), pp. 311–324.

[10] M. Liefvendahl and G. Kreiss, *Analytical and numerical investigation of the resolvent for plane Couette flow*, SIAM J. Appl. Math., 63 (2003), pp. 801–817.

[11] W. M. F. Orr, *The stability or instability of the steady motions of a perfect liquid and of a viscous liquid. part* i*: A perfect liquid. part* ii*: A viscous liquid*, Proc. Roy. Irish Acad. Sect. A, 27 (1907), pp. 9–138.

[12] V. A. Romanov, *Stability of plane-parallel Couette flow*, Funct. Anal. Appl., 7 (1973), pp. 137–146.

[13] P. J. Schmid and D. S. Henningson, *Stability and Transition in Shear Flows*, Appl. Math. Sci. 142, Springer-Verlag, New York, 2001.

[14] A. Sommerfeld, *Ein beitrag zur hydrodynamischen erklärung der turbulenten flüssigkeitbewegungen*, in Atti del 4. Congr. Internat. dei Mat. III, 1908, pp. 116–124, Roma.

[15] H. B. Squire, *On the stability for three-dimensional disturbances of viscous flow between parallel walls*, Proc. Roy. Soc. London Ser. A, 142 (1933), pp. 621–628.

[16] N. Tillmark and P. H. Alfredsson, *Experiments on transition in plane Couette flow*, J. Fluid Mech., 235 (1992), pp. 89–102.

[17] L. N. Trefethen, A. E. Trefethen, S. C. Reddy, and T. A. Driscoll, *Hydrodynamic stability without eigenvalues*, Science, 261 (1993), pp. 578–584.

SIAM J. APPL. MATH.
Vol. 65, No. 2, pp. 684–701

BOOMING AND CRASHING POPULATIONS AND EASTER ISLAND*

BILL BASENER† AND DAVID S. ROSS‡

Abstract. The population of Easter Island grew steadily for some time and then suddenly decreased dramatically. This is not the sort of behavior predicted by the usual logistic differential equation model of an isolated population or by the predator-prey model for a population using resources. We present a mathematical model that predicts this type of behavior when the growth rate of the resources, such as food and trees, is less than the rate at which resources are harvested. Our model is expressed mathematically as a system of two first-order differential equations, both of which are generalized logistic equations. Numerical solution of the equations, using realistic parameters, predicts values very close to archaeological observations of Easter Island. We analyze the model by using a coordinate transformation to blow up a singularity at the origin. Our analysis reveals surprisingly rich dynamics including a degenerate Hopf bifurcation.

Key words. Easter Island, population dynamics

AMS subject classifications. 92D02, 91D02, 37N02

DOI. 10.1137/S0036139903426952

1. The model.

At one time inhabitants of Easter Island prospered. They were sufficiently sophisticated, artistically and technologically, to build and transport the enormous mysterious statues for which the island is famous. Yet when westerners first came in contact with the island in the late eighteenth century, the inhabitants lived meagerly in flimsy huts and there were no trees left on the island. The island is extremely isolated, surrounded by over 1000 miles of ocean. Archaeological records indicate that a small group, about 50 to 150 people, sailed to Easter Island between 400 and 700 AD. The population grew to about 10,000 between 1200 and 1500 AD. It is thought that at this time the inhabitants built the biggest statues, had large boats, sailed on the ocean for fishing, and used the abundant large trees for building. The inhabitants overused the resources to the point of starvation and the island's human population decreased drastically. As a consequence of the population's actions, the large trees and other resources completely disappeared from the island. For a more detailed discussion of the history of Easter Island, see [4] and [8].

Neither of the standard elementary types of population models, logistic models and predator-prey models, predicts this sort of growth and decline. We present a system of differential equations for an isolated population that uses self-replenishing resources (such as trees, plants, and animals) which exhibits this booming and crashing behavior. We prove that if the population uses resources too quickly relative to the rate at which the resources replenish themselves, then the population will increase and then disappear in finite time. If the population uses the resources more slowly, then the population and resources do not disappear. A thorough characterization of solutions for various parameter values is given in Figure 5.

*Received by the editors April 30, 2003; accepted for publication (in revised form) April 19, 2004; published electronically January 27, 2005.
http://www.siam.org/journals/siap/65-2/42695.html
†Department of Mathematics and Statistics, Rochester Institute of Technology, College of Science, 85 Lomb Memorial Drive, Rochester, NY 14623 (wfbsma@vmsmail.rit.edu).
‡Department of Mathematics and Statistics, Rochester Institute of Technology, College of Science, 85 Lomb Memorial Drive, Rochester, NY 14623, and the Archimedes Project, Kaiser Permanente (dsrsma@rit.edu).

For our model, let P be a population and let R be the amount of resources. Our model is given by (1). To derive these equations, we assume that the resources would equilibrate in the absence of people. So when the population is zero the equation for the resources should be the standard logistic equation with $c, K > 0$. As in the standard logistic model we call c the growth rate of the resources and K the carrying capacity. The constant c has units of inverse time; it is the fraction by which the resource supply would increase per unit time were the resource supply far from the island's carrying capacity. The carrying capacity K has the same units as R; it is the maximum resource supply that the island can support.

The term $-hP$ accounts for the harvesting of the resources. The constant h, the harvesting constant, has units of reciprocal time; it is on the order of the reciprocal of the average lifetime of members of the population. The population P has units of persons, as does R; one unit of the resources is the amount of resources required to support a member of the population through his or her lifetime. We assume that the resources are accessible so that the amount of harvesting is proportional only to the population. This is a reasonable assumption for people on a small island.

At any given time the size of the population that our island can support depends on the amount of resources on the island. Given our choice of units, the island has the capacity to support R people. The evolution of the population is described by a logistic equation with the carrying capacity equal to R.

$$\begin{aligned} \frac{\mathrm{d}R}{\mathrm{d}t} &= cR\left(1-\frac{R}{K}\right) - hP, \\ \frac{\mathrm{d}P}{\mathrm{d}t} &= aP\left(1-\frac{P}{R}\right). \end{aligned} \tag{1}$$

The positive constant a has units of inverse time. The quantity aP is the net growth rate of the population in circumstances in which resources are abundant. Observe that when there are no resources ($R = 0$) the carrying capacity for the population is zero. This makes sense, but it causes mathematical trouble in the form of a singularity on the P-axis. The P/R term in (1) places our model in the class of ratio-dependent models, a class that has recently received much attention in the population biology literature. (See [9].) In fact, a discrete predator-prey model analogous to ours has been used by Eberhardt in [5].

The main virtues of this model are that it incorporates a variable carrying capacity for the population and that it is based on a simple but sensible account of the interaction between a population and its resources. Moreover, the predictions of this model match archeological data for the population of Easter Island; the predictions of standard models, such as the logistic model and the Lotka–Volterra model, do not. Of course a model's prediction matching data is not sufficient, though it is necessary, to establish the model's validity.

Our model is notable for the singularity in (1) when $R = 0$. Other models of populations similar to that of Easter Island do not involve such singularities; recent contributions to the literature have favored modified Lotka–Volterra models. Anderies [1] presents a general form for such models:

$$\begin{aligned} \frac{\mathrm{d}R}{\mathrm{d}t} &= \rho(R) - H(R,P), \\ \frac{\mathrm{d}P}{\mathrm{d}t} &= G(H,R)P. \end{aligned} \tag{2}$$

Here, $\rho(R)$ is the growth rate of the resources, $H = H(R, P)$ is the rate at which resources are harvested, and $G(H, R)P$ is the growth rate of the population; that is, G is the difference between the per person birth and death rates.

Brander and Taylor [3] choose the logistic form $\rho(R) = cR(1 - R/K)$ that we use in (1). Their harvesting rate is proportional to PR, while ours is simply proportional to P. For G, they use the linear function $G(H, R) = (b - d + \phi R)$, where b is the birth rate, d is the death rate, and ϕ is a constant. (The function $P(t)$, in Anderies's work and in the work of some other researchers in this field, is the labor pool, whereas we have considered, instead, the entire population.)

The harvesting model of Brander and Taylor accounts for the fact that as resources become scarce, less of the resources will be harvested per person. In our model, by contrast, the same amount of resources is harvested per person in all conditions. Consequently, our model does not capture details of low-resource conditions. The harvesting model, in which the harvesting rate is proportional to the amount of resources, seems to err in the other direction; it probably produces an underestimate of the harvesting rate in conditions of scarce resources. The truth is probably somewhere between the two models. While scarcity should diminish the harvesting rate, there will be a tendency for members of the population to maintain their standard of living at the cost of depleting resources. For conditions of plenty, our model seems sensible, and the assumption that the harvesting rate is proportional to the resources probably overestimates the harvesting rate.

Brander and Taylor use a population growth rate model in which the difference between the per person birth and death rates, $(b - d + \phi R)$, which is negative in the absence of resources, increases linearly with resources. In our population growth model, the difference between the per person birth and death rates is $a(1 - P/R)$, which is proportional to the unused fraction of the island's carrying capacity.

The per person growth rate of Brander and Taylor has the familiar mathematical form of an exponential decay model in conditions of scarcity. In the absence of resources, the population in the Brander and Taylor model dies out exponentially. Our model has the population, along with the resources, die out exponentially in some cases and in finite time in other cases. The appealing feature, in our model, of allowing the population to die out in finite time comes at the cost of an unbounded per person death rate. As resources increase, the model of Brander and Taylor has the difference between the per person birth and death rates become arbitrarily large. In our model, when resources are more than sufficient for the population, the difference between the per person birth and death rates approaches a finite positive constant.

Brander and Taylor derive their model in the framework of neoclassical economics; they justify the form of their harvesting rate by maximizing a Cobb–Douglas utility function. Anderies [2] takes a similar approach. He improves on their model by introducing a more general type of utility function, a Stone–Geary utility function. In this way, Anderies allows for a structural change in the culture when resources are scarce. He derives a continuous per person harvesting rate that is constant in conditions of scarcity and approaches a smaller constant asymptotically like $1/R$ in conditions of great abundance. This extra level of detail allows Anderies to fit the population data of Easter Island better than Brander and Taylor. (See graphs in [1] and [2].)

We have not embedded our model in neoclassical economic theory; we have simply made some plausible assumptions. We shall show, in section 2, that with reasonable values of the parameters, our model fits the archaeological data for Easter Island

closely. In the rest of the paper, we elaborate upon what we consider our model's other virtue: its exceptionally rich dynamics. We expect that this feature of the model makes it valuable as an example of the sorts of behavior that even a simple two-dimensional population model can exhibit.

Discussions of mathematical models like ours—models of the interaction of a human population and its resources—often include speculation about implications of the analysis for the population of the earth as a whole. We shall do this too, but with misgivings; models like ours do not capture the causes of the growth and advancement of modern technological societies. For example, one of the premises of our model is that resources grow and flourish independent of humans, that the only effect that the humans have on the resources is that the humans harvest them. This is a simplification even for the case of the Easter Islanders, who probably engaged in some of the cultivation of resources that is a hallmark of technological civilizations. For modern civilization, even the idea of resources as something given, apart from humans, is wrong; human ingenuity turns natural materials and phenomena into resources. Finally, at the most abstract level, models like ours do not even address the essential issues of the survival of a species that does things like construct mathematical models of its interaction with its environment.

That said, our model suggests a scenario not often considered for the overpopulation of the Earth. If a population overuses its resources (for our model, if $h > c$), the population will become large while the resources decrease. This situation results in a gradual exponential population growth for an extended period of time and then a sudden catastrophic elimination of the population. (See Figure 2.)

In section 2 we compare numerical approximations of solutions to archaeological data of Easter Island. In section 3 we prove our main theorem and describe general behavior of solutions.

2. Archaeological data of Easter Island and the world population. In this section we compare the population of Easter Island and the population predicted by a numerical solution of (1). We also provide a projection of the world population under the assumption the humans are overusing their resources. It is well accepted that numerical models do not provide accurate numbers for projecting human populations, in part because the constants (growth rate, etc.) for human populations depend on ever-changing social and technological factors. However, mathematical models do provide the approximate "shape" of the graph of a human population. We provide the Easter Island model in part as confirmation that the shape of solutions to (1) is reasonable for the human population and apply a solution with this shape to the world's human population.

A good summary of Easter Island history is given in Cohen's excellent text [4] on global population:

> The best current estimate is that the population began with a boatload of settlers in the first half millennium after Christ, perhaps around 400 A.D. The population remained low until about A.D. 1100. Growth then accelerated and the population then doubled every century until about 1400. Slower growth continued until at most 6000 to 8000 people occupied the island around 1600. The maximum population may have reached 10,000 people in A. D. 1680. A Decline then set in. Jean François de Galaup Comte de La Pérouse, who visited the island in 1786, estimated a population of 2000, and this estimate is now accepted as roughly correct.

The graph of population as a function of time for a numerical approximation to system 1 is shown in Figure 1. Note that the solution matches Cohen's historical estimate until around 1780. However, the population of Easter Island did not actually disappear as it does in the model. We expect that once the population became small enough, factors other than those considered in the model became important for the population. For example, records suggest that the people on Easter Island changed their diet to smaller animals and grasses after their larger ecosystem was destroyed.

In the numerical solution graphed in Figure 1 we used $a = .0044$, which is consistent with historical observations of developing countries prior to the second world war. We took the island's carrying capacity, K, to be $70,000$. It has been estimated (see [4]) that the amount of fertile land needed to supply food for one person is approximately 350 square meters, varying to a great degree depending on the type of land and climate. The area of Easter Island is approximately 166,000,000 square meters. If all of it were fertile and if it were farmed efficiently, there would be enough food for 475,000 people. Since only some of the land is farmable, this makes our approximation of $K = 70,000$ reasonable. The values $c = 0.001$ and $h = 0.025$ are more difficult to justify; we chose these values to fit the data. Note, however, that h is on the order of the reciprocal of a lifespan as suggested in section 1.

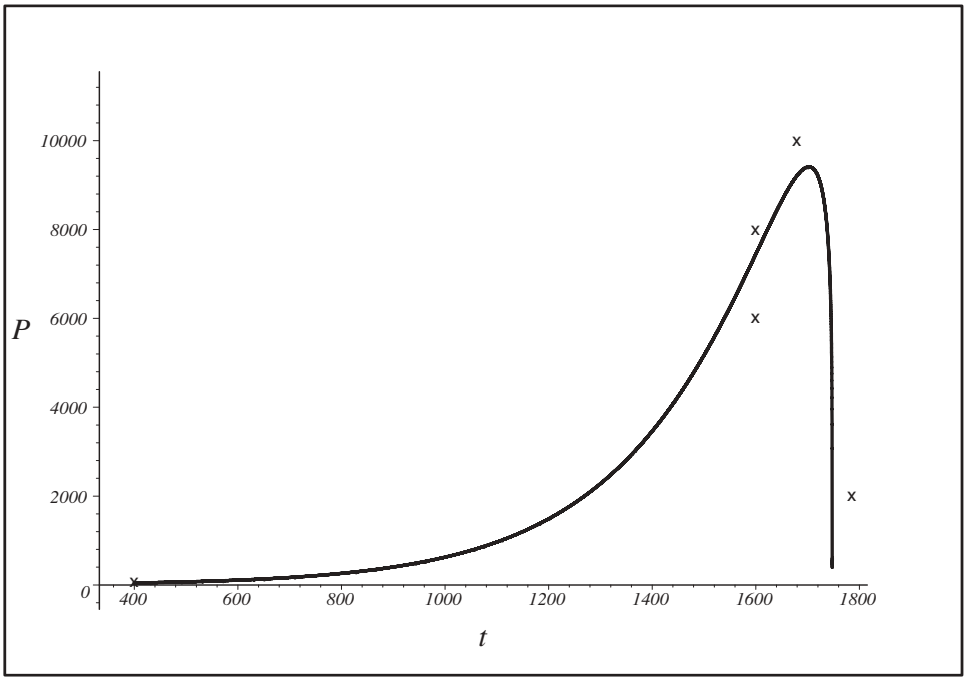


FIG. 1. *The graph of population versus time for a solution to* (1) *modeling the population of Easter Island. Each "x" is a data point approximated using archaeology.*

We do not claim that Figure 1 proves that the population of Easter Island evolved according to the dynamics of (1). But we think it suggests that these equations do provide a reasonable model for an isolated population with limited self-replenishing resources.

A numerical approximation of the world's population using (1) is shown in Figure 2. All of the units are in billions. We assume that the Earth's population in

the year 2000 is 6 (billion). For this approximation we use the carrying capacity of the Earth as $K = 1000$. Approximations to the carrying capacity of the Earth vary widely, as do definitions of what the carrying capacity means. Estimations vary from 1 billion to 1,000 billion, (see [4]), and we choose the upper limit. The growth rate of the Earth's human population has been in the range from 1.73 to over 2 (again, see [4]). We use a conservative estimate of $a = 1.5$. We choose c and h to model a situation where humans are barely overusing resources, $h = 0.6,\ c = 0.5$. The model suggests that the Earth's human population will grow steadily until it reaches a maximum of 350 billion in the year 2350, and then over the next 20 years the population will decrease until either extinction or another model, such as small local farmers, becomes appropriate. As stated earlier, we make no claim to the accuracy of these numbers other than that the prospect of a collapse of a population, instead of a gradual leveling off, is an important scenario to consider. Recall that by Figure 5, the long term behavior of the solution, extinction or equilibrium, depends only on the harvesting rate and the growth rate of the resources.

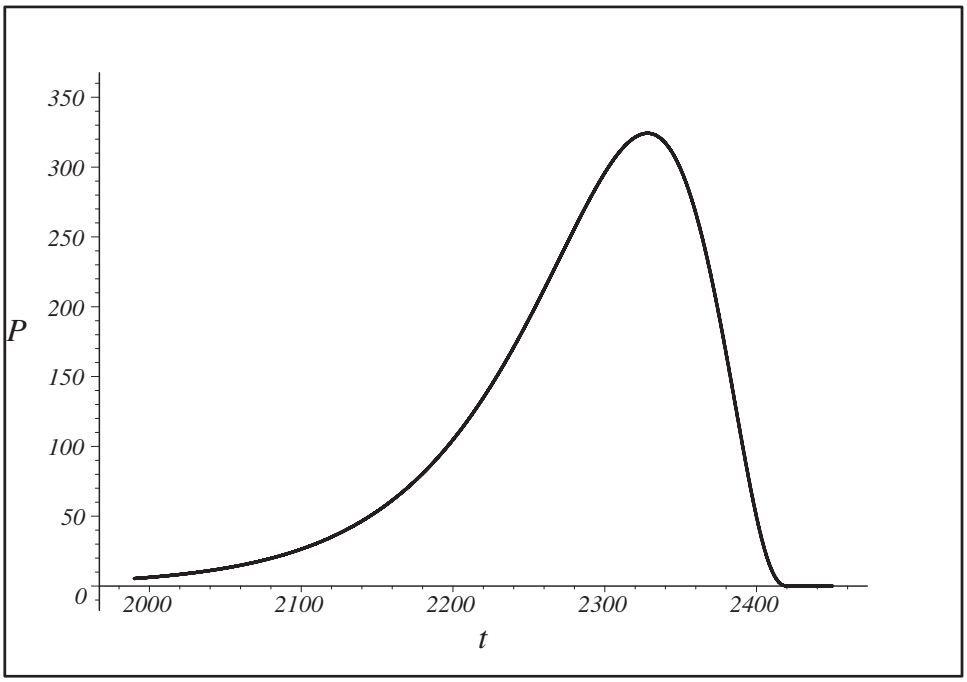


FIG. 2. *The graphs of population and resources versus time for a solution to* (1) *modeling the world's population.*

3. Analysis of the equations. Solutions of (1) fall into three qualitatively different categories: solutions that are asymptotic to an equilibrium point with $P, R > 0$, solutions that approach the singularity at $P = R = 0$, and periodic solutions. A characterization of solutions for various values of h and c is given in Figure 5.

The nullclines provide some insight into the behavior of the system. The nullcline on which $\mathrm{d}P/\mathrm{d}t = 0$ consists of the lines

$$P = 0, \quad P = R.$$

The nullcline on which $\mathrm{d}R/\mathrm{d}t = 0$ is the parabola $P = (c/h)R\left(1 - \frac{R}{K}\right)$ or

$$P = \frac{-c}{hK}R^2 + \frac{c}{h}R$$

(this can be determined by setting $\mathrm{d}R/\mathrm{d}t = 0$ and solving for P). These nullclines are shown, along with the direction field and a numerically integrated solution, in Figures 3 and 4. The parabola $P = (\frac{-c}{hK})R^2 + \frac{c}{h}R$ always intersects the line $P = 0$ at $(0,0)$ and $(K,0)$. The point $(K,0)$ is an equilibrium point, but the point $(0,0)$ is a singular point. The parabola $P = (\frac{-c}{hK})R^2 + \frac{c}{h}R$ intersects the line $P = R$ at $(0,0)$ and, if $c > h$, at $(\frac{K}{c}(c-h), \frac{K}{c}(c-h))$. A characterization of solutions is given in Figure 5.

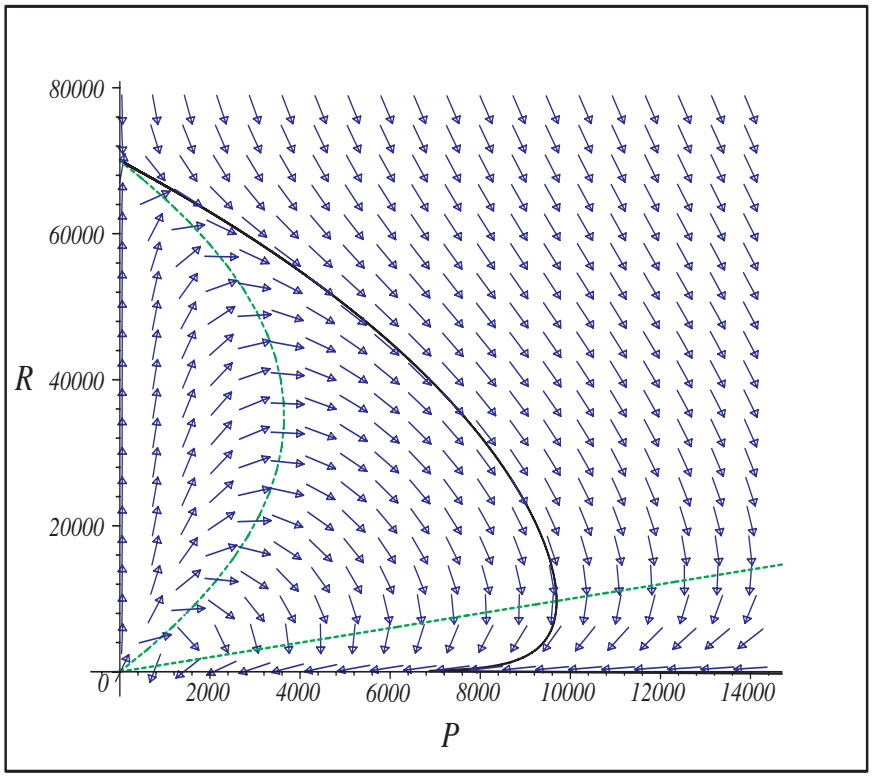


FIG. 3. *The direction field for* (1) *is shown along with the nullclines and one numerically integrated solution. The constants are* $a = .004, c = .01, K = 25000, h = .015$. *The initial condition for the solution is* $P = 75, R = K$, *which were approximately the values when settlers first landed on Easter Island.*

We simplify the equations without loss of generality by rescaling time so that $a = 1$. This puts the differential equations in the form

$$\frac{\mathrm{d}P}{\mathrm{d}t} = P\left(1 - \frac{P}{R}\right), \tag{3}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = cR\left(1 - \frac{R}{K}\right) - hP. \tag{4}$$

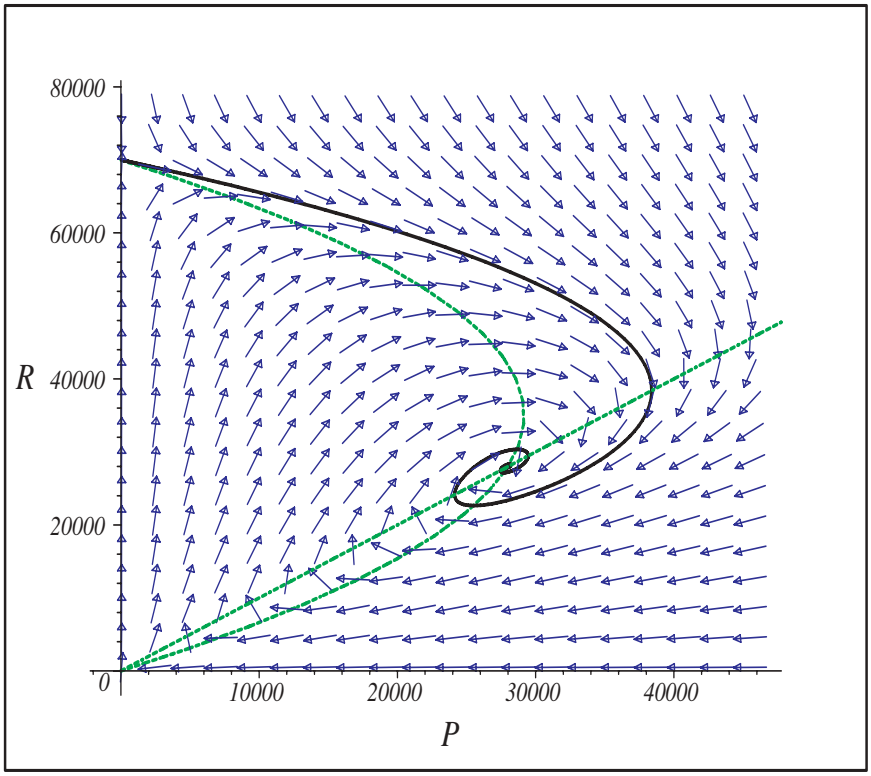


FIG. 4. *The direction field for* (1) *is shown along with the nullclines and one numerically integrated solution. The constants are* $a = .004, c = .01, K = 25000, h = .005$. *The initial condition for the solution is* $P = 75, R = K$, *which were approximately the values when settlers first landed on Easter Island. (Note that the singularity along the* P*-axis causes improperly drawn vectors along the* P*-axis.)*

We are most concerned about solutions that approach the origin, solutions that correspond to the disappearance of both the resources and the population. Standard local analysis near $(0, 0)$ is not possible because of the singularity there. We blow up this singularity through a change of variables. Let

$$z = P, \qquad \xi = P/R. \tag{5}$$

The equations in these new coordinates are

$$z' = z(1 - \xi), \qquad \xi' = (h - 1)\xi^2 + (1 - c)\xi + \frac{c}{K}z. \tag{6}$$

Note that the new system is free of singularities. We are most interested in values of the new coordinates for which P and R are both positive. For these values the change of variables is invertible. Note that the change of variables takes the first quadrant in P, R-coordinates to the first quadrant in z, ξ-coordinates, it takes vertical lines to themselves, and it is the identity mapping ($z = P, \xi = R$) along the parabola

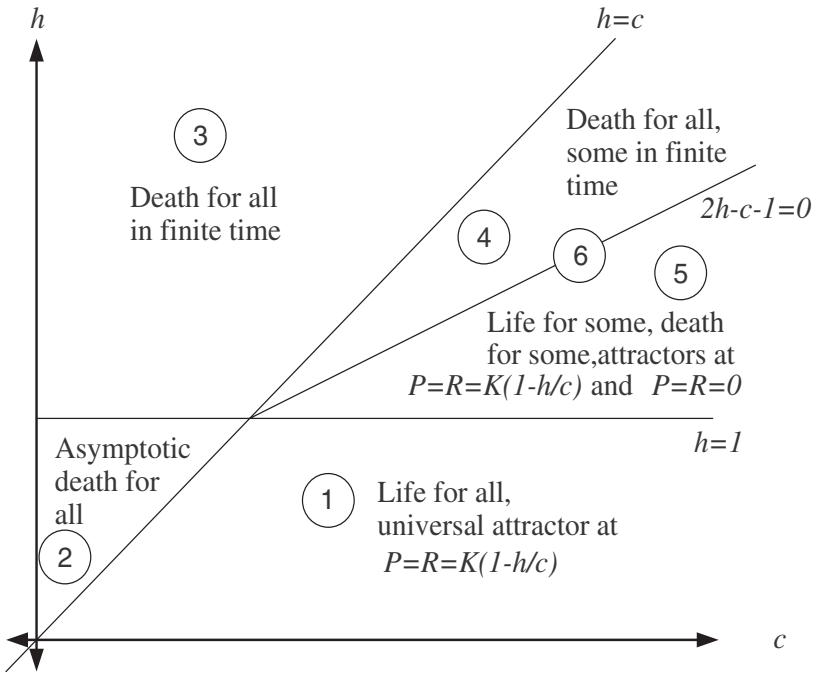


FIG. 5. *The long-term behavior of solutions depending on the growth rate of resources, c, and the harvesting rate, h. The number in each region indicates the corresponding proposition.*

$P = R^2$. Also, the region near $(0,0)$ with $P, R > 0$ has been "blown up" to the region near the positive ξ-axis. In particular, if an orbit approaches (0,0) with P and R asymptotically proportional to each other, the corresponding orbit in (z, ξ) will approach the point $(0, \Theta)$, where Θ is the asymptotic proportion. If P approaches 0 faster than R, we have $\xi \to 0$, and if P approaches 0 slower than P, we get $\xi \to \infty$. Behavior near the positive P-axis is obscured. However, behavior here is easy to understand in P, R-coordinates: the equation for the resources reduces to a logistic equation and the population grows.

We denote the first quadrant $\{(z, \xi) \,|\, z > 0, \xi > 0\}$ by Ω and denote the region $\{(z, \xi) \,|\, z \geq 0, \xi > 0\}$ by Ω^*. Although the positive ξ-axis does not correspond to distinct values of P and R—the whole axis corresponds to $P = R = 0$—we need it to analyze asymptotic behavior. Our main tool will be the Lyapunov function $\lambda : \Omega \to \mathbb{R}$,

$$\lambda(z, \xi) = z^{2h-2} \left(\frac{K}{2c} \xi^2 - \frac{K}{c} \xi + \frac{z}{2h-1} + \frac{K}{2c} - \frac{K(1-h/c)}{2h-2} \right),$$

the properties of which will be established in Lemma 2. First we establish some qualitative behavior of the system. By linearizing the system about the point $\{(z, \xi)\}$ we obtain the Jacobian

$$J = \begin{bmatrix} 1-\xi & z \\ \dfrac{c}{K} & 2(h-1)\xi + 1 - c \end{bmatrix}.$$

The equilibrium points of our system are

- $(0,0)$, at which J has eigenvalues 1 and $1-c$;
- $(0, \frac{c-1}{h-1})$, at which J has eigenvalues $\frac{h-c}{h-1}$ and $c-1$;
- $(K(1-h/c),1)$, at which J has eigenvalues $\frac{2h-c-1\pm\sqrt{(2h-c-1)^2-4(c-h)}}{2}$.

The behavior of solutions depends in a surprisingly complex way on the constants h and c. This dependence is summarized in Figure 5. Our main tools in establishing this characterization are Lemma 2, which establishes the properties of a Lyapunov function, and Lemma 3.

We shall begin our analysis with some lemmas about basic structural features of the system. We shall then use these lemmas to characterize the qualitative behavior of the system for various values of h and c.

LEMMA 1. *The regions Ω and Ω^* are both positive invariant. That is, if a solution is in one of these regions initially, then it remains in the region as long as it exists.*

Proof. The ξ-axis is invariant and the vector field is pointing into the first quadrant along the positive z-axis. Specifically, if $z>0$ and $\xi=0$, then $z'=z$ and $\xi'=\frac{c}{K}>0$. If $z=0$ and $\xi>0$, then $z'=0$ and $\xi'=(h-1)\xi^2+(1-c)\xi$. ∎

Note that the regions are not negative invariant and that it is possible that orbits become unbounded in finite time.

LEMMA 2. *Let*

$$\lambda(z,\xi)=z^{2h-2}\left(\frac{K}{2c}\xi^2-\frac{K}{c}\xi+\frac{z}{2h-1}+\frac{K}{2c}-\frac{K(1-h/c)}{2h-2}\right).$$

(a) *If $2h-c-1=0$, λ is constant on trajectories in Ω.*
(b) *If $2h-c-1<0$, λ is strictly decreasing on trajectories in Ω that are not equilibria.*
(c) *If $2h-c-1>0$, λ is strictly increasing on trajectories in Ω that are not equilibria.*

Proof. A direct (but not short) computation yields

$$\lambda'=(2h-c-1)(K/c)(\xi-1)^2z^{2h-2},$$

where $'$ denotes the derivative with respect to time. Statement (a) follows directly from this computation.

To prove (b), assume $2h-c-1<0$. Note that $\lambda'<0$ except when $\xi=1$. By differentiating twice more, we obtain

$$\begin{aligned}
\lambda'' = {} & (2h-c-1)\frac{K}{c}[2(\xi-1)(\xi')z^{2h-2}+(\xi-1)^2(2h-2)z^{2h-1}z'],\\
\lambda''' = {} & (2h-c-1)\frac{K}{c}[2(\xi')(\xi')z^{2h-2}+2(\xi-1)(\xi'')z^{2h-2}\\
& +2(\xi-1)(\xi')(2h-2)z^{2h-1}z'+2(\xi-1)(\xi')(2h-1)z^{2h-1}z'\\
& +(\xi-1)^2(2h-2)(2h-1)z^{2h}(z')^2+2(\xi-1)(\xi')(2h-1)z^{2h-1}z'').
\end{aligned}$$

When $\xi=1$, we have $\lambda''=0$ and λ''' is strictly negative unless $\xi'=0$. Since the only points with $\xi=1$ and $\xi'=0$ are equilibria, statement (b) follows. Statement (c) is proven similarly. ∎

If x_e is an equilibrium point, a function L defined on a neighborhood of x_e is called a Lyapunov function if it has a minimum at x_e and is strictly decreasing on all trajectories other than x_e. The existence of a Lyapunov function establishes x_e as an attractor or a stable equilibrium (see [7].) When the level sets of L are compact and

x_e is a global minimum, the point x_e is a global attractor. Since the level sets of λ are not all compact, we use topological methods to understand global behavior.

We say that an orbit $\gamma(t)$ is *positively bounded* if the positive orbit $O^+(\gamma) = \{\gamma(t) \,|\, t > 0\}$ is bounded. We say that γ is *positively unbounded* if $O^+(\gamma)$ is unbounded. Lemma 2 allows us to prove Lemma 3, which characterizes the long term behavior of positively bounded solutions.

LEMMA 3. *Suppose that* $2h-c-1 \neq 0$. *Any positively bounded solution beginning in* Ω *is asymptotic to an equilibrium point (in* Ω^**).*

Proof. The ω-limit set of a solution curve is defined to be

$$\omega(\gamma) = \cap_{s>0}\overline{\cup_{t>s}\gamma(t)},$$

where the overbar denotes closure. It is a standard result [7] that a point p is in $\omega(\gamma)$ if and only if there is a sequence $\{t_n\}$ with $t_n \to \infty$ as $n \to \infty$ such that $\gamma(t_n) \to p$. This follows directly from the definition of $\omega(\gamma)$. Another standard result [7] is that any positively bounded solution has a nonempty ω-limit set. This result follows from the Bolzano–Weierstrass theorem applied to the set $\{\gamma(n) \,|\, n \in \mathbb{N}\}$. Define the ω_Ω-limit set of γ by $\omega_\Omega(\gamma) = \omega(\gamma) \cap \Omega$.

For this proof let $\gamma(t) = (z(t), \xi(t))$ denote a positively bounded solution of the differential equation beginning from an initial condition in Ω. Since λ is strictly monotonic on trajectories and continuous on Ω, it must be constant on the ω_Ω-limit set of any solution. Hence the ω_Ω-limit set of any solution is the (possibly empty) union of equilibrium points.

We claim that if $\omega_\Omega(\gamma)$ contains an equilibrium point $p \in \Omega$, then $\omega(\gamma) = p$. For any $\delta > 0$ there is a time $s \in \mathbb{R}$ such that $\gamma(t) \in B_\delta(p)$ for all $t > s$. Were this not so, then for every $\delta > 0$ the set $\partial(B_\delta(p)) \cap \gamma$ (where $\partial(B_\delta(p)) = \{(z, \xi) \,|\, z^2 + \xi^2 = \delta\}$) would be infinite and hence have a limit point in $\partial(B_\delta(p))$, which we call p_δ. Then $p_\delta \in \omega(\gamma)$ for each δ, and the limit of p_δ as $\delta \to 0$ is p. Since our equilibrium points are isolated, infinitely many of these points are nonequilibrium points, which is impossible.

We have shown that $\omega(\gamma)$ either is a single equilibrium point in Ω or is contained in $\Omega^* - \Omega$. We claim that if $\omega(\gamma) \subseteq \Omega^* - \Omega$, then it consists of a single equilibrium point. Let $p \in \Omega^* - \Omega$ such that p is not an equilibrium point. Since $\Omega^* - \Omega$ is the positive ξ-axis, p is either in the stable manifold or in the unstable manifold of an equilibrium point in $\Omega^* - \Omega$. If p is in the stable or unstable manifold of a sink or source, then it cannot be an ω-limit point of an orbit in Ω. Suppose that p is in the unstable manifold of an equilibrium point $p_e \in \Omega^* - \Omega$ and $p \in \omega(\gamma)$ for some $\gamma(t)$. The unstable manifold of p_e in Ω consists of an orbit extending into Ω. By the Lambda lemma (see [6]), this orbit is in $\omega(\lambda)$, contradicting our earlier assertion that the only points of $\omega(\gamma)$ in Ω are equilibrium points. Similarly, p cannot be in the stable manifold of a saddle. Hence, an ω-limit set of a positively bounded orbit in Ω is a single equilibrium point. ∎

In the next two lemmas we characterize unbounded solutions.

LEMMA 4. *If* $h < 1$, *then all orbits are positively bounded.*

Proof. Suppose that $\gamma(t)$ is positively unbounded. We claim that either $\xi \to \infty$ or $z \to \infty$. Otherwise, there exist an $M > 0$ and a sequence $t_1, t_2, \ldots$ with $t_n \to \infty$ as $n \to \infty$ such that $||\gamma(t_n)|| < M$ for all n. Then by the Bolzano–Weierstrass theorem the set $\{\gamma(t_n) \,|\, n \in \mathbb{N}\}$ would have a limit point $x_0 \in B_M((0,0))$. This limit point would be in the ω-limit set of γ. By the argument in the proof of Lemma 3, $\gamma(t)$ would have to be asymptotic to x_0 and γ would be bounded.

The Poincaré sphere is a standard tool for analyzing the behavior of a two-dimensional differential equation near infinity. (See [7, p. 169].) A simple calculation using the Poincaré sphere shows that for $h < 1$ there are no orbits asymptotic to infinity. In the notation of [7], $P_2(X,Y) = -XY$, $Q_2(X,Y) = (h-1)Y^2$, and hence the only equilibrium points on the circle at infinity are $\pm(1,0,0)$ and $\pm(0,1,0)$, and none of these have a stable manifold which intersects Ω. □

LEMMA 5. *If $h > 1$, then any positively unbounded orbit has the property that $z(t) \to 0$ and $\xi(t) \to \infty$ as $t \to t_*$ for some finite t_*. Moreover, if $\xi > \max\{1, \frac{c-1}{h-1}\}$ at any time along an orbit, then the orbit is positively unbounded.*

Proof. Assume $h > 1$. Consider an orbit with initial condition (z_0, ξ_0) with $z > 0$ and $\xi > \max\{1, \frac{c-1}{h-1}\}$. We will show that any such orbit has the property that $z(t) \to 0$ and $\xi(t) \to \infty$ as $t \to t_* < \infty$. For such an orbit,

$$\begin{aligned}\xi' &= (h-1)\xi\left(\xi + \frac{1-c}{h-1}\right) + \frac{c}{K}z \\ &> (h-1)\xi\left(\xi + \frac{1-c}{h-1}\right) > 0.\end{aligned}$$

Hence, $\xi(t) > \max\{1, \frac{c-1}{h-1}\}$ and $\xi'(t) > (h-1)\xi(\xi + \frac{1-c}{h-1}) > 0$ for all $t > 0$.

The differential inequality $\xi' \geq (h-1)\xi(\xi + \frac{1-c}{h-1})$ implies that ξ behaves like a solution of $x' = x^2$; it approaches infinity in finite time. More precisely, this differential inequality can be integrated to yield

$$\left(\frac{\xi(\xi_0 + \frac{1-c}{h-1})}{\xi_0(\xi + \frac{1-c}{h-1})}\right)^{\frac{1}{1-c}} > e^t,$$

which implies that $\xi \to \infty$ before $t = (\ln(\xi_0 + \frac{1-c}{h-1}) - \ln(\xi_0))/(1-c) < \ln(\frac{h-c}{h-1})/(1-c) < \infty$.

We now show that $z(t) \to 0$ for this orbit. Since $z' = z(1-\xi)$ and $\xi \to \infty$, it follows that $z(t)$ is eventually monotonic decreasing. Since $z(t)$ is bounded below by 0, $z(t)$ converges to some value $z_* \geq 0$. Since $\xi' = (h-1)\xi^2 + (1-c)\xi + \frac{c}{K}z$ and z is bounded, $\xi(t)$ is eventually monotonic increasing. Without loss of generality, assume that the initial condition for this orbit is (z_0, ξ_0) and that z is monotonic decreasing and ξ is monotonic increasing for all $t \geq 0$. The positive orbit $(z(t), \xi(t))$, $t \geq 0$, then corresponds to the graph of a function $z = f(\xi)$, $f : (\xi_0, \infty) \to (z_*, z_0)$. Then

$$\frac{\mathrm{d}f}{\mathrm{d}\xi} = \frac{\mathrm{d}z/\mathrm{d}t}{\mathrm{d}\xi/\mathrm{d}t} < 0,$$

and by the fundamental theorem of calculus

$$\begin{aligned}z_0 - z_* &= \int_{\xi_0}^{\infty} f'(\xi)\mathrm{d}\xi \\ &= \int_{\xi_0}^{\infty} \frac{f(\xi)(1-\xi)}{(h-1)\xi^2 + (1-c)\xi + \frac{c}{K}f(\xi)}\mathrm{d}\xi.\end{aligned} \tag{7}$$

Note that $f'(\xi) < 0$ because $\xi' > 0$ and $z' < 0$. Since $z_* \leq f(\xi) \leq z_0$ for all ξ,

$$\frac{f(\xi)(1-\xi)}{(h-1)\xi^2 + (1-c)\xi + \frac{c}{K}f(\xi)} < \frac{z_*(1-\xi)}{(h-1)\xi^2 + (1-c)\xi + \frac{c}{K}z_0}.$$

Observe that

$$\lim_{\xi\to\infty} \frac{\frac{z_*(1-\xi)}{(h-1)\xi^2+(1-c)\xi+\frac{c}{K}z_*}}{\frac{-1}{\xi}} \geq \frac{z_*}{(h-c)} \geq 0,$$

with equality if and only if $z_* = 0$. If $z_* \neq 0$, the limit comparison test for indefinite integrals implies that the integral in (7) diverges because the integral

$$\int_{\xi_0}^{\infty} \frac{-1}{\xi} \mathrm{d}\xi$$

diverges. This is a contradiction since $z_0 - z_*$ is finite. Hence $z(t) \to z_* = 0$.

Now consider an orbit with initial condition (z_0, ξ_0) with $z_0 > 0$ and $\xi_0 \leq \max\{1, \frac{c-1}{h-1}\}$. We will show that if $\xi(t)$ is bounded, then so is $z(t)$. This will complete the proof of the $h > 1$ portion of the lemma. Suppose, to obtain a contradiction, that $z(t)$ is unbounded for some orbit $(z(t), \xi(t))$. Since the set $\xi' = 1$ is a parabola opening to the left, there is some M such that if $z > M$, then $\xi' > 1$. Since $z(t)$ is unbounded and $z'(t) = z(\xi - 1) < z$, there are times $t_a < t_b$ with $z(t) > M$ for all $t \in [t_a, t_b]$ and $t_b - t_a > \max\{1, \frac{c-1}{h-1}\}$. Since Ω is positive invariant, $\xi(t_a) > 0$. Hence,

$$\xi(t_b) > \xi(t_b) - \xi(t_a) = \int_{t_a}^{t_b} \xi' \mathrm{d}t > \int_{t_a}^{t_b} 1 \mathrm{d}t > t_b - t_a > \max\left\{1, \frac{c-1}{h-1}\right\}.$$

Therefore, by the previous assertion, $\xi(t) \to \infty$ and $z(t) \to 0$. This proves that $z(t)$ is bounded for positive time for all orbits. □

In the course of this proof, we have established an upper bound on the time a population takes to disappear if the orbit begins in Ω with $\xi > \max\{1, \frac{c-1}{h-1}\}$. Specifically, $\xi \to \infty$ before $t = (\ln(\xi_0 + \frac{1-c}{h-1}) - \ln(\xi_0))/(1-c) < \ln(\frac{h-c}{h-1})/(1-c)$. This implies that $P(t), R(t) \to 0$ while $t < \ln(\frac{h-c}{h-1})/(1-c) < \infty$.

PROPOSITION 1. *Suppose that $h < c$ and $h < 1$. Then all solutions beginning in Ω are asymptotic to the stable equilibrium at $(z, \xi) = (\Gamma, 1)$, where $\Gamma = K(1 - h/c)$. Hence, in P, R-coordinates, all solutions beginning in the first quadrant are asymptotic to the equilibrium at $(P, R) = (\Gamma, \Gamma)$.*

Proof. Since $h < 1$, by Lemma 4 every solution is bounded and approaches an equilibrium solution. For these parameter values the equilibrium solutions in Ω^* are $(0, 0)$, $(K(1 - h/c), 1)$, and possibly $(0, \frac{c-1}{h-1})$. The only equilibrium with a nonempty stable manifold in Ω is $(K(1 - h/c), 1)$. Hence, all solutions are asymptotic to this equilibrium. □

PROPOSITION 2. *If $h > c$ and $h < 1$, then $(z, \xi) \to (0, \frac{c-1}{h-1})$ as $t \to \infty$ for all solutions. Hence, $(P, R) \to (0, 0)$ as $t \to \infty$ with*

$$\frac{P}{R} \sim \frac{c-1}{h-1}.$$

Proof. Since $h < 1$, by Lemma 4 all solutions are bounded. The only equilibrium point in Ω^* with a nonempty stable manifold in Ω is $(0, \frac{c-1}{h-1})$. By Lemma 3, all solutions are asymptotic to $(0, \frac{c-1}{h-1})$. □

An alternative proof of Proposition 2 is based on the Lyapunov function

$$L = \frac{1}{2}\left(\xi + \frac{1-c}{h-1}\right)^2 + \frac{c}{K}z$$

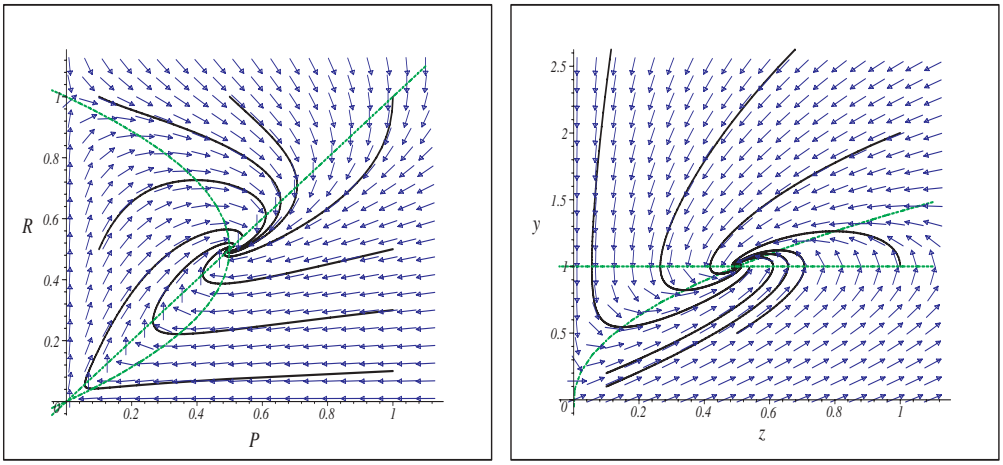


FIG. 6. *The phase planes with* $c = 1$, $h = 0.5$, *and* $K = 1$ *satisfying the hypothesis of Proposition* 1. *The nullclines are indicated by dashed lines.*

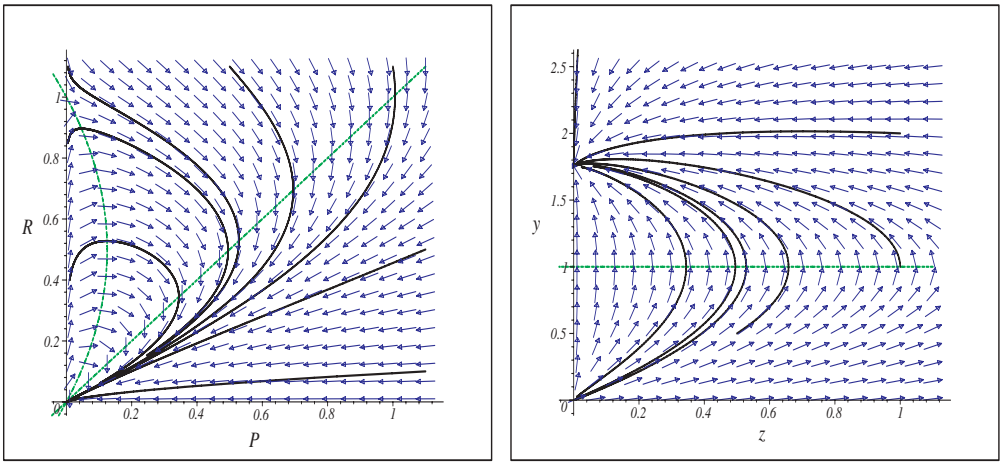


FIG. 7. *The phase planes with* $c = 0.3$, $h = 0.6$, *and* $K = 1$ *satisfying the hypothesis of Proposition* 2. *The same set of initial conditions is used for the solutions shown in each coordinate system. The nullclines are indicated by dashed lines.*

with

$$\frac{\mathrm{d}L}{dt} = (h-1)\xi\left(\xi + \frac{1-c}{h-1}\right)^2 + \frac{c}{K}\left(\frac{h-c}{h-1}\right)z,$$

which is negative when $c < h < 1$. Moreover, the level sets of L are compact. This constitutes an alternate proof that $(0, \frac{c-1}{h-1})$ is a global attractor.

PROPOSITION 3. *If* $h > c$ *and* $h > 1$, *then* $z \to 0$ *and* $\xi \to \infty$ *in finite time for all solutions. Hence,* (P, R) *goes to the singularity at* $(0, 0)$ *in finite time with*

$$\frac{P}{R} \to \infty.$$

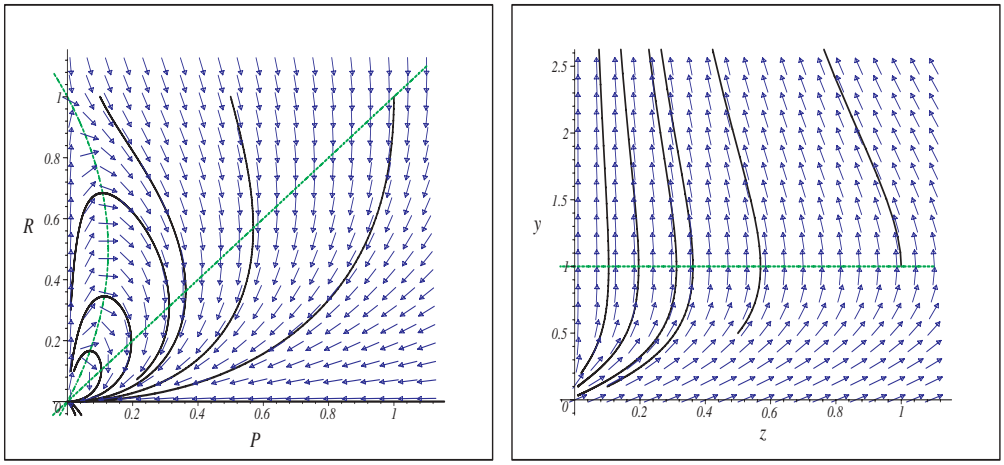


FIG. 8. *The phase planes with $c = 1$, $h = 2$, and $K = 1$ satisfying the hypothesis of Proposition 3. The nullclines are indicated by dashed lines.*

Proof. For our parameter values the only equilibrium solutions in Ω^* are $(0,0)$ and possibly $(0, \frac{c-1}{h-1})$. The equilibrium $(0,0)$ is unstable and $(0, \frac{c-1}{h-1})$ is unstable if $(0, \frac{c-1}{h-1}) \in \Omega^*$. Therefore, there are no solutions in Ω that are asymptotic to an equilibrium solution. By Lemma 3, all solutions are unbounded. By Lemma 5, $z(t) \to 0$ and $\xi(t) \to \infty$ in finite time for all solutions. □

PROPOSITION 4. *Suppose that $h < c$, $h > 1$, and $2h - c - 1 > 0$. For almost every solution, $z(t) \to 0$ and $\xi(t) \to \infty$ in finite time. There is one solution with $\xi(t) \to \frac{c-1}{h-1}$ and one unstable equilibrium solution at $(K(1 - h/c), 1)$.*

In P, R-coordinates, almost every solution goes to the singularity at $(0,0)$ in finite time with

$$\frac{P}{R} \to \infty.$$

There is one solution that is asymptotic to $(0,0)$ with $P/R \sim \frac{c-1}{h-1}$ and one unstable equilibrium solution at $(K(1 - h/c), K(1 - h/c))$.

Proof. For our parameter values the equilibrium solutions in Ω^* are $(0,0)$, $(0, \frac{c-1}{h-1})$, and $(K(1 - h/c), 1)$. The equilibrium at $(0,0)$ is a saddle with its stable manifold contained in $\Omega^* - \Omega$. The equilibrium at $(0, \frac{c-1}{h-1})$ is a saddle, and its stable manifold is a solution extending into Ω. The equilibrium at $(K(1 - h/c), 1)$ is unstable. Therefore, all solutions except $(K(1 - h/c), 1)$ and the stable manifold $(0, \frac{c-1}{h-1})$ are unbounded by Lemma 3. By Lemma 5, $z(t) \to 0$ and $\xi(t) \to \infty$ in finite time for these solutions. □

PROPOSITION 5. *Suppose that $h < c$, $h > 1$, and $2h - c - 1 < 0$. Let A denote the region*

$$0 < z < \frac{-K(2h-1)}{2c}\left(\xi^2 - 2\xi + \frac{2h-c-1}{h-1}\right).$$

All orbits that intersect A approach the equilibrium point $(\Gamma, 1)$ asymptotically; in fact, these solutions constitute the basin of attraction of this sink. The single solution stable manifold of $(0, \frac{c-1}{h-1})$ is a separatrix. All other solutions have the property that $z \to 0$, $\xi \to \infty$ in finite time.

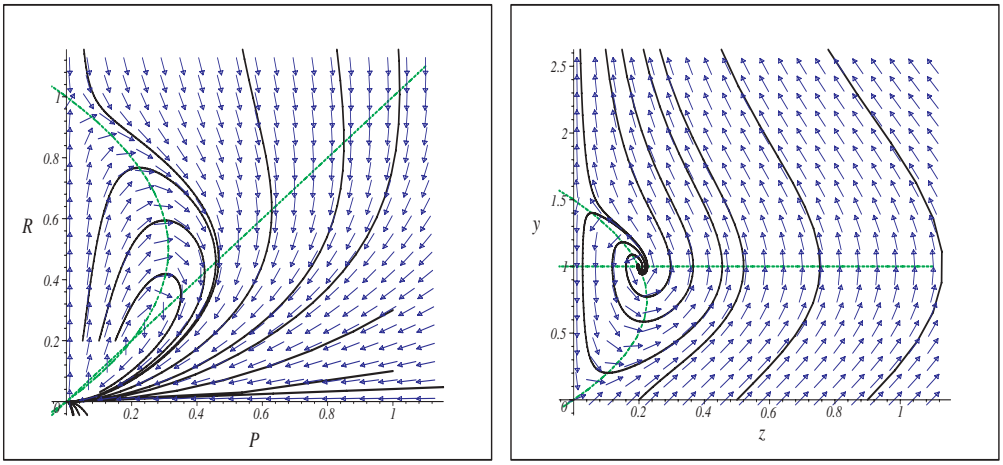


FIG. 9. *The phase planes with $c = 2.5$, $h = 2$, and $K = 1$ satisfying the hypothesis of Proposition 4. The nullclines are indicated by dashed lines.*

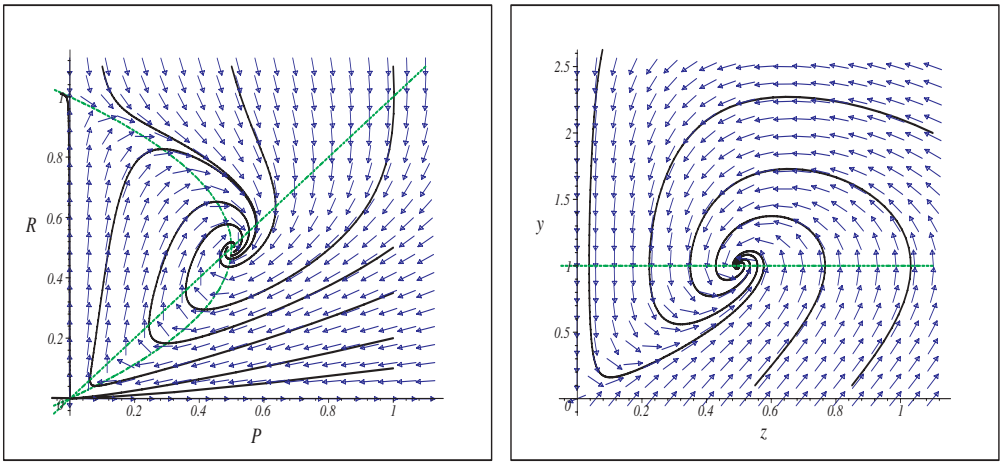


FIG. 10. *The phase planes with $c = 3$, $h = 1.5$, and $K = 1$ satisfying the hypothesis of Proposition 5. The nullclines are indicated by dashed lines.*

In P, R-coordinates, there is an open set of orbits which are asymptotic to the equilibrium at $(P, R) = (\Gamma, \Gamma)$. There is an open set of orbits which approach $(0, 0)$ in finite time. There is a single solution that is asymptotic to $(0, 0)$ with $P/R \sim \frac{c-1}{h-1}$ as $t \to \infty$, and this solution is the separatrix between the two open sets.

Proof. By solving $\lambda = 0$ we obtain

$$z = \frac{-K(2h-1)}{2c}\left(\xi^2 - 2\xi + \frac{2h-c-1}{h-1}\right).$$

It is easy to see that $\lambda < 0$ on A and $(\Gamma, 1) \in A$. Observe that the only equilibrium point in $\overline{A}$ whose stable manifold has a nonempty intersection with A is $(\Gamma, 1)$. Since $\lambda'(t) < 0$ for all orbits in Ω by Lemma 2, all orbits are forward asymptotic to $(\Gamma, 1)$.

The stable manifold of $(0, \frac{c-1}{h-1})$ in Ω is a single orbit. Denote this orbit by $\alpha(t)$.

Near $(0, \frac{c-1}{h-1})$, there is a well-defined notion of above and below α. Orbits just above α follow the unstable manifold of $(0, \frac{c-1}{h-1})$ up the ξ-axis. Eventually $\xi(t) > \frac{c-1}{h-1}$ for these solutions. Since $2h-c-1 < 0$, we have $c-1 > 2(h-1)$ and $\frac{c-1}{h-1} = \max\{1, \frac{c-1}{h-1}\}$. Therefore, $\xi(t) \max\{1, \frac{c-1}{h-1}\}$ for each of these solutions and $z(t) \to 0$, $\xi(t) \to \infty$ by Lemma 5.

Orbits just below α follow the unstable manifold of $(0, \frac{c-1}{h-1})$ down the ξ-axis into the region A, and these orbits are asymptotic to $(\Gamma, 1)$. Therefore, α is the separatrix between orbits asymptotic to $(\Gamma, 1)$ and ones that approach $(0, \infty)$ in finite time. By Lemmas 3 and 5, every orbit is asymptotic to $(\Gamma, 1)$, $(0, \frac{c-1}{h-1})$, or $(0, \infty)$. □

PROPOSITION 6. *Suppose that $h < c$, $h > 1$, and $2h - c - 1 = 0$. Let A denote the region*

$$0 < z < \frac{-K(2h-1)}{2c}\left(\xi^2 - 2\xi\right).$$

There is a heteroclinic orbit from $(0,0)$ to $(0, \frac{c-1}{h-1})$. The boundary of A consists of this orbit, a heteroclinic orbit in $\Omega^* - \Omega$ from $(0,0)$ to $(0, \frac{c-1}{h-1})$, and these two equilibrium points. All orbits inside A are periodic. All orbits outside of A approach $(0, \infty)$ in finite time.

In P, R-coordinates, there is a heteroclinic orbit from $(K, 0)$ to $(0, 0)$. This orbit together with the orbit in the R-axis from $(0,0)$ to $(K, 0)$ and these two equilibrium points forms a limit cycle. Orbits within this limit cycle are all periodic. Orbits outside of this limit cycle approach $(0, 0)$ in finite time.

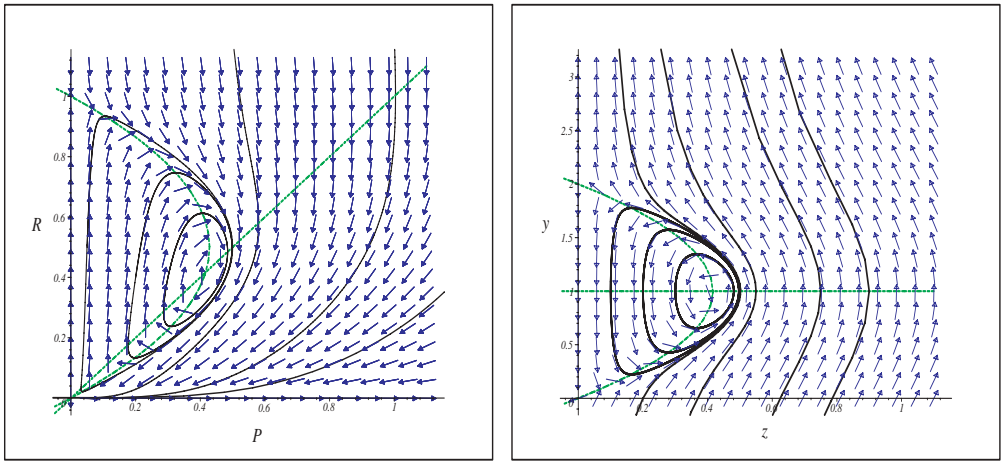


FIG. 11. *The phase planes with $c = 6$, $h = 3.5$, and $K = 1$ satisfying the hypothesis of Proposition 6. The nullclines are indicated by dashed lines.*

Proof. For this case, $\lambda'(t) = 0$ by Lemma 2, so λ is an integral of the system. We want to show that it is a nondegenerate integral. The gradient of λ is

$$\nabla\lambda = \left((2h-2)z^{2h-3}\left(\frac{K}{2c}\xi^2 - \frac{K}{c}\xi + \frac{z}{2h-1} + \frac{K}{2c} - \frac{K(1-h/c)}{2h-2}\right) + \frac{z^{2h-2}}{2h-1}, z^{2h-2}\frac{K}{c}(\xi - 1)\right).$$

Since $\partial\lambda/\partial\xi = z^{2h-2}\frac{K}{c}(\xi - 1)$, if $\nabla\lambda = 0$ in Ω, then $\xi = 1$. Substituting $\xi = 1$ into $\partial\lambda/\partial z = 0$ gives $z = \Gamma$. Hence, the only point in Ω where $\nabla\lambda = \mathbf{0}$ is the equilibrium $(\Gamma, 1)$, and λ is nondegenerate.

From (3) with $2h - c - 1 = 0$, solving $\lambda = 0$, we get

$$z = \frac{-K(2h-1)}{2c}\left(\xi^2 - 2\xi\right).$$

This is a parabola opening to the left. It intersects the ξ-axis at $\xi = 0, \frac{c-1}{h-1}$. Since λ is constant along solutions, this parabola is a heteroclinic orbit.

By direct computation, $\lambda(1, \Gamma) = \frac{-\Gamma^{2h-1}}{(2h-2)(2h-1)}$. The orbits with $\frac{-\Gamma^{2h-1}}{(2h-2)(2h-1)} < \lambda < 0$ are nested periodic orbits in A that enclose convex regions containing $(1, \Gamma)$. This follows from differentiation of λ.

There are no equilibria outside $\overline{A}$. Hence, by the Poincaré–Bendixson theorem, all orbits outside of A are unbounded. By Lemma 5, all of these orbits approach $(0, \infty)$ in finite time. □

In conclusion we note that our system undergoes a degenerate Hopf bifurcation when $h < c$, $h > 1$, and $2h - c - 1$ changes sign. The equilibrium at $(K(1 - h/c), 1)$ changes from a spiral source for $2h - c - 1 > 0$ to a spiral sink for $2h - c - 1 > 0$. When such a transition occurs through a classic Hopf bifurcation, a single periodic orbit emerges from (or contracts to) the equilibrium point. The Hopf bifurcation theorem (see [7]) identifies a large class of conditions under which such bifurcations occur. Our system falls through the cracks of the theorem; the theorem applies to all systems except those whose Taylor coefficients at the equilibrium satisfy a certain equation, and the Taylor coefficients of our system satisfy that equation.

REFERENCES

[1] J. M. Anderies, *On modeling human behavior and institutions in simple ecological economic systems*, Ecological Economics, 35 (2000), pp. 393–412.

[2] J. M. Anderies, *Econimic development, demographics, and renewable resources: A dynamical systems approach*, Environment and Development Economics, 8 (2003), pp. 219–246.

[3] J. A. Brander and M. S. Taylor, *The simple economics of Easter Island: A Ricardo-Malthus model of renewable resource use*, American Economic Review, 88 (1998), pp. 119–138.

[4] J. Cohen, *How Many People Can the Earth Support?*, W. W. Norton and Company, New York, 1995.

[5] L. L. Eberhardt, *Is wolf predation ratio-dependent?*, Canadian J. Zoology, 76 (1997), pp. 380–386.

[6] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press, Cambridge, UK, 1995.

[7] L. Perko, *Differential Equations and Dynamical Systems*, Springer-Verlag, New York, 1993.

[8] C. L. Redman, *Human Impact on Ancient Environments*, University of Arizona Press, Tucson, AZ, 1999.

[9] P. Turchin, *Complex Population Dynamics: A Theoretical/Empirical Synthesis*, Princeton University Press, Princeton, NJ, 2003.

SIAM J. APPL. MATH.
Vol. 65, No. 2, pp. 702–719

PROJECTIVE GEOMETRY OF HUMAN MOTION, WITH AN APPLICATION TO INJURY RISK*

H. LAURIE† AND R. PENNE‡

Abstract. We give an exposition of Plücker vectors for a system of joint axes in projective 3-space. We use Plücker vectors to analyze dependencies among joint axes and in particular to show that two rotational joints rigidly joined by a bar and each with 3 degrees of freedom always form a 5-dimensional system. We introduce the concept of reduced redundancy in a dependent set of projective Lines and argue that reduced redundancy in the axes of a body position increases injury risk. We apply this to a simple two-joint model of bowling in cricket and show by analysis of some experimental data that reduced redundancy around ball release is observed in some cases.

Key words. Plücker coordinates, human motion, reduced redundancy, injury risk

AMS subject classifications. 51N15, 51505, 70E17, 70E60, 92C10, 92C50

DOI. 10.1137/S0036139903429658

1. Introduction. A variety of techniques exist for the mathematical analysis of human motion, including techniques that are also used in robotics [15, 2]. However, to our knowledge, nobody has yet employed the formalisms and insights of projective geometry, well known in robotics [18, 3].

We are motivated by the analysis of certain complicated athletic effects achieved by throwlike motions, such as a topspin serve in tennis and an away-swinger in cricket. It is clear that the brief interval ending in the release of the ball is crucial: after release, the ball is in free fall, except for some aerodynamic and gyroscopic effects. Thus the athlete must release the ball in a particular state of motion (translational as well as rotational).

Several questions arise: By what movements of the joints does a given athlete achieve a given effect? Is there more than one way to achieve a given effect? Do some effects require motions that are inherently more risky than others, and if so, can this risk be characterized analytically?

Many of these questions can be illuminated by using techniques from the mathematics of robotics, in which the following are possible: a simple description which unifies all aspects of the motion of the athlete and the ball, a representation in which rotation and translation are easily combined, and a level of generality at which all cases of reduced mobility can be found (and explicitly calculated).

In this report, we analyze what appears to us be a simple, interesting case: the motion of a cricketer's arm (much simplified) near the moment of delivery. We regard the hips as fixed, the torso as rigid, and the waist and shoulder as joints, each of which provides 3 degrees of freedom. Alert readers will notice that we ignore the elbow, wrist, and fingers, as well as any contact motion of the ball in the hand prior to release. For our purposes, it is assumed that the requirements of a given delivery have prescribed the motion at the center of the wrist. However, we will see that in our

*Received by the editors June 11, 2003; accepted for publication (in revised form) June 11, 2004; published electronically January 27, 2005.
http://www.siam.org/journals/siap/65-2/42965.html
†Department of Mathematics and Applied Mathematics, UCT, Rondebosch, 7701, South Africa (henri@pc001a.mth.uct.ac.za).
‡Industrial Sciences and Technology, Karel de Grote-Hogeschool, Antwerpen, Belgium (rudi.penne@kdg.be).

model in all positions there are only 5 degrees of freedom for the motion of the ball. Consequently, our system of 6 axes possesses intrinsic redundancy by design! It can be regarded as the solution of nature to the human desire to accomplish complicated motions. Indeed, kinematic redundancy offers an opportunity to distribute stress over many joints (see also [15]). Our main result is to prove the existence of special positions where *reduced redundancy* occurs: that is, for a given motion the amount of rotation about one or more joint axes is fixed, while there is some freedom in distributing motion about the remaining joint axes. These positions should not be confused with standard kinematic singularities, as no decrease of mobility is involved; 5 degrees of freedom are always maintained. We interpret reduced redundancy as a source of injury risk.

The paper is organized as follows: First, we describe the use of Plücker coordinates as a unified framework for computations concerning rotations and translations in many linked joints. Second, we describe a simplified model for the motion of an arm in the act of releasing a legal cricket delivery, a motion associated with risk of overuse injury [12, 14, 11]. Third, we analyze the degrees of freedom for the motion of the ball in this model, for the general case as well as all special cases; this includes a full analysis of reduced redundancy. Fourth, we discuss the analysis from two points of view: the prevention of injury and the forbidden motions of the wrist. Finally, we show that reduced redundancies indeed occur in real bowling actions by analyzing data from two bowlers with an injury history.

2. Plücker coordinates for human motion. Human motion is the result of rotations around joint axes, at least, infinitesimally in the first approximation (that is, neglecting the play in the joints and the deformation of bone, cartilage, and ligament). However, the desired motion of the end effector (in our case, the cricket ball) will in general have components of both rotation and translation. In projective geometry, translation can be rendered as rotation about an axis at infinity. In this view, all motions are rotations, and Plücker coordinates are merely a convenient way of describing them.

2.1. Projective points. In projective geometry, we identify all points on a line through the origin in $\mathbb{R}^4$ with a projective Point in the corresponding projective space $\mathbb{P}^3$. Thus the vector $\mathbf{x} = (\lambda a, \lambda b, \lambda c, \lambda d) \in \mathbb{R}^4$ corresponds to $\mathbf{p} \in \mathbb{P}^3$ for all $\lambda \neq 0$. Such a 4-vector is referred to as a set of homogeneous coordinates for $\mathbf{p}$. By the usual convention, the hyperplane $H{:}x_4 = 1$ in $\mathbb{R}^4$ is considered as (a copy of) affine 3-space. All Points of $\mathbb{P}^3$ which correspond to lines intersecting this hyperplane are called *finite points*, and these are identified with the affine point of H where they intersect. So for finite points,

$$(a, b, c, d) \sim (a/d, b/d, c/d, 1) \sim (a/d, b/d, c/d).$$

Notice that some lines through the origin in $\mathbb{R}^4$ do not intersect the hyperplane H, and therefore some projective Points are not finite. They are said to lie *at infinity*, and they are represented by homogeneous coordinates with 0 as the fourth coordinate: $(a, b, c, 0)$.

Similarly, planes through the origin of $\mathbb{R}^4$ correspond to Lines in $\mathbb{P}^3$. If such a plane is parallel to the hyperplane H, then it represents a Line at infinity. Finally, each 3-dimensional subspace of $\mathbb{R}^4$ is associated with a Plane in $\mathbb{P}^3$. The Plane corresponding to $x_4 = 0$ is the Plane at infinity of $\mathbb{P}^3$, and it contains all Points at infinity.

2.2. Plücker coordinates. From an algebraic point of view, a chosen set of homogeneous coordinates for a Point $\mathbf{p} \in \mathbb{P}^3$ represents a vector in the vector space $V = \mathbb{R}^4$. Now we can consider the *exterior algebra* built on V:

$$\wedge V = V^{(0)} \oplus V^{(1)} \oplus V^{(2)} \oplus V^{(3)} \oplus V^{(4)},$$

which enables us to make computations with scalars ($\mathbb{R} = V^{(0)}$), vectors ($V = V^{(1)}$), but also with more complicated objects called *antisymmetric tensors*, and this in the same framework. The *exterior product* $\wedge$ is a bilinear, antisymmetric operation on $\wedge V$, such that for $\mathbf{A} \in V^{(i)}$ and $\mathbf{B} \in V^{(j)}$ we get $\mathbf{A} \wedge \mathbf{B} \in V^{(i+j)}$ if $i + j \leq 4$ or $\mathbf{A} \wedge \mathbf{B} = \mathbf{0}$ otherwise (also in the case $i + j \leq 4$ it can happen that $\mathbf{A} \wedge \mathbf{B} = \mathbf{0}$ in $V^{(i+j)}$).

Example 1. The elements in $V^{(2)}$ (the so-called 2-tensors) are products $\mathbf{p} \wedge \mathbf{q}$ of vectors $\mathbf{p}$ and $\mathbf{q}$ in V, or linear combinations of these. Notice that $\mathbf{p} \wedge \mathbf{q} = \mathbf{0}$ in $V^{(2)}$ if $\mathbf{p}$ and $\mathbf{q}$ represent the same projective point, due to the antisymmetry.

For the reader who is not familiar with the exterior algebra it suffices to know for our purposes that each tensor can be regarded as just some vector, and $V^{(i)}$ as a real vector space of dimension $\binom{4}{i}$. For example, $V^{(2)} \cong \mathbb{R}^6$. Furthermore, using the standard basis of V, there is a canonical way to construct a basis for $V^{(i)}$. The corresponding coordinates arising in this manner for tensors are called *Plücker coordinates*.

Let us be more specific in the case of 2-tensors, because they will be needed most in this article. If $\mathbf{L} \in V^{(2)}$ then we have 6 Plücker coordinates for $\mathbf{L}$, by convention labeled by double-indices:

$$\mathbf{L} = (L_{12}, L_{13}, L_{14}, L_{23}, L_{24}, L_{34}).$$

In the special case that $\mathbf{L} = \mathbf{p} \wedge \mathbf{q}$ with $\mathbf{p} = (p_1, p_2, p_3, p_4)$ and $\mathbf{q} = (q_1, q_2, q_3, q_4)$, there is an easy rule to obtain the coordinates for $\mathbf{L}$: $L_{ij} = p_i q_j - p_j q_i$.

$$\mathbf{p} \wedge \mathbf{q} = \begin{pmatrix} p_1 q_2 - p_2 q_1 \\ p_1 q_3 - p_3 q_1 \\ p_1 q_4 - p_4 q_1 \\ p_2 q_3 - p_3 q_2 \\ p_2 q_4 - p_4 q_2 \\ p_3 q_4 - p_4 q_3 \end{pmatrix},$$

the elements of which are the 2×2 minors of the matrix $\left(\begin{smallmatrix} p_1 & p_2 & p_3 & p_4 \\ q_1 & q_2 & q_3 & q_4 \end{smallmatrix}\right)$ in lexicographic order. If $\mathbf{p}$ and $\mathbf{q}$ represent different projective Points, then $\mathbf{L} = \mathbf{p} \wedge \mathbf{q}$ represents the projective Line through these two Points. Of course, many other 2-tensors represent the same Line in $\mathbb{P}^3$. Indeed, we can use a multiple of $\mathbf{p}$ or $\mathbf{q}$ without changing the involved projective Points, or we can even choose another pair of Points on the same Line. Fortunately, the new 2-tensor $\mathbf{L}' = \mathbf{p}' \wedge \mathbf{q}'$ will always be a multiple of $\mathbf{L}$: $\mathbf{L}' = \alpha \mathbf{L}$. We conclude that the Plücker coordinates of $\mathbf{L}$ can be considered as a 6-tuple of homogeneous coordinates for the projective Line represented by $\mathbf{L}$. Notice that Lines at infinity are characterized by having Plücker coordinates with $L_{14} = L_{24} = L_{34} = 0$.

Because not every 2-tensor in $V^{(2)}$ can be written as the exterior product of two vectors in V, not every 6-tuple of Plücker coordinates represents a Line in $\mathbb{P}^3$. More precisely, one can prove that $(L_{12}, L_{13}, L_{14}, L_{23}, L_{24}, L_{34})$ corresponds to a projective line if and only if it differs from zero and the *Grassmann–Plücker relation* is satisfied:

$$L_{14} L_{23} - L_{24} L_{13} + L_{34} L_{12} = 0. \tag{GP}$$

Thus, "most" 6-tuples in $\mathbb{R}^6$ are not the Plücker coordinates of a projective Line. However, there is an interesting theorem, *Poinsot's central axis theorem*, which says that each 2-tensor $\mathbf{A}$ not obeying (GP) can be expressed as a sum $\mathbf{A} = \mathbf{L} + \mathbf{M}$ such that

1. $\mathbf{L}$ corresponds to a finite Line (not at infinity),
2. $\mathbf{M}$ corresponds to a Line at infinity,
3. Every affine Plane through $\mathbf{M}$ is perpendicular to $\mathbf{L}$.

Let us give one more illustration of Plücker coordinates. Given is a 2-tensor $\mathbf{A} = (A_{12}, A_{13}, A_{14}, A_{23}, A_{24}, A_{34})$ and a vector $\mathbf{p} =(p_1, p_2, p_3, p_4)$. Then $\mathbf{P} = \mathbf{A} \wedge \mathbf{p}$ belongs to the 4-dimensional space $V^{(3)}$: $\mathbf{P} = (P_{123}, P_{124}, P_{134}, P_{234})$, with

$$\begin{aligned} P_{123} &= A_{12}p_3 - A_{13}p_2 + A_{23}p_1, \\ P_{124} &= A_{12}p_4 - A_{14}p_2 + A_{24}p_1, \\ P_{134} &= A_{13}p_4 - A_{14}p_3 + A_{34}p_1, \\ P_{234} &= A_{23}p_4 - A_{24}p_3 + A_{34}p_2. \end{aligned}$$

In particular, if $\mathbf{A}$ represents a projective Line, which moreover does not contain the projective Point represented by $\mathbf{p}$, then $\mathbf{P}$ represents the projective Plane determined by this Line and this Point. If the Point lies on the Line, then $\mathbf{P} = \mathbf{0}$. In any case, if a 3-tensor differs from zero, it will represent a Plane in $\mathbb{P}^3$. Furthermore, it is the Plane at infinity if and only if $P_{124} = P_{134} = P_{234} = 0$. On the other hand, if $\mathbf{P} \in V^{(3)}$ represents a finite Plane, the vector $(P_{234}, -P_{134}, P_{124}) \in \mathbb{R}^3$ is perpendicular to the associated affine plane.

For a good introduction to Plücker coordinates and antisymmetric tensors, including the formal definitions, we refer to [18].

2.3. Dependencies among lines. A set of Lines in $\mathbb{P}^3$, finite or at infinity, is called independent (resp., dependent) if the corresponding 2-tensors are linearly independent (resp., dependent) in $V^{(2)}$ or, equivalently, if the corresponding Plücker coordinates are linearly independent (resp., dependent) in $\mathbb{R}^6$. These concepts are defined in algebraic terms; nevertheless, the possible dependencies among projective Lines have a transparent geometric characterization. We refer to [5] for a complete description of this. We quote only those situations that will be relevant for our analysis.

- Two Lines can only be dependent when they coincide.
- Three Lines are dependent if and only if they lie in the same Plane *and* go through the same Point.
- Four Lines are dependent if and only if at least one of the following cases occur:
  1. Three of the four Lines are dependent.
  2. The four Lines lie in the same Plane.
  3. The four Lines go through the same Point.
  4. Two of the Lines lie in a Plane α, intersecting in Point p, and the remaining two Lines lie in a Plane β, intersecting in Point q, such that the Planes α and β meet in the line pq.
  5. The four Lines belong to the same system of rulers on a quadratic surface.

In particular, if we are given two parallel Lines (intersecting at infinity), then a linear combination of their Plücker coordinates will always represent a Line in the

unique Plane through the given Lines, and which either lies at infinity, or which is parallel to the given Lines. Four parallel Lines in 3-space are always dependent.

Example 2. As an illustration, let us consider a situation of 4 Lines, with $\mathbf{L}_1, \mathbf{L}_2, \mathbf{L}_3$ concurrent (but not coplanar) through Point $\mathbf{p}$, and $\mathbf{L}_4$ not containing $\mathbf{p}$. Clearly, these 4 Lines are independent. By taking linear combinations of the 3 concurrent Lines we can generate any Line $\mathbf{L}$ through $\mathbf{p}$. Furthermore, if $\mathbf{L}$ happens to intersect $\mathbf{L}_4$ (in $\mathbf{q}$, say), then linear combinations of $\mathbf{L}$ and $\mathbf{L}_4$ generate Lines through $\mathbf{q}$, lying in the Plane determined by $\mathbf{L}$ and $\mathbf{L}_4$. If $\mathbf{L}$ and $\mathbf{L}_4$ do not intersect each other, then we cannot obtain new Lines by combining them (a violation of (GP)). We conclude that the Lines which depend on $\mathbf{L}_1, \mathbf{L}_2, \mathbf{L}_3, \mathbf{L}_4$, are exactly those that contain $\mathbf{p}$ or that lie in the Plane through $\mathbf{L}_4$ and $\mathbf{p}$.

Next, we will elaborate on a special case which will be important for the applications in this paper.

THEOREM 2.1. *Let* $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3$ *and* $\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3$ *be two triples of concurrent Lines in* $\mathbb{P}^3$ *through different Points* $\mathbf{w}$ *and* $\mathbf{s}$. *Assume moreover that* $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3$ *are not coplanar, neither are* $\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3$. *Then the set* $\{\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3, \mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3\}$ *always has rank* 5. *Or more explicitly, these* 6 *Lines are always dependent but always contain a subset of* 5 *Lines which is independent.*

Proof. Choose a Plücker vector $\mathbf{P}$ to represent the Line $\mathbf{sw}$. By abuse of notation, we let $\mathbf{W}_1, \ldots, \mathbf{S}_3$ stand for the Plücker vectors of the corresponding lines as well. Then there exist linear combinations

$$\mathbf{P} = \alpha_1\mathbf{W}_1 + \alpha_2\mathbf{W}_2 + \alpha_3\mathbf{W}_3,$$
$$\mathbf{P} = \beta_1\mathbf{S}_1 + \beta_2\mathbf{S}_2 + \beta_3\mathbf{S}_3,$$

which gives rise to the claimed dependency:

$$\alpha_1\mathbf{W}_1 + \alpha_2\mathbf{W}_2 + \alpha_3\mathbf{W}_3 - \beta_1\mathbf{S}_1 - \beta_2\mathbf{S}_2 - \beta_3\mathbf{S}_3 = \mathbf{0}.$$

Next, we observe that at least one of $\{\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3\}$ does not pass through $\mathbf{w}$, say $\mathbf{S}_1$. From the example above we learn that $\{\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3, \mathbf{S}_1\}$ is a set of independent Lines, and moreover, the only Lines which are dependent on these 4 Lines are Lines through $\mathbf{w}$, or Lines in the Plane determined by $\mathbf{S}_1$ and $\mathbf{w}$. Because the triple $\{\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3\}$ is assumed to be nonplanar, it is impossible that both $\mathbf{S}_2$ and $\mathbf{S}_3$ depend on $\{\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3, \mathbf{S}_1\}$, which completes the proof. ∎

2.4. Describing kinematics by Plücker coordinates. For a more extended exposition of the material presented in this paragraph, we refer to [18] and [4]. Consider a motion of a rigid body B in 3-space. Then, every point p of B traces a path, $p = p(t)$. If the motion is sufficiently *smooth* from a mathematical point of view, we can compute the derivative at a certain time t_0, giving us the infinitesimal motion of B at $t = t_0$. This results in a velocity vector $v_p = \dot{p}(t_0)$ for every point p of B. The rigidity of B can be translated into the statement that for every pair of its points $\{p, q\}$ the distance between these points must remain constant during the motion,

$$||p(t) - q(t)||^2 = \text{constant},$$

or infinitesimally (*preserved distance property*),

$$(v_p - v_q) \cdot (p - q) = 0. \tag{PDP}$$

From now on, when we use the term "motion," we always mean an infinitesimal rigid motion: the assignment of a velocity vector to every point of B such that (PDP) is

satisfied. Thus, we associate a vector v_p to every point p of B, taking (PDP) into account.

One important example of such a motion is a spatial rotation about the origin. Here, there is always a line A involved, the so-called axis of rotation, containing the origin. Points on A remain fixed (zero velocity vector), but for other points p the velocity v_p is perpendicular to the plane determined by A and p. As a matter of fact, the rotation is specified by a vector ω along A, such that $v_p = \omega \times p$ (vector cross product). The length of ω is called the angular velocity, and together with the distance of p from the axis A, it determines the length of v_p.

Another fundamental motion is a translation along a given vector v. Here, we have a constant velocity: for every point p we put $v_p = v$.

A crucial theorem says that every rigid motion is the composition of rotations and translations, or infinitesimally, the velocity vectors can be written as the sum of rotation velocities and/or translation velocities.

Consider a rotation about some axis A, not necessarily containing the origin. If we embed affine 3-space into $\mathbb{P}^3$, as described in section 2.1, then we can associate with A a projective line $\mathbf{A}$, and hence a Plücker vector $\mathbf{P}_A$. For each point p in $\mathbb{R}^3$ we choose the standard homogeneous coordinates for the associated projective point $\mathbf{p}$ (having $p_4 = 1$). Now we can define the "motion of $\mathbf{p}$" as the following 3-tensor:

$$M(\mathbf{p}) = \mathbf{M} = \mathbf{P}_A \wedge \mathbf{p} \in V^{(3)}.$$

To see that this makes sense, consider a vector $\mathbf{M} = (M_{123}, M_{124}, M_{134}, M_{234})$ of Plücker coordinates. This determines the vector $v_p = (M_{234}, -M_{134}, M_{124}) \in \mathbb{R}^3$, which is zero if $p \in A$, or else it is perpendicular to the plane determined by p and A. And indeed, as one can prove that (PDP) holds for these vectors, they represent a rotation about axis A. The unused coordinate M_{123} in $M(\mathbf{p})$ is determined by the fact that this 3-tensor corresponds to a plane through $\mathbf{p}$. Of course, the magnitude of the vectors v_p depends on the chosen Plücker coordinates $\mathbf{P}_A$ for $\mathbf{A}$, but then again there are an infinite number of possible rotations about axis A in $\mathbb{R}^3$. One can say that the magnitude of the chosen Plücker vector accounts for the involved angular velocity. We conclude that the 2-tensor $\mathbf{P}_A$ encodes both the rotation axis A and the angular velocity. Therefore, it is called the *center* of the motion. Taking a multiple of $\mathbf{P}_A$ does not change the axis, only the angular velocity. If you are interested in the velocity of a specific point p under this motion, just perform the exterior product $\mathbf{P}_A \wedge \mathbf{p}$, using standard homogeneous coordinates for $\mathbf{p}$.

Now that we have put spatial rotations in the setting of projective geometry, we can extend the notion of rotation axis. Indeed, we can take $\mathbf{A}$ to be a line at infinity, so if $\mathbf{P} = \mathbf{P}_A$, then $\mathbf{P}_{14} = \mathbf{P}_{24} = \mathbf{P}_{34} = \mathbf{0}$. If we copy the previous computations for some point p, we observe, surprisingly, that the last three Plücker coordinates of $M(\mathbf{p})$ do not depend on p. Therefore, we see that v_p is a constant vector if we perform a rotation about an axis at infinity, which must be a translation! More precisely, $v_p = (\mathbf{P}_{23}, -\mathbf{P}_{13}, \mathbf{P}_{12})$, a vector which is perpendicular to any plane in $\mathbb{R}^3$ whose projective extension contains the given axis at infinity $\mathbf{A}$. For the sake of uniformity, we will again call the 2-tensor $\mathbf{P}_A$ the center of the motion, and the 3-tensor $M(\mathbf{p})$ the motion itself of the point p.

Our arguments will directly take place in $\mathbb{P}^3$ or $\wedge\mathbb{R}^4$, but readers who like to switch to affine space now and then should remember

$$\begin{aligned} p = (p_1, p_2, p_3) &\longrightarrow \mathbf{p} = (p_1, p_2, p_3, 1), \\ v_p = (M_{234}, -M_{134}, M_{124}) &\longleftarrow M(\mathbf{p}) = (M_{123}, M_{124}, M_{134}, M_{234}). \end{aligned}$$

In this setting, the zero-tensor in $V^{(3)}$ corresponds to the zero velocity.

As mentioned before, composing two motions comes down to adding the velocity vectors in each point p. Let the corresponding centers of motion be denoted by $\mathbf{C}_1$ and $\mathbf{C}_2$, Plücker vectors in $\mathbb{R}^6$. Then the resulting motion of p equals

$$\mathbf{C}_1 \wedge \mathbf{p} + \mathbf{C}_2 \wedge \mathbf{p} = (\mathbf{C}_1 + \mathbf{C}_2) \wedge \mathbf{p}$$

due to a basic property of the exterior product. Now we can consider $\mathbf{C} = \mathbf{C}_1 + \mathbf{C}_2$ to be the center of the composite motion. This means that every Plücker vector $\mathbf{P}$ in $\mathbb{R}^6$ can play the part of a center of some motion. More precisely, if $\mathbf{P}$ represents a projective Line (satisfying (GP)), then it gives rise to a rotation (finite line) or a translation (line at infinity); otherwise it is the center of a composition of rotations and translations. As a consequence of Poinsot's central axis theorem (section 2.2) we can be even more specific in the latter case. To this end, we define a *screw motion* as the composition of a rotation (infinitesimal, of course) about some axis, and a translation (ditto) along the same axis. **If a motion is not a pure translation or rotation, then it is a screw motion.**

From now on, Plücker coordinates of 2-tensors (the space $\mathbb{R}^6$) are interpreted as centers of infinitesimal rigid motions.

3. A simple model for bowling a cricket delivery. Biomechanical models for cricket motions are not that rare, but few exist for bowling [1, 11]. For our purposes, a model simpler than either of these will suffice. We make the following assumptions about motion just prior to delivery:

1. There is no rotation in the elbow (as is required in a legal delivery).
2. There is no rotation in the wrist, and the state of motion of the ball upon release prescribes the motion of the so-called tool center (a term from robotics), which we take to be the wrist.
3. The spine is taken as rigid but free to rotate as if its base is attached to the pelvis in a ball joint (i.e., we ignore deformation of the torso), and the shoulder is rigidly joined to the spine.
4. The joint axes of both joints pass through the center of the joint.

For greater realism, one might add more joints; for example, it is known that the shoulder does rotate relative to the torso [8], and the ball might leave the hand in a contact motion. This is not conceptually difficult but is computationally and experimentally challenging. The same applies to relaxing assumption 4, to allow noncoincident joint axes. Still more challenging would be the direct modeling of muscle groups (as in [6, 16]), as this would increase the number of axes of rotation substantially, and one might be hard put to identify an axis of rotation for every muscle group, particularly those with attachments over more than one joint.

3.1. Introducing the joint axes of our model. With assumptions 1 to 4, the system reduces to two joints, which we call the waist (w) and the shoulder (s). Although in general w may be in motion, there is no loss of generality if we place w at the fixed origin and identify its joint axes with axes of a reference coordinate system XYZ. They are interpreted as follows: for a person standing, X points horizontally forward, Y horizontally points to the left, and Z points vertically along the spine, in our case upwards.

We choose units of length so that the right shoulder joint s is at $(0, -1, 1)$ in the system; since the torso is rigid, it stays there. The three joint axes through s follow the usual convention: We choose S_1 as the axis that passes through shoulder and

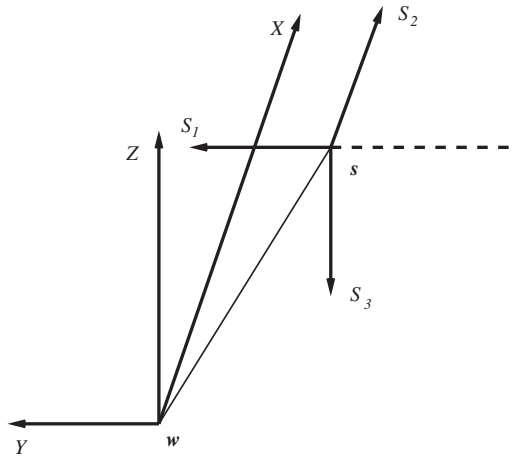


FIG. 3.1. *Configuration of joint axes in our simplified model of a right-handed cricket bowler, facing away from the reader.* **w** *is the waist joint and* **s** *is the shoulder joint. Distances are normalized so that the shoulder is at* $(0, -1, 1)$ *in the waist joint axes. The dashed line corresponds to the arm in standard position: horizontal, palm down. Note that the shoulder axes move with the arm so that* S_3 *is always pointing in the same direction as the palm.*

elbow, in the direction of the shoulder, and S_2, S_3 perpendicular to each other and to S_1 so that when the arm is extended sideways horizontally wrist down, S_2 points forwards and S_3 points downwards. This means that the $S_1S_2S_3$ system moves with the arm, and in particular that S_3 is always perpendicular to the palm. The general configuration is illustrated in Figure 3.1.

We note that some of our results below depend on the choice of shoulder axes. In particular, we find a case where rotation about the s_3 axis plays a significant role in predicting injury risk. This would be indefensible if all we knew of the shoulder joint was that it had 3 degrees of rotational freedom, because our result would disappear under many other apparently equivalent choice of axes.

However, we do know more about how the shoulder moves and about the motion of the arm of a fast bowler near the point of release. In that context, the $s_1s_2s_3$ system as described above is preferred and has intrinsic interpretation for two reasons. First, s_1 is an anatomically intrinsic axis in all rotations of the shoulder, because of the role of the rotator cuff, which are the only muscles that cause rotation around s_1. Second, during the final phase of the delivery of a fast ball in cricket, the bowler's arm moves in a plane. Near the moment of release, the direction of s_3 is tangential to this motion (since these bowlers aim for high speed deliveries), so the plane of motion is the s_1s_3 plane, and in that plane the motion is a pure rotation around the s_2 axis. By orthogonality to both the s_1 and s_2 axes, the s_3 axis is also intrinsic.

3.2. Plücker coordinates of the 6 joint axes. The positions of the waist and the shoulder are given by

$$w = \begin{pmatrix} 0 & 0 & 0 & 1 \end{pmatrix} \qquad \text{and} \qquad s = \begin{pmatrix} 0 & -1 & 1 & 1 \end{pmatrix}.$$

The directions of the joint axes are

$$\begin{aligned} W_1 &= \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}, \\ W_2 &= \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix}, \\ W_3 &= \begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix}, \\ S_1 &= \begin{pmatrix} a_1 & b_1 & c_1 & 0 \end{pmatrix}, \\ S_2 &= \begin{pmatrix} a_2 & b_2 & c_2 & 0 \end{pmatrix}, \\ S_3 &= \begin{pmatrix} a_3 & b_3 & c_3 & 0 \end{pmatrix}. \end{aligned}$$

The six centers of rotation are then

$$\begin{aligned} P_1 = W_1 \wedge w &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}, \\ P_2 = W_2 \wedge w &= \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \\ P_3 = W_3 \wedge w &= \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \\ P_4 = S_1 \wedge s &= \begin{pmatrix} a_1 & b_1 & c_1 & 0 \\ 0 & -1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} -a_1 & a_1 & a_1 & b_1 + c_1 & b_1 & c_1 \end{pmatrix}, \\ P_5 = S_2 \wedge s &= \begin{pmatrix} a_2 & b_2 & c_2 & 0 \\ 0 & -1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} -a_2 & a_2 & a_2 & b_2 + c_2 & b_2 & c_2 \end{pmatrix}, \\ P_6 = S_3 \wedge s &= \begin{pmatrix} a_3 & b_3 & c_3 & 0 \\ 0 & -1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} -a_3 & a_3 & a_3 & b_3 + c_3 & b_3 & c_3 \end{pmatrix}. \end{aligned}$$

We collect these in the columns of the motion matrix M:

$$M = \begin{pmatrix} 0 & 0 & 0 & -a_1 & -a_2 & -a_3 \\ 0 & 0 & 0 & a_1 & a_2 & a_3 \\ 1 & 0 & 0 & a_1 & a_2 & a_3 \\ 0 & 0 & 0 & b_1 + c_1 & b_2 + c_2 & b_3 + c_3 \\ 0 & 1 & 0 & b_1 & b_2 & b_3 \\ 0 & 0 & 1 & c_1 & c_2 & c_3 \end{pmatrix}.$$

All the information regarding configurations of the joints and possible motions can be found by analyzing M. More precisely, infinitesimally, the motion of the wrist is a composition of a rotation about w and a rotation about s (in our model). Thus, the center of this motion is a linear combination of the six Plücker coordinates which we assigned to the six given axes. This motivates us to define the column space of the matrix M to be the *motion space* (of the wrist in the given position of the human body), $\mathbb{MS}$. Recall from Theorem 2.1 that the matrix M always has rank equal to 5, implying a constant dimension of 5 for the motion space. Notice that we never obtain the full $\mathbb{R}^6$ as motion space in our model; this would require including further rotations in our model, such as about the elbow or the wrist.

3.3. Possible motions under the model. Suppose the human body (in particular, the torso and the bowling arm) is in a certain position. If one intends to propel the ball in some specific way, then this is accomplished by performing an infinitesimal motion with the hand. In our model, the only way to realize a hand motion is by means of rotations about the waist (3 joint axes) and/or about the shoulder (3 joint

axes). Every (infinitesimal) rotation about one of these 6 axes is given by an appropriate multiple of the corresponding Plücker vector. We conclude that the motion of the ball is controlled by a 2-tensor which is a linear combination of the 6 Plücker vectors of our model; that is, it belongs to the column space of the matrix M ($\mathbb{MS}$). In particular, a linear combination which gives the zero 2-tensor corresponds to not moving at all (the zero center of motion).

Clearly, the first two rows of M are equal in magnitude but opposite in sign. This implies that every possible motion is represented by a 2-tensor with opposite Plücker coordinates in the first two places,

$$\mathbf{B} = (-a, a, b, c, d, e),$$

or equivalently, a possible motion is a point of $\mathbb{R}^6$ in the hyperplane $\mathbb{H} : p_{12} = -p_{13}$, so $\mathbb{MS} \subset \mathbb{H}$. Furthermore, since both spaces have dimension 5, we can state that $\mathbb{MS} = \mathbb{H}$.

Example 3. Try to perform a *pure* translation with your hand along the Z-axis (the direction of the spine) by only using the waist joint and the shoulder joint. You will not succeed! The algebraic proof for this goes as follows. Each translation along Z is represented by a set of Plücker coordinates of the line at infinity of the XY-plane. This means that it is a multiple of

$$(1, 0, 0, 0) \wedge (0, 1, 0, 0) = (1, 0, 0, 0, 0, 0),$$

which is not a possible motion, because it does not belong to $\mathbb{H}$.

Example 4. In an analogous fashion we see that a *pure* translation along the Y-axis is not possible. Indeed, such a translation is always represented by a multiple of $(0, 1, 0, 0, 0, 0)$, the Plücker vector for the line at infinity of the XZ-plane.

Example 5. However, a translation along the X-axis appears to be possible (this is the direction perpendicular to the plane of the torso; fortunately for cricketers, this direction is the one they want the ball to go). Indeed, the corresponding 2-tensor is a multiple of $(0, 0, 1, 0, 0, 0)$, the line at infinity of YZ; hence it belongs to $\mathbb{MS}$. But how can this be accomplished in practice? Let L be the line through s and parallel to Y. Because the Plücker vector of L is a linear combination of the Plücker vectors of S_1, S_2, and S_3, any rotation about L can be realized. Notice that L lies in the YZ-plane, as does the shoulder joint s in our model. Because Y and L intersect at infinity, an appropriate linear combination of their Plücker vectors yields the line at infinity of YZ. We conclude that a pure translation along X can be realized by composing a rotation about Y and a rotation about L.

Forbidden motions are interesting for two reasons. First, they are a simple way of describing what is possible. Second, they have an associated injury risk: attempting forbidden motion will introduce extremely large stresses, and coming close to forbidden motion (in the sense of a path through the motion space) may also require large stresses, a well-known phenomenon in robotics [3].

3.4. Critical positions of the human body. In our model, the possible motions are supported by six joint axes, each with a natural physical interpretation. Giving the spatial positions of these six axes determines what we will call the "position of the human body." In the previous sections we explained that our analysis of cricket bowling comes down to exploring the linear relations between the Plücker coordinates of these six axes. To simplify computations we are entitled to make the fixed coordinate frame coincide with the three waist axes, and that is what we did in

section 3.2, yielding the simple structure of matrix M. So, in fact, by describing a body position we will mean the specification of the relative position of the shoulder axes with respect to the waist axes. It needs 3 parameters to be specified in order to fix the orientation of $S_1S_2S_3$ relative to $W_1W_2W_3$ $(= XYZ)$, for example the 3 Euler angles. Thus the very limited positions of the human body relevant to this paper can be regarded as points in a 3-dimensional space **Pos**.

3.5. Redundancy and supports. As a consequence of Theorem 2.1 we know that, in each position of the body, our six joint axes span a 5-dimensional space ($\mathbb{MS}$). We say that our kinematic system has a *generic redundancy*. Further, still in each position, basic linear algebra teaches us that we have a 1-dimensional space of linear dependencies between our six 2-tensors $(6 - 5 = 1)$. Redundancy in a model for human motion is also treated in [15], where the emphasis is also on the potential for fatigue management but the operational definition and mathematical treatment are different.

DEFINITION 3.1. *The* support *of a linear dependency among a set A of vectors is the subset of A consisting of exactly those vectors with nonzero coefficient in this dependency.*

In a given position of the body, each (nontrivial) linear dependency of the six joint axes is a multiple of every other one. Thus, we can define merely the "support of a body position" without specifying the linear dependency. Notice that, whatever position we are in, we always use the same notation for our six joints axes; hence the support can always be considered as a subset of $J = \{X, Y, Z, S_1, S_2, S_3\}$. This can be mathematically encoded in a map:

$$\mathbf{supp} : \mathbf{Pos} \to 2^J : p \mapsto \mathbf{supp}(p).$$

Before proceeding, let us explain the relevance of the this concept. Suppose the body is in some position p. Let $\mathbf{M} = (M_{12}, M_{13}, M_{14}, M_{23}, M_{24}, M_{34})$ be the motion in $\mathbb{MS}$ that we want to perform. This is achieved by finding appropriate coefficients (angular velocities):

$$\mathbf{M} = \alpha\mathbf{X} + \beta\mathbf{Y} + \gamma\mathbf{Z} + \sigma_1\mathbf{S}_1 + \sigma_2\mathbf{S}_2 + \sigma_3\mathbf{S}_3,$$

where the bold font reminds us of the fact that we switched to Plücker vectors (or 2-tensors). Now suppose that $\mathbf{supp}(p) = \{Y, Z, S_3\}$, corresponding to the following relation:

$$\lambda\mathbf{Y} + \mu\mathbf{Z} + \nu\mathbf{S}_3 = \mathbf{0}$$

with nonzero coefficients λ, μ, ν. Then we can realize the same motion $\mathbf{M}$ as

$$\mathbf{M} = \alpha\mathbf{X} + (\beta + k\lambda)\mathbf{Y} + (\gamma + k\mu)\mathbf{Z} + \sigma_1\mathbf{S}_1 + \sigma_2\mathbf{S}_2 + (\sigma_3 + k\nu)\mathbf{S}_3$$

with k an arbitrary constant. This means that the efforts done by Y, Z, and S_3 can be traded among each other, while the contributions by X, S_1, and S_2 are given by fixed coefficients with no chance for compensation. From this we learn two important things:

1. The concept of redundancy of joint axes is inherent to the human body. It is the solution supplied by nature to distribute the necessary efforts among the several joints for achieving a certain motion.
2. Positions in which the human body has abundant support are less strenuous than positions with limited support.

3.6. Critical positions. Now we arrive at the core of this paper. We will classify the possible supports in our model. A position of the human body is called *critical* if the support is smaller than expected, that is, smaller than in generic positions. We say that a critical position suffers from *redundancy with reduced support* or shortly, *reduced redundancy*. Our first observation says that the required work for joint axis X can never be compensated by one of the other five axes.

THEOREM 3.2. *For each position* $p \in \mathbf{Pos}$ *we have that* $X \notin \mathbf{supp}(p)$.

Proof. Since the shoulder joint s is assumed to lie in the YZ-plane, the Line $\mathbf{L} = sw$ is a linear combination of $\mathbf{Y}$ and $\mathbf{Z}$. And of course, $\mathbf{L}$ is a linear combination of $\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3$; hence the set $\{\mathbf{S_1}, \mathbf{S_2}, \mathbf{S_3}, \mathbf{Y}, \mathbf{Z}\}$ is dependent. Because the motion space $\mathbb{MS}$ has dimension 5 in every position, $\mathbf{X}$ cannot be a linear combination of $\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3, \mathbf{Y}, \mathbf{Z}$; hence, it does not belong to the support. □

THEOREM 3.3. *Let* p *be a position of the human body. We distinguish three cases for the Line* $\mathbf{L} = sw$.

1. *The Line* $\mathbf{L}$ *is not contained in a plane determined by any two Lines of* $\{\mathbf{S_1}, \mathbf{S_2}, \mathbf{S_3}\}$. *In this case*
$$\mathbf{supp}(p) = \{\mathbf{S_1}, \mathbf{S_2}, \mathbf{S_3}, \mathbf{Y}, \mathbf{Z}\}.$$
2. *The Line* $\mathbf{L}$ *does not coincide with a line of* $\{\mathbf{S_1}, \mathbf{S_2}, \mathbf{S_3}\}$, *but it lies in the plane generated by two of them (*$\mathbf{L} \in \mathbf{S_i S_j}$*). Then*
$$\mathbf{supp}(p) = \{\mathbf{S_i}, \mathbf{S_j}, \mathbf{Y}, \mathbf{Z}\}.$$
3. *The Line* $\mathbf{L}$ *coincides with one of* $\{\mathbf{S_1}, \mathbf{S_2}, \mathbf{S_3}\}$ *(i.e.,* $\mathbf{L} = \mathbf{S_i}$*). Then*
$$\mathbf{supp}(p) = \{\mathbf{S_i}, \mathbf{Y}, \mathbf{Z}\}.$$

Proof. The claims are an immediate consequence of what is said in section 2.3.

In case 3, if $\mathbf{L} = \mathbf{S_i}$, then the Lines $\mathbf{Y}, \mathbf{Z}, \mathbf{S_i}$ are concurrent and coPlanar, and so they are dependent. The support cannot be smaller, because this would mean that at least two of these lines coincide.

In case 2, either Lines $\mathbf{S_i}, \mathbf{S_j}, \mathbf{Y}, \mathbf{Z}$ are coPlanar or the pairs $\{\mathbf{Y}, \mathbf{Z}\}$ and $\{\mathbf{S_i}, \mathbf{S_j}\}$ determine two Planes that meet in the line $\mathbf{sw}$ through their intersections. In both cases, the four Lines are dependent. Furthermore, no three of them are concurrent, implying that the support is not smaller.

In case 1, we can rule out the five possibilities for the dependency of four lines (section 2.3). We refer to Theorem 3.2 for the claim that $\{\mathbf{S_1}, \mathbf{S_2}, \mathbf{S_3}, \mathbf{Y}, \mathbf{Z}\}$ is a dependent set. □

Remark. Cases 2 and 3 of the previous theorem correspond to the critical positions of our model.

4. Reduced redundancy as injury risk. It is known that high levels of fitness are attained in many cricketers [13]; nevertheless, injuries are fairly common [10] and fatigue may play a significant role [7]. This is not the place to review the mechanisms of overuse injury (the interested reader is referred to [17] as a starting point). We adopt the common perspective that overuse injuries start as microinjuries such as bruised bone and microtorn ligament. We suggest that overuse is more likely in situations of reduced redundancy. In such cases, no compensation that reduces the strain on a microinjured site is possible. The subject, in repeating the action, is condemned to repeating, at the same intensity, a motion that already caused a microinjury. By contrast, the ability to achieve a desired motion with a range of different

joint rotations amounts to having the option of avoiding a motion that has caused a microinjury. The probability of overuse injury should decrease; hence, redundancy should correlate with reducing the risk of overuse injury. If so, then bowlers whose body position at ball release has more reduced redundancy than others should be at higher risk of injury, because such bowlers are less able to adapt. We also assume that microinjury is more likely in fatigued tissues, and hence adopt the view that reducing the probability of overuse injury is equivalent to reducing fatigue.

4.1. The role of the various joint axes. We interpret a joint axis that does not belong to the support in a given body position as a "necessary" axis of that position.

The joint axis X is through the "waist" joint and perpendicular to the pelvis; it is more or less parallel to the direction of the ball around the time of release. It is always a necessary axis, so for a particular desired motion, the amount of sideways bending of the spine is prescribed.[1] One implication of this is that injury risk due to this motion cannot be modified.

We noted above that a largely supported redundancy should help to reduce fatigue. Similarly, if an axis is necessary then no fatigue management can reduce the rate of tiring in structures involved in rotations around it. While the human body will have many more joint axes, our analysis suggests that bowlers will find it hard to compensate for fatigue related to rotation around the X-axis. Anecdotal evidence suggests that bowlers may attempt compensation by "falling over" as they tire. However, studies on changes in bowling action over long spells [9] have not reported rotation around this axis, so no scientific judgment is possible.

In critical positions we even suffer from reduced redundancy. The calculations for reduced redundancy depend on the choice of shoulder axes. We argued above that the s_1 axis is anatomically an intrinsic axis of rotation and that s_2 is dynamically an intrinsic axis of rotation for fast bowlers, because the motion of the arm is in the s_1s_3 plane around the time of delivery of a fast ball.

In the bowling of a cricket fast ball, the worst-case scenario of reduced redundancy is that S_1 passes through the waist joint, which corresponds to case 3 above and implies that rotation about the other two shoulder axes are prescribed in all motions. Let us consider the simplest (and also most common) example: a straight arm. For such bowlers, the most risky action is one in which wrist, elbow, shoulder, and waist all lie on the same line very near or at the moment of delivery. Their ability to modify the amount of rotation will be limited to axes S_1, Y, and Z; thus one expects overuse injury related to such rotations to be less common. So for them tradeoffs are only possible among axial rotations of the arm, twisting of the spine, and bending forward at the waist. On the other hand, coaches need to be aware that changing the rotation in one of these axes will cause compensation in the other two axes.

We note that this situation is avoided by releasing the ball either behind or in front of the plane of the torso (more on this below) by a round-arm action, where S_1 is nearer to horizontal, and by a very upright action, where S_1 is nearer to vertical. Vertical action is usually encouraged by coaches but in some cases may tend to align the wrist with shoulder and waist and so increase injury risk.

Furthermore, Y and Z are never necessary, so that the amount of twisting and bending (backwards/forwards, that is) can be modified. Thus, in case of excessive

[1]Some care is needed here: In our model, bending of the spine is approximated by rotation of an inflexible spine in the "waist" joint. It may be that more than one pattern of rotations of vertebrae can achieve the desired rotation.

rotation in these directions at the waist, it should be possible to modify the bowler's action to reduce these, no matter what the configuration of their joints at the moment of delivery. For instance, excessive twisting around the Z-axis during the delivery stride is currently regarded as a major source of injury risk (the "mixed" action, which starts with hips and shoulders facing forwards; then the shoulders rapidly rotate and counterrotate—see [9, 12] and many others). Our study suggests that bowlers using a mixed action should be able to change action with relative ease.

Finally, is it possible to deliver a cricket ball with a maximally supported redundancy? Yes, but such actions are unusual and discouraged by coaches. The Line $\mathbf{L}$ in the analysis above corresponds to the line through waist and shoulder; it is required that this line be perpendicular to none of the shoulder joint axes. For instance, suppose that at delivery, the X and S_3 axes are parallel (certainly an aim in some deliveries by fast- and medium- pace bowlers). Then the wrist should not be in the plane formed by spine and shoulder (otherwise case 2 applies: $\mathbf{L}$ perpendicular to $\mathbf{S}_2$ or, equivalently, $\mathbf{L}$ is a linear combination of $\mathbf{S_1}$ and $\mathbf{S_3}$). So these bowlers should deliver such balls from behind or in front of the torso (the former seems to be common). The other axes have similar requirements. $\mathbf{L}$ perpendicular to $\mathbf{S}_1$ would be an excessively round-arm action and perhaps unlikely (though it could occur in the slinging action of some fast bowlers). $\mathbf{L}$ perpendicular to $\mathbf{S}_3$ is perhaps harder to avoid but should still be rare; for instance a round-arm action with the palm down at the moment of release, which might occur in some spin bowling actions.

4.2. An example. We give an analysis of the action of two medium-fast bowlers, both from the youth academy and hence at risk of injury, as potentially elite medium-fast or fast bowlers. The data were kindly provided by Janine Gray of Sports Science Institute of South Africa, who collected the data on these two bowlers as part of a larger study. Both subjects were 17 years old and free of injury at the time the data were collected. Bowler B had a long history of injuries, some of them from noncricket activities. In particular, he had suffered a stress injury to the lower back, which was seen as due to cricket. Bowler A had never been injured. Their historical workloads were different—Bowler B had played cricket from early boyhood, while Bowler A was a recent recruit to the game.

For each bowler, reflectors were attached to the body surface. Under stroboscopic lighting (frequency 120 Hz), video cameras recorded the positions of the reflectors at intervals (interval length about 8 milliseconds). The following reflectors were used in the calculation below: two on the wrist, one on the shoulder, and three on the waist. The three waist coordinates were assumed to lie at the vertices of a symmetric trapezium, and the center of its circumrectangle was calculated to give w, the center of the waist joint. In calculating s, the center of the shoulder joint, we assumed that the shoulder is fixed relative to the waist, so a simple correction allowed us to move from the position of the reflector on the acromion to s. The midpoint of the two reflectors on the wrist provided the position of r, the center of the wrist. In Figures 4.1 and 4.2 we depict aspects of the raw data: wrist position as a function of time for both bowlers.

Calculation of redundancy then proceeds as follows. The simple subtraction $w-s$ and normalization gave us the unit vector $\mathbf{l}$, which gives the direction of the Line $\mathbf{L}$ through waist and shoulder. Similarly, $s-r$ gives $\mathbf{s_1}$, the direction of S_1. Since the wrist reflectors lie in the S_1S_2-plane as does $\mathbf{s_1}$, simple orthogonalization gives $\mathbf{s_2}$, and $\mathbf{s_3}$ is then available as the cross-product of $\mathbf{s_1}$ and $\mathbf{s_2}$. The dot products of $\mathbf{l}$ with the $\mathbf{s_i}$ are then calculated, giving the direction cosines of $\mathbf{l}$ in the $S_1S_2S_3$ axes. When

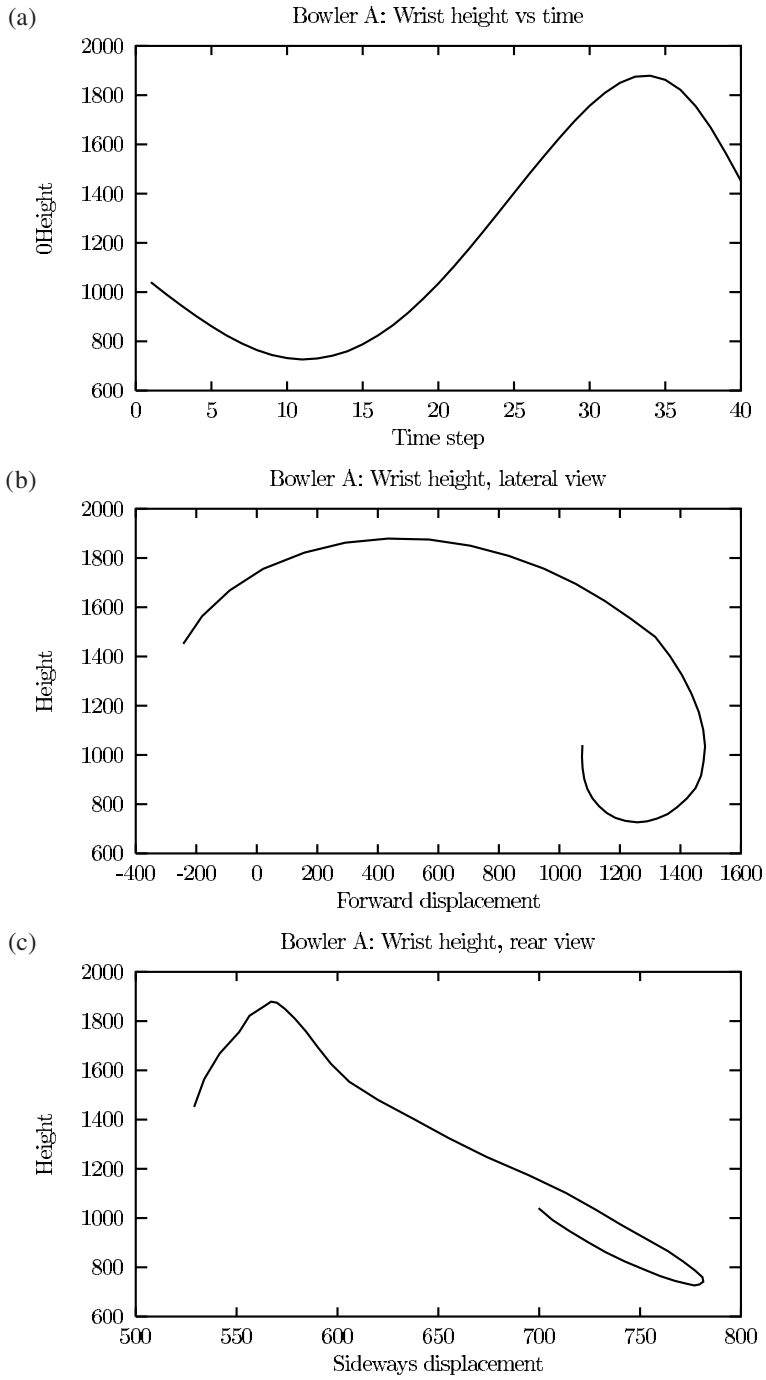


FIG. 4.1. *Wrist position as a function of time for Bowler A. The three diagrams give different aspects.* (a) *Wrist height as a function of time.* (b) *Wrist path in the YZ-plane (movement is leftward on diagram).* (c) *Wrist path in the XZ-plane. Height and displacement in millimeters; time steps* 0.83 *milliseconds apart.*

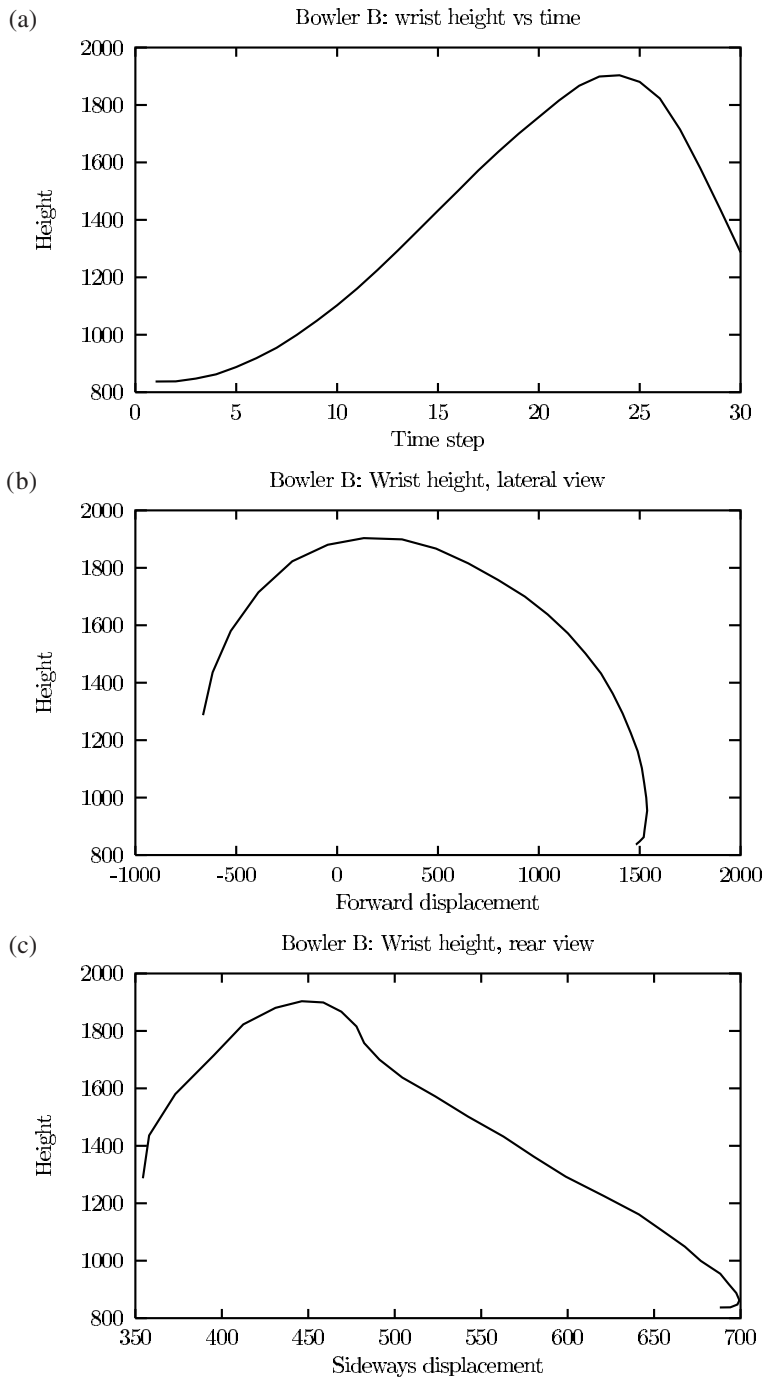


FIG. 4.2. *Wrist position as a function of time for Bowler B. The three diagrams give different aspects.* (a) *Wrist height as a function of time.* (b) *Wrist path in the* YZ*-plane.* (c) *Wrist path in the* XZ*-plane. Height and displacement in millimeters; time steps* 0.83 *milliseconds apart.*

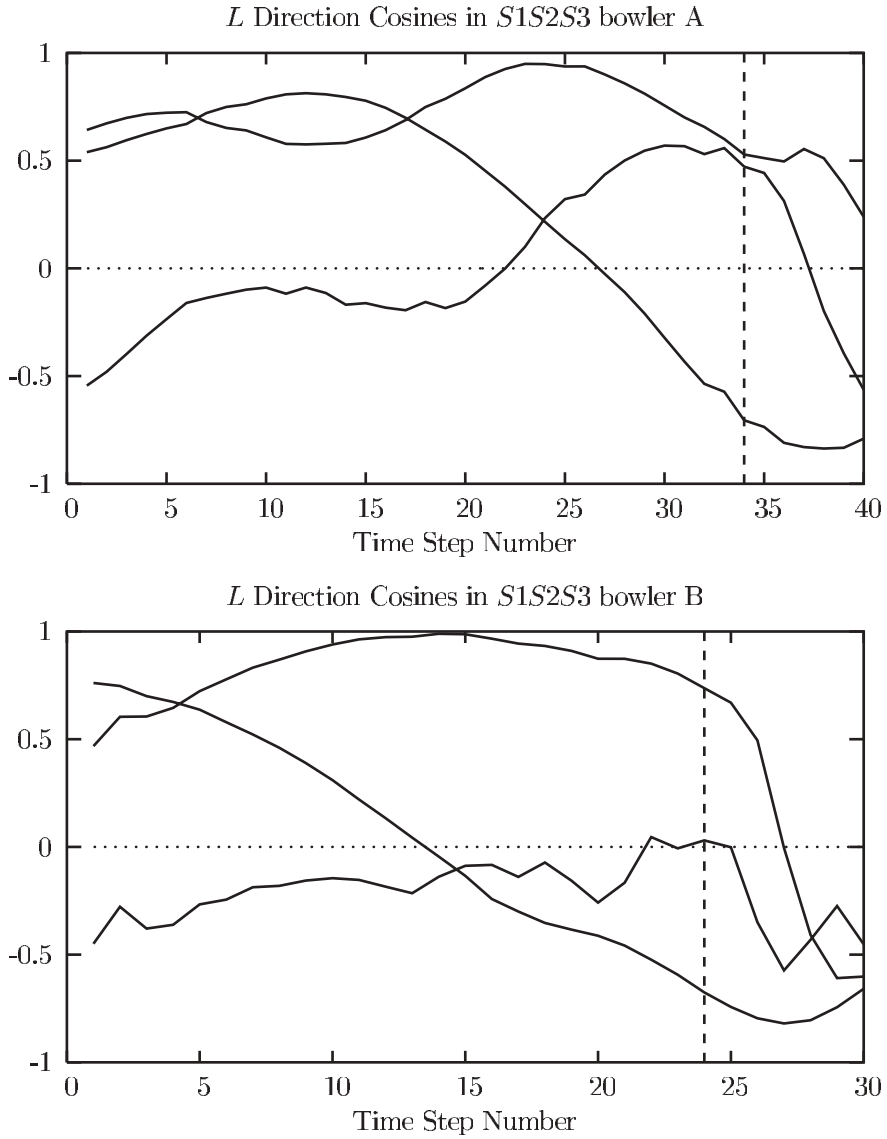


FIG. 4.3. *Reduced redundancy as revealed by the direction of the waist-shoulder line in shoulder coordinates. Curves give direction cosines of L, the line from waist to shoulder, in the axes S1, S2, and S3 of the shoulder; they are plotted against time steps (one unit of time is a few milliseconds). Vertical lines indicate the approximate points of release. Whenever one of the direction cosines goes to zero, reduction of redundancy occurs and the corresponding axis is absent from the support of the body position. Bowler A maintains full support until well after the moment of delivery, but Bowler B loses the S_3 axis from the support of the motion for about* 15 *milliseconds on either side of the moment of release.*

reduced redundancy occurs, then $\mathbf{l}$ is perpendicular to one or two of the $\mathbf{s_i}$—that is, one of the direction cosines is zero. This is easily spotted on a graph of direction cosines vs. time (see Figure 4.3).

The plots in Figure 4.3 cover approximately 0.4 s in time and forward motion of about 2 meters in space. Note that for Bowler A all three direction cosines stay well

away from zero in the period prior to release, but that for Bowler B the S_3 axis goes to zero about 15 milliseconds before release, and stays there for about 30 milliseconds. With respect to our choice of axes, Bowler B operates with reduced redundancy around the time of release of the ball, but not Bowler A. This suggests that Bowler B may be less able to modify his action to cope with fatigue. This is consistent with their injury history, as Bowler B indeed has had more injuries than Bowler A. However, it is also true that Bowler B has had much more opportunity for overuse than Bowler A, due to a far longer playing career. We hope to track both subjects to test whether indeed Bowler A will remain relatively free from overuse injury, as we predict.

REFERENCES

[1] A. Burnett, C. Barrett, R. Marshall, B. Elliott, and R. Day, *Three-dimensional measurement of lumbar spine kinematics for fast bowlers in cricket*, Clin. Biomech., 13 (1998), pp. 574–593.

[2] L. Chèze, C. Gutierrez, R. San Marcelino, and J. Dimmet, *Biomechanics of upper limb using robotic techniques*, Human Movement Sci., 15 (1996), pp. 477–496.

[3] J. Craig, *Introduction to Robotics*, Addison-Wesley, 1989.

[4] H. Crapo and W. Whiteley, *Statics of frameworks and motions of panel structures: A projective geometric introduction*, Struct. Topol., 6 (1982), pp. 43–82.

[5] A. Dandurand, *The rigidity of compound spatial grids*, Struct. Topol., 10 (1984), pp. 41–56.

[6] P. Eberhard, T. Spägele, and A. Gollhofer, *Investigations for the dynamical analysis of human motion*, Multibody Sys. Dynam., 3 (1999), pp. 1–20.

[7] B. Elliot, *Back injuries and the fast bowler in cricket*, J. Sports Sci., 18 (2000), pp. 983–991.

[8] B. Elliott, R. Wallis, S. Sakurai, D. Lloyd, and T. Besier, *The measurement of shoulder alignment in cricket fast bowling*, J. Sports Sci., 20 (2002), pp. 407–510.

[9] P. Glazier, G. Paradisis, and S.-M. Cooper, *Anthropometric and kinematic influences on release speed in men's fast-medium bowling*, J. Sports Sci., 18 (2000), pp. 1013–1021.

[10] T. Leary and J. A. White, *Acute injury incidence in professional county club cricket players*, British J. Sports Med., 34 (2000), pp. 145–147.

[11] D. Lloyd, J. Alderson, and B. Elliott, *An upper limb kinematic model for the examination of cricket bowling: A case study of Mutiah Muralitheran*, J. Sports Sci., 18 (2000), pp. 975–982.

[12] A. McGrath and C. Finch, *Bowling Cricketing Injuries Over: A Review of the Literature*, Tech. Report 105, Monash University Accident Research Centre, 1996.

[13] T. Noakes and J. Durandt, *Physiological requirements of cricket*, J. Sports Sci., 18 (2000), pp. 919–929.

[14] M. Portus, P. Sinclair, S. Burke, D. Moore, and P. Farhart, *Cricket fast bowling performance and technique and the influence of selected physical factors during an 8-over spell*, J. Sports Sci., 18 (2000), pp. 999–1011.

[15] V. Potkonjak, S. Tzafestas, D. Kostic, and G. Djordjevic, *Human-like behaviour of robot arms: General considerations and the handwriting task—Part* I: *Mathematical description of human-like motion: distributed positioning and virtual fatigue*, Robotics and Computer Integrated Manufacturing, 17 (2001), pp. 305–315.

[16] A. van den Bogert and H. Schamhardt, *Multi-body modelling and simulation of animal locmotion*, Acta Anatomica, 146 (1993), pp. 95–102.

[17] A. van Mechelen, *Running injuries. A review of the epidemiological literature*, Sports Medicine, 14 (1992), pp. 320–335.

[18] N. White, *Grassmann-Cayley algebra and robotics*, J. Intelli. Robot. Sys., 11 (1994), pp. 91–107.

SIAM J. APPL. MATH.
Vol. 65, No. 2, pp. 720–736

ON THE GEOMETRIC FLOW OF KIRCHHOFF ELASTIC RODS*

CHUN-CHI LIN† AND HARTMUT R. SCHWETLICK‡

Abstract. Recently, rod theory has been applied to the mathematical modeling of bacterial fibers and biopolymers (e.g., DNA) to study their mechanical properties and shapes (e.g., supercoiling). In static rod theory, an elastic rod in equilibrium is the critical point of an elastic energy. This induces a natural question of how to find elasticae. In this paper, we focus on how to find the critical points by means of gradient flows. We relate a geometric function of curves to the isotropic Kirchhoff elastic energy of rods so that the generalized elastic curves are the centerlines of elastic rods in equilibrium. Thus, the variational problem for rods is formulated in curve geometry. This problem turns out to be a generalization of curve-straightening flows, which induce nonlinear fourth-order evolution equations. We establish the long time existence of length-preserving gradient flow for the geometric energy. Furthermore, by studying the asymptotic behavior, we show that the limit curves are the centerlines of the Kirchhoff elastic rods in equilibrium.

Key words. geometric flows, fourth order, Kirchhoff elastic rods

AMS subject classifications. 35K55, 53C44, 74B20, 74L15

DOI. 10.1137/S0036139903431713

1. Introduction.

Recently, rod theory has been applied to the mathematical modeling of bacterial fibers and biopolymers (e.g., DNA) to study their mechanical properties and shapes (e.g., supercoiling). In static rod theory, an elastic rod in equilibrium is the critical point of an elastic energy. This induces a natural question of how to find elasticae. In our project, we ask the question, Starting from a given rod configuration Γ in $\mathbb{R}^3$, can we find the critical points of a Kirchhoff elastic energy, or the so-called elasticae, by means of geometric gradient flows? In order to keep the model problem in this paper simple, we consider only a special isotropic Kirchhoff elastic energy. For more general rod theory, readers are referred to [1].

Suppose $f : I = \mathbb{R}/\mathbb{Z} \to \mathbb{R}^3$ is the centerline of a closed rod. Let $\gamma = |\partial_x f|$, $ds = \gamma\, dx$ be the arclength element, and $\partial_s = \gamma^{-1}\partial_x$ be the arclength differentiation. Denote by $T = \partial_s f$ the unit tangent vector and by $\kappa = \partial_s^2 f$ the curvature vector of f. A rod configuration Γ is a framed curve described by $\{f(s); T(s), M_1(s), M_2(s)\}$, where the material frame $\{T, M_1, M_2\}$ forms an orthonormal frame field along f. Thus, we can write the skew-symmetric system

$$\begin{pmatrix} T' \\ M_1' \\ M_2' \end{pmatrix} = \begin{pmatrix} 0 & m_1 & m_2 \\ -m_1 & 0 & m \\ -m_2 & -m & 0 \end{pmatrix} \begin{pmatrix} T \\ M_1 \\ M_2 \end{pmatrix},$$

with arbitrary functions $m_1(s)$, $m_2(s)$, and $m(s)$. Consider the Kirchhoff elastic energy $\mathcal{E}$ of an isotropic rod Γ, defined by

$$\mathcal{E}[\Gamma] := \int_I \left(\alpha \cdot (m_1^2 + m_2^2) + \beta \cdot m^2\right) ds,$$

*Received by the editors July 17, 2003; accepted for publication (in revised form) June 3, 2004; published electronically January 27, 2005. This work was supported by the Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany, and the Taiwan NSC grant 92-2119-M-003-017.
http://www.siam.org/journals/siap/65-2/43171.html

†Department of Mathematics, National Taiwan Normal University, Taipei 116, Taiwan (chunlin@math.ntnu.edu.tw).

‡Max Planck Institute for Mathematics in the Sciences, Inselstr. 22-26, Leipzig D-04103, Germany (schwetlick@mis.mpg.de).

with material constants $\alpha > 0$ and $\beta \geq 0$. The term involving α gives the bending energy, while the term involving β gives the twisting energy.

Whenever a smooth curve f has no inflection points, the Frenet frame field $\{T, N, B\}$ along f is well defined. By using the Frenet frame field, it can be easily verified that

$$\mathcal{E}[\Gamma] = \int_I \left(\alpha |\kappa|^2 + \beta m^2\right) ds \tag{1.1}$$

(e.g., see [7]). A natural frame is an orthonormal frame field along a given curve f, which is uniquely determined by its initial data at a point and the skew-symmetric system,

$$\begin{pmatrix} T' \\ U' \\ V' \end{pmatrix} = \begin{pmatrix} 0 & u & v \\ -u & 0 & 0 \\ -v & 0 & 0 \end{pmatrix} \begin{pmatrix} T \\ U \\ V \end{pmatrix}$$

(see [3] or [7, p. 607]). A natural frame can be thought as a frame without twisting. As we denote by θ the angle from U to M_1 with $\theta(0) = 0$, one can verify that m is equal to the twisting rate, i.e., $m(s) = \theta'(s)$. Whenever f contains no inflection points, the Frenet frame is well defined along f. Denote by ϕ the angle from U to N; then it is easy to verify that the torsion of the curve satisfies $\tau = \phi'$. Denote by $\Psi := \theta - \phi$ the angle from N to M_1 and let $\triangle\Psi := \Psi(L) - \Psi(0)$, where L is the total length of f. By these notations, we have

$$Tw[\Gamma] = \int_I m \, ds = \triangle\Psi + \int_I \tau \, ds. \tag{1.2}$$

It is worth mentioning here that whenever f contains neither self-intersection nor inflection points, applying the so-called Fuller–Calugareanu–White formulas provides another approach to derive (1.2). This approach is less general and less direct, but it reveals the topological meaning of $\triangle\Psi$, although the total twisting number of Γ, $Tw[\Gamma]$, and the total torsion of f, $\int_I \tau \, ds$, are not topological invariants. We thus set up the boundary value problem by prescribing a real number, $\triangle\Psi$, which is called the endpoint condition of rod configurations in the rest of this paper. From above, we would like to emphasize that the bending energy and twisting energy interact as rod configurations achieving the critical points of the elastic energy. More precisely, the twisting depends on the centerlines of rods as well. Otherwise, the twisting energy and bending energy can be considered separately, and the resulting centerlines of rod elasticae would simply be curve elasticae.

In [7], Langer and Singer proposed to study the generalized elastic curves by introducing the geometric functional $\widetilde{\mathcal{F}}$ of curves $f : I \to \mathbb{R}^3$,

$$\widetilde{\mathcal{F}}[f] := \lambda_3 \mathcal{K}[f] + \lambda_2 \mathcal{T}[f] + \lambda_1 \mathcal{L}[f], \tag{1.3}$$

where

$$\mathcal{K}[f] := \int_I \frac{1}{2} |\kappa|^2 \, ds, \ \mathcal{T}[f] := \int_I \tau \, ds, \ \mathcal{L}[f] := \int_I ds,$$

and λ_i in (1.3) are Lagrange multipliers for $i = 1, 2$. According to their formulation, a generalized elastic curve f in equilibrium is a critical point of the elastic energy $\widetilde{\mathcal{F}}$

among the class of curves with fixed total torsion $\mathcal{T}[f] = T_0$ and length $\mathcal{L}[f] = L$. As long as λ_i together with the fixed total torsion T_0 fit certain relations, they showed that f is the centerline of an isotropic elastic rod in equilibrium. The problems considered in this paper and in (1.3) are closely related to curve straightening flows. To the authors' knowledge, curve straightening flows have been studied by Wen [9], Polden [8], Koiso [6], and Dziuk, Kuwert, and Schätzle [4]. At the beginning, we tried to apply the method used in the problems of curve straightening flows to the geometric functional $\widetilde{\mathcal{F}}$ proposed in [7]. However, an essential difficulty coming from the constraint of fixing the total torsion fails this approach. Namely, after multiplying the term of the first variation of the total torsion $\mathcal{T}[f]$ by its Lagrange multiplier, the method of L^2 curvature estimates combined with Gagliardo–Nirenberg-type interpolation inequalities used in the problems of curve straightening flows fails, because this term has higher power of derivatives in total than those from $\mathcal{K}[f]$.

In order to resolve the difficulty mentioned above, we propose another approach based on Theorem 1.1 below. We learn from [5] and [7] that a symmetric elastic rod (or, equivalently, an isotropic elastic rod) must have a constant twisting rate. Observe that among all isotropic rod configurations Γ with constant twisting rate $m = \frac{\mathcal{T}[f]+\triangle\Psi}{L}$, and fixed length L but without inflection points, we have the identity

$$\mathcal{E}[\Gamma] = \mathcal{G}_{\triangle\Psi,L}[f] := 2\alpha\mathcal{K}[f] + \frac{\beta}{L}\left(\mathcal{T}[f] + \triangle\Psi\right)^2.$$

Theorem 1.1 basically means that the equilibrium elastic rods must stay in the subclass of rod configurations with constant twisting rate and fixed length L. Thus, in order to find closed elastic rods of $\mathcal{E}$, we work with the geometric functional

$$\mathcal{F}[f] := \mathcal{G}_{\triangle\Psi,L}[f] + \lambda_1 \cdot (\mathcal{L}[f] - L), \tag{1.4}$$

where λ_1 is the Lagrange multiplier. It turns out that working with the functional $\mathcal{G}_{\triangle\Psi,L}$ of curves with fixed length L is more suitable than working directly with the rod energy $\mathcal{E}$ in our geometric approach.

THEOREM 1.1. *Let $f : I = \mathbb{R}/\mathbb{Z} \to \mathbb{R}^3$ be the centerline of a closed rod Γ. Assume f contains no inflection points. Then, subject to variations of fixed length L and endpoint condition $\triangle\Psi$ in* (1.2)*, Γ is an equilibrium of the elastic energy $\mathcal{E}$ if and only if f is a critical point of the geometric functional $\mathcal{F}$ and the twisting rate is equal to the constant $\frac{\triangle\Psi+\mathcal{T}[f]}{L}$.*

The inflection points in the above theorem simply mean points of zero curvature. We exclude the situation of the limit curves containing inflection points because our argument in Theorem 1.1 relies on the formulation of Frenet frames, which are ill-defined at an inflection point. By the first variational formulas in Lemma 3.1, we obtain, for the length-preserving L^2 gradient flow of $\mathcal{F}$ the evolution equation,

$$\partial_t f = \lambda_3 \cdot \left(-\nabla_s^2\kappa - \frac{|\kappa|^2}{2}\kappa\right) + \lambda_2(t) \cdot \nabla_s(T \times \kappa) + \lambda_1(t) \cdot \kappa, \tag{1.5}$$

where $f : [0,\infty) \times I \to \mathbb{R}^3$ has smooth initial data f_0. Here, the covariant derivative

$\nabla_s\eta$ denotes the normal component of $\partial_s\eta$, i.e., $\nabla_s\eta = \partial_s\eta - \langle\partial_s\eta, T\rangle T$, and

$$\lambda_1(t) := \frac{2\alpha\int_I \langle\kappa, \nabla_s^2\kappa + \frac{|\kappa|^2}{2}\kappa\rangle\, ds - \frac{2\beta}{L}(\mathcal{T}[f] + \triangle\Psi)\int_I \langle\kappa, \nabla_s(T\times\kappa)\rangle\, ds}{\int_I |\kappa|^2\, ds}, \tag{1.6}$$

$$\lambda_2(t) := \frac{2\beta}{L}(\mathcal{T}[f] + \triangle\Psi), \qquad L,\ \beta > 0, \tag{1.7}$$

$$\lambda_3 := 2\alpha, \qquad \alpha > 0. \tag{1.8}$$

Notice that $\lambda_1(t)$ in (1.6) is chosen so that $\frac{d}{dt}\mathcal{L}[f_t] = 0$ for all time. The following theorem is the main result of this paper.

THEOREM 1.2. *For any real number $\triangle\Psi$ and any smooth initial closed curve f_0, there exists a smooth solution to the L^2-gradient flow in* (1.5) *until the appearance of inflection points. With the assumption of no inflection points appearing during the flow, the curves subconverge to f_∞, an equilibrium of the energy functional $\mathcal{F}$, after reparametrization by arclength and translation. Furthermore, if f_∞ contains no inflection points, then f_∞ is the centerline of an equilibrium Kirchhoff elastic rod with constant total twisting rate $\frac{\mathcal{T}[f_\infty]+\triangle\Psi}{L}$.*

This paper is arranged as follows. In section 2 we introduce further notation and collect the results needed from [4]. Since most of these preliminaries follow the lines in [4], the reader is recommended to consult [4] for further details. In section 3 we present the proof of the main results. Finally, section 4 is devoted to the numerical treatment of the problem. We explain the algorithm we have used and show several computational results.

2. Preliminaries.

LEMMA 2.1 (Lemma 2.1 in [4]). *Suppose ϕ is any normal field along f and $f : [0,\epsilon)\times I \to \mathbb{R}^n$ is a time-dependent curve satisfying $\partial_t f = V + \varphi T$, where V is the normal velocity and $\varphi = \langle T, \partial_t f\rangle$. Then the following formulas hold:*

$$\nabla_s\phi = \partial_s\phi + \langle\phi,\kappa\rangle T, \tag{2.1}$$

$$\partial_t(ds) = (\partial_s\varphi - \langle\kappa, V\rangle)\ ds, \tag{2.2}$$

$$\partial_t\partial_s - \partial_s\partial_t = (\langle\kappa, V\rangle - \partial_s\varphi)\ \partial_s, \tag{2.3}$$

$$\partial_t T = \nabla_s V + \varphi\cdot\kappa, \tag{2.4}$$

$$\partial_t\phi = \nabla_t\phi - \langle\nabla_s V + \varphi\ \kappa, \phi\rangle T, \tag{2.5}$$

$$\nabla_t\kappa = \nabla_s^2 V + \langle\kappa, V\rangle\kappa + \varphi\cdot\nabla_s\kappa, \tag{2.6}$$

$$(\nabla_t\nabla_s - \nabla_s\nabla_t)\phi = (\langle\kappa, V\rangle - \partial_s\varphi)\nabla_s\phi + \langle\kappa,\phi\rangle\nabla_s V - \langle\nabla_s V,\phi\rangle\cdot\kappa. \tag{2.7}$$

LEMMA 2.2 (Lemma 2.2 in [4]). *Suppose $f : [0,\widehat{T})\times I \to \mathbb{R}^n$ moves in a normal direction with velocity $\partial_t f = V$, ϕ is a normal vector field along f, and $\nabla_t\phi + \nabla_s^4\phi = Y$. Then*

$$\frac{d}{dt}\frac{1}{2}\int_I |\phi|^2\ ds + \int_I \left|\nabla_s^2\phi\right|^2\ ds = \int_I \langle Y,\phi\rangle\ ds - \frac{1}{2}\int_I |\phi|^2\langle\kappa, V\rangle\ ds. \tag{2.8}$$

Furthermore, $\psi = \nabla_s\phi$ satisfies the equation

$$\nabla_t\psi + \nabla_s^4\psi = \nabla_s Y + \langle\kappa,\phi\rangle\nabla_s V - \langle\nabla_s V,\phi\rangle\kappa + \langle\kappa, V\rangle\psi. \tag{2.9}$$

For normal vector fields $\phi_1,\ldots,\phi_k$ along f, we denote by $\phi_1 * * * \phi_k$ a term of the type

$$\phi_1 * * * \phi_k = \begin{cases} \langle\phi_{i_1},\phi_{i_2}\rangle\cdots\langle\phi_{i_{k-1}},\phi_{i_k}\rangle & \text{for } k \text{ even},\\ \langle\phi_{i_1},\phi_{i_2}\rangle\cdots\langle\phi_{i_{k-2}},\phi_{i_{k-1}}\rangle\cdot\phi_{i_k} & \text{for } k \text{ odd}, \end{cases}$$

where $i_1, \dots, i_k$ is any permutation of $1, \dots, k$. Slightly more generally, we allow some of the ϕ_i to be functions, in which case the $*$-product reduces to multiplication. For a normal vector field ϕ along f, we denote by $P_\nu^\mu(\phi)$ any linear combination of terms of the type $\nabla_s^{i_1}\phi * \cdots * \nabla_s^{i_\nu}\phi$ with universal constant coefficients, where $\mu = i_1 + \cdots + i_\nu$ is the total number of derivatives. Notice that the following formulas hold:

$$\begin{cases} \nabla_s\left(P_b^a(\phi) * P_d^c(\phi)\right) = \nabla_s P_b^a(\phi) * P_d^c(\phi) + P_b^a(\phi) * \nabla_s P_d^c(\phi), \\ \\ P_b^a(\phi) * P_d^c(\phi) = P_{b+d}^{a+c}(\phi),\ \nabla_s P_d^c(\phi) = P_d^{c+1}(\phi). \end{cases} \tag{2.10}$$

Similarly, we denote by $Q_\nu^\mu(\kappa)$ the linear combination of $\partial_s^{i_1}\kappa * * * \partial_s^{i_\nu}\kappa$, where $i_1 + \cdots + i_\nu = \mu$.

The following lemma states the important interpolation inequality for higher order curvature functionals.

LEMMA 2.3 (Proposition 2.5 in [4]). *For any term $P_\nu^\mu(\kappa)$ with $\nu \geq 2$ which contains only derivatives of κ of order at most $k-1$, we have*

$$\int_I \left|P_\nu^\mu(\kappa)\right| \, ds \leq c\, \mathcal{L}[f]^{1-\mu-\nu} \|\kappa\|_2^{\nu-\gamma} \|\kappa\|_{k,2}^{\gamma}\ , \tag{2.11}$$

where $\gamma = \left(\mu + \frac{1}{2}\nu - 1\right)/k$, $c = c(n,k,\mu,\nu)$, and

$$\|\kappa\|_{k,p} := \sum_{i=0}^{k} \left\|\nabla_s^i \kappa\right\|_p, \qquad \left\|\nabla_s^i\kappa\right\|_p := \mathcal{L}[f]^{i+1-\frac{1}{p}} \left(\int_I \left|\nabla_s^i\kappa\right|^p \, ds\right)^{\frac{1}{p}}.$$

Moreover, if $\mu + \frac{1}{2}\nu < 2k+1$, then $\gamma < 2$ and we have for any $\varepsilon > 0$,

$$\int_I \left|P_\nu^\mu(\kappa)\right| \, ds \leq \varepsilon \int_I |\nabla_s^k \kappa|^2 \, ds + c\varepsilon^{\frac{-\gamma}{2-\gamma}} \left(\int_I |\kappa|^2 \, ds\right)^{\frac{\nu-\gamma}{2-\gamma}} + c\left(\int_I |\kappa|^2 \, ds\right)^{\mu+\nu-1}. \tag{2.12}$$

LEMMA 2.4 (Lemma 2.6 in [4]). *We have the identities*

$$\nabla_s\kappa - \partial_s\kappa = |\kappa|^2 T, \tag{2.13}$$

$$\nabla_s^m\kappa - \partial_s^m\kappa = \sum_{i=1}^{[\frac{m}{2}]} Q_{2i+1}^{m-2i}(\kappa) + \sum_{i=1}^{[\frac{m+1}{2}]} Q_{2i}^{m+1-2i}(\kappa)\, T. \tag{2.14}$$

LEMMA 2.5 (Lemma 2.7 in [4]). *Assume the bounds $\|\kappa\|_{L^2} \leq \Lambda_0$ and $\|\nabla_s^m\kappa\|_{L^1} \leq \Lambda_m$ for $m \geq 1$. Then for any $m \geq 1$ one has*

$$\left\|\partial_s^{m-1}\kappa\right\|_{L^\infty} + \|\partial_s^m\kappa\|_{L^1} \leq c_m(\Lambda_0, \dots, \Lambda_m). \tag{2.15}$$

3. Proof of the main results. The formulas in the next lemma can be directly verified by applying the general formulas in Lemma 2.1. Thus, we skip the detail of the computation and leave the verification to the reader.

LEMMA 3.1. *Let $f : I = \mathbb{R}/\mathbb{Z} \to \mathbb{R}^3$ represent a smooth curve in $\mathbb{R}^3$ without inflection points. Then, for any variation $f_\varepsilon(x) = f(x) + \varepsilon W(x)$, where f, $W \in$*

$C^\infty(I)$, *one has the following:*

$$\begin{aligned}
\tfrac{d}{d\varepsilon}\big\lfloor_{\varepsilon=0}\mathcal{L}[f_\varepsilon] &= -\int_I \langle \kappa, W\rangle \, ds + [\langle T, W\rangle]_0^L,\\
\tfrac{d}{d\varepsilon}\big\lfloor_{\varepsilon=0}\mathcal{T}[f_\varepsilon] &= -\int_I \langle \nabla_s (T\times\kappa), W\rangle \, ds\\
&\quad + \Big[\langle \nabla_s^2 (W - \langle W, T\rangle T) + \langle W, T\rangle \cdot \nabla_s \kappa, \tfrac{B}{|\kappa|}\rangle + \langle W, T\times\kappa\rangle\Big]_0^L,\\
\tfrac{d}{d\varepsilon}\big\lfloor_{\varepsilon=0}\mathcal{K}[f_\varepsilon] &= \int_I \langle \nabla_s^2\kappa + \tfrac{|\kappa|^2}{2}\kappa, W\rangle \, ds\\
&\quad + \Big[\langle T, W\rangle \cdot \tfrac{|\kappa|^2}{2} + \langle \kappa, \nabla_s (W - \langle W, T\rangle T)\rangle - \langle \nabla_s\kappa, W\rangle\Big]_0^L.
\end{aligned}$$

Proof of Theorem 1.1. If we perturb the rod configuration Γ of a given elastic rod in equilibrium without perturbing the centerline f, then

$$0 = \delta\mathcal{E}[\Gamma] = \beta\cdot\delta\int_I m^2 \, ds = 2\beta\cdot\int_I m\cdot(\delta m) \, ds. \tag{3.1}$$

By the endpoint condition in (1.2) and the formula $m(s) = \theta'(s)$, we conclude that m is a constant and

$$m = \frac{(\triangle\Psi + \mathcal{T}[f])}{L} = L^{-1}\int_I m \, ds.$$

Thus, any closed Kirchhoff elastic rod in equilibrium with endpoint condition $\triangle\Psi$ and length L belongs to the subclass of rod configurations $\mathcal{A}_{\triangle\Psi,L}$, where

$$\mathcal{A}_{\triangle\Psi,L} := \left\{\Gamma : m(s) = \frac{\triangle\Psi + \mathcal{T}[f]}{L},\ \mathcal{L}[f] = L\right\}.$$

Observe that for any rod configurations $\Gamma\in\mathcal{A}_{\triangle\Psi,L}$, we have

$$\mathcal{E}[\Gamma] = \mathcal{G}_{\triangle\Psi,L}[f]. \tag{3.2}$$

Now, perturbations of Γ preserving the length in the subclass of rod configurations $\mathcal{A}_{\triangle\Psi,L}$ induce the variational equation

$$\delta\left(\mathcal{G}_{\triangle\Psi,L}[f] + \lambda_1\cdot(\mathcal{L}[f] - L)\right) = 0,$$

where λ_1 is the Lagrange multiplier.

Conversely, by assuming that f is the critical point of $\mathcal{F}$, we have

$$\delta_L\mathcal{G}_{\triangle\Psi,L}[f] = 0,$$

where δ_L denotes perturbations of preserving the length. The rod configuration Γ has constant twisting rate $\frac{(\triangle\Psi+\mathcal{T}[f])}{L}$; therefore $\Gamma\in\mathcal{A}_{\triangle\Psi,L}$. Thus,

$$\delta_L\mathcal{E}[\Gamma] = \delta_L\mathcal{G}_{\triangle\Psi,L}[f] = 0. \qquad \square$$

Proof of Theorem 1.2. The proof is motivated by the arguments in [4]. Recalling that no inflection point is on the initial curve, the short time existence is a standard

argument. We thus skip it here and focus on the long time existence and asymptotic behavior.

To prove global bounds we wish to estimate higher Sobolev norms of the curvature. Their evolution is given by

$$\nabla_t \nabla_s^m \kappa = -\nabla_s^4 \nabla_s^m \kappa + \text{ tensors of lesser order.}$$

Therefore we arrive at

$$\frac{d}{dt}\frac{1}{2}\int_I \mid \nabla_s^m \kappa \mid^2 \, ds + \int_I \mid \nabla_s^{m+2} \kappa \mid^2 \, ds = \text{ terms of lesser order.}$$

It will not be necessary to compute the error terms explicitly; it is sufficient to keep track of their scaling. In other words, we have to know the order of the derivatives involved. Using the notation introduced before, the next lemma characterizes the error terms coming from the twist term, i.e., dealing with the new situation that the total torsion is included in our energy.

LEMMA 3.2. *For $m \geq 2$, we have the formula*

$$\begin{aligned}
\nabla_s^m (T \times \kappa) \;=\; & T \times \nabla_s^m \kappa + \sum_{a_1,b_1,c_1,d_1} \left[P_{b_1}^{a_1}(\kappa) \times P_{d_1}^{c_1}(\kappa) \right]^{\perp} \\
& + \sum_{i=1,2} \sum_{a_2^{(i)},b_2^{(i)},c_2^{(i)},d_2^{(i)},e_2^{(i)},f_2^{(i)}} \left[\left(P_{b_2^i}^{a_2^i}(\kappa) \times P_{d_2^i}^{c_2^i}(\kappa) \right) * P_{f_2^i}^{e_2^i}(\kappa) \right]^{\perp} \\
& + \sum_{i=1,2} \sum_{a_3^{(i)},b_3^{(i)},c_3^{(i)},d_3^{(i)}} \left(\left(T \times P_{b_3^{(i)}}^{a_3^{(i)}}(\kappa) \right) * P_{d_3^{(i)}}^{c_3^{(i)}}(\kappa) \right),
\end{aligned}$$

where the sums are taken such that $(a_1 + c_1) + (b_1 + d_1)/2 = m$, $(a_2^{(i)} + c_2^{(i)} + e_2^{(i)}) + (b_2^{(i)} + d_2^{(i)} + f_2^{(i)})/2 = m - i$, and $(a_3^{(i)} + c_3^{(i)}) + (b_3^{(i)} + d_3^{(i)})/2 = m - i + 1/2$ for $i \in \{1, 2\}$.

Proof of Lemma 3.2. We first need the following formulas, which easily can be verified by applying (2.1). Assume $P_b^a(\phi)$ and $P_d^c(\phi)$ are normal vector fields; then

$$\text{(3.3)} \quad \begin{cases} \nabla_s \left(P_b^a(\phi) \times P_d^c(\phi) \right) = & [(\nabla_s P_b^a(\phi) + \langle \kappa, P_b^a(\phi) \rangle \cdot T) \times P_d^c(\phi) \\ & + P_b^a(\phi) \times (\nabla_s P_d^c(\phi) + \langle \kappa, P_d^c(\phi) \rangle \cdot T)]^{\perp}, \\ \nabla_s \left(T \times P_d^c(\phi) \right) = [\kappa \times P_d^c(\phi)]^{\perp} + T \times P_d^{c+1}(\phi), \end{cases}$$

where $[\cdots]^{\perp}$ denotes its normal component and $\times$ denotes the exterior product in $\mathbb{R}^3$. Notice that in (3.3), we use $+$ instead of $-$ for our convenience because the sign is meaningless as using universal constant coefficients in those terms, $P_\beta^\alpha(\phi)$.

Now the proof is an induction argument. As $m = 2$,

$$\nabla_s^2 (T \times \kappa) = T \times \nabla_s^2 \kappa + (\kappa \times \nabla_s \kappa)^{\perp} = T \times \nabla_s^2 \kappa + \left(P_1^0(\kappa) \times P_1^1(\kappa) \right)^{\perp}.$$

As $m \geq 3$, we apply (2.1), (2.10), and (3.3) in the following calculation:

$$\begin{aligned}
\nabla_s^m (T \times \kappa) \;&= \nabla_s \Bigg\{ T \times \nabla_s^{m-1} \kappa + \sum_{a_1,b_1,c_1,d_1} \left[P_{b_1}^{a_1}(\kappa) \times P_{d_1}^{c_1}(\kappa) \right]^{\perp} \\
&\quad + \sum_{i=1}^{2} \sum_{a_2^{(i)},b_2^{(i)},c_2^{(i)},d_2^{(i)},e_2^{(i)},f_2^{(i)}} \left[\left(P_{b_2^i}^{a_2^i}(\kappa) \times P_{d_2^i}^{c_2^i}(\kappa) \right) * P_{f_2^i}^{e_2^i}(\kappa) \right]^{\perp} \\
&\quad + \sum_{i=1}^{2} \sum_{a_3^{(i)},b_3^{(i)},c_3^{(i)},d_3^{(i)}} \left(\left(T \times P_{b_3^{(i)}}^{a_3^{(i)}}(\kappa) \right) * P_{d_3^{(i)}}^{c_3^{(i)}}(\kappa) \right) \Bigg\} \\
&= \nabla_s \Bigg\{ I_1 + I_2 + \sum_{i=1}^{2} I_3^{(i)} + \sum_{i=1}^{2} I_4^{(i)} \Bigg\}.
\end{aligned}$$

1.

$$\nabla_s I_1 = \nabla_s \left[T \times \nabla_s^{m-1}\kappa\right] = T \times \nabla_s^m \kappa + \left[P_1^0(\kappa) \times P_1^{m-1}(\kappa)\right]^{\perp}.$$

2.

$$\begin{aligned}
\nabla_s I_2 &= \sum_{a_1,b_1,c_1,d_1} \nabla_s \left[P_{b_1}^{a_1}(\kappa) \times P_{d_1}^{c_1}(\kappa)\right]^{\perp} \\
&= \sum_{a_1,b_1,c_1,d_1} \nabla_s [P_{b_1}^{a_1}(\kappa) \times P_{d_1}^{c_1}(\kappa)] - \langle P_{b_1}^{a_1}(\kappa) \times P_{d_1}^{c_1}(\kappa), T\rangle \cdot \kappa \\
&= \sum_{a_1,b_1,c_1,d_1} [((P_{b_1}^{a_1+1}(\kappa) + P_{b_1+1}^{a_1}(\kappa)\, T) \times P_{d_1}^{c_1}(\kappa)) \\
&\quad + (P_{b_1}^{a_1}(\kappa) \times \left(P_{d_1}^{c_1+1}(\kappa) + P_{b_1+1}^{a_1}(\kappa)\, T\right)]^{\perp} + \left(T \times P_{b_1}^{a_1}(\kappa)\right) * P_{d_1+1}^{c_1}(\kappa) \\
&= \sum_{a,b,c,d} [P_b^a(\kappa) \times P_d^c(\kappa)]^{\perp} + \sum_{A,B,C,D} \left(T \times P_B^A(\kappa)\right) * P_D^C(\kappa),
\end{aligned}$$

where $(a+c)+(b+d)/2 = (a_1+c_1)+(b_1+d_1)/2+1$ and $(A+C)+(B+D)/2 = (a_1+c_1)+(b_1+d_1)/2+1/2$.

3.

$$\begin{aligned}
\nabla_s I_3^{(i)} &= \sum_{a_2^{(i)},b_2^{(i)},c_2^{(i)},d_2^{(i)},e_2^{(i)},f_2^{(i)}} \nabla_s \left[\left(P_{b_2^i}^{a_2^i}(\kappa) \times P_{d_2^i}^{c_2^i}(\kappa)\right) * P_{f_2^i}^{e_2^i}(\kappa)\right]^{\perp} \\
&= \sum_{a_2^{(i)},b_2^{(i)},c_2^{(i)},d_2^{(i)},e_2^{(i)},f_2^{(i)}} \nabla_s [(P_{b_2^{(i)}}^{a_2^{(i)}}(\kappa) \times P_{d_2^{(i)}}^{c_2^{(i)}}(\kappa)) * P_{f_2^{(i)}}^{e_2^{(i)}}(\kappa)] \\
&\qquad - \langle (P_{b_2^{(i)}}^{a_2^{(i)}}(\kappa) \times P_{d_2^{(i)}}^{c_2^{(i)}}(\kappa)) * P_{f_2^{(i)}}^{e_2^{(i)}}(\kappa), T\rangle \cdot \kappa \\
&= \sum_{a_2^{(i)},b_2^{(i)},c_2^{(i)},d_2^{(i)},e_2^{(i)},f_2^{(i)}} \nabla_s [(P_{b_2^{(i)}}^{a_2^{(i)}}(\kappa) \times P_{d_2^{(i)}}^{c_2^{(i)}}(\kappa))] * P_{f_2^{(i)}}^{e_2^{(i)}}(\kappa) \\
&\qquad + (P_{b_2^{(i)}}^{a_2^{(i)}}(\kappa) \times P_{d_2^{(i)}}^{c_2^{(i)}}(\kappa)) * \nabla_s P_{f_2^{(i)}}^{e_2^{(i)}}(\kappa) \\
&\qquad - \langle (P_{b_2^{(i)}}^{a_2^{(i)}}(\kappa) \times P_{d_2^{(i)}}^{c_2^{(i)}}(\kappa)) * P_{f_2^{(i)}}^{e_2^{(i)}}(\kappa), T\rangle \cdot \kappa \\
&= \sum_{a,b,c,d,e,f} [(P_b^a(\kappa) \times P_d^c(\kappa)) * P_f^e(\kappa)]^{\perp} + \sum_{A,B,C,D} \left(T \times P_B^A(\kappa)\right) * P_D^C(\kappa),
\end{aligned}$$

where $(a+c+e)+(b+d+f)/2 = (a_2^{(i)}+c_2^{(i)}+e_2^{(i)})+(b_2^{(i)}+d_2^{(i)}+f_2^{(i)})/2+1$ and $(A+C)+(B+D)/2 = (a_2^{(i)}+c_2^{(i)})+(b_2^{(i)}+d_2^{(i)})/2+1/2$.

4.

$$\begin{aligned}
\nabla_s I_4^{(i)} &= \sum_{a_3^{(i)},b_3^{(i)},c_3^{(i)},d_3^{(i)}} \nabla_s [(T \times P_{b_3^{(i)}}^{a_3^{(i)}}(\kappa)) * P_{d_3^{(i)}}^{c_3^{(i)}}(\kappa)] \\
&= \sum_{a_3^{(i)},b_3^{(i)},c_3^{(i)},d_3^{(i)}} \{\partial_s [(T \times P_{b_3^{(i)}}^{a_3^{(i)}}(\kappa)) * P_{d_3^{(i)}}^{c_3^{(i)}}(\kappa)]\}^{\perp} \\
&= \sum_{a_3^{(i)},b_3^{(i)},c_3^{(i)},d_3^{(i)}} \{(P_1^0(\kappa) \times P_{b_3^{(i)}}^{a_3^{(i)}}(\kappa) + (T \times P_{b_3^{(i)}}^{a_3^{(i)}+1}(\kappa))) * P_{d_3^{(i)}}^{c_3^{(i)}}(\kappa)\}^{\perp} \\
&\qquad + (T \times P_{b_3^{(i)}}^{a_3^{(i)}}(\kappa)) * P_{d_3^{(i)}}^{c_3^{(i)}+1}(\kappa) \\
&= \sum_{a,b,c,d} [(P_b^a(\kappa) \times P_d^c(\kappa)) * P_f^e(\kappa)]^{\perp} + \sum_{A,B,C,D} \left(T \times P_B^A(\kappa)\right) * P_D^C(\kappa),
\end{aligned}$$

where $(a+c+e)+(b+d+f)/2=(a_3^{(i)}+c_3^{(i)})+(b_3^{(i)}+d_3^{(i)})/2+1/2$ and $(A+C)+(B+D)/2=(a_3^{(i)}+c_3^{(i)})+(b_3^{(i)}+d_3^{(i)})/2+1$.

The proof is finished by summing up all of these terms from part 1 through part 4. ∎

LEMMA 3.3 (corresponding to Lemma 2.3 in [4]). *Suppose*

$$\partial_t f = -\nabla_s^2\kappa+\sigma|\kappa|^2\kappa+\lambda_1\kappa+\lambda_2\nabla_s(T\times\kappa),$$

where σ, $\lambda_i\in\mathbb{R}$. Then, for $m\geq 0$, the derivatives of the curvature $\phi_m=\nabla_s^m\kappa$ satisfy

(3.4)
$$\begin{aligned}&\nabla_t\phi_m+\nabla_s^4\phi_m\\&=P_3^{m+2}(\kappa)+\sigma\cdot\left(P_3^{m+2}(\kappa)+P_5^m(\kappa)\right)+\lambda_1\cdot\left(\nabla_s^{m+2}\kappa+P_3^m(\kappa)\right)\\&\quad+\lambda_2\cdot\left(\nabla_s^{m+3}(T\times\kappa)+\nabla_s^{m+1}(T\times\kappa)*P_2^0(\kappa)+\cdots+\nabla_s^1(T\times\kappa)*P_2^m(\kappa)\right).\end{aligned}$$

The statement is still true when $\lambda_i=\lambda_i(t)$ depends on time.

Proof of Lemma 3.3. The case of $m=0$ follows from (2.6) and the definition of $\partial_t f$,

$$\begin{aligned}\nabla_t\kappa &= -\nabla_s^4\kappa+\sigma\cdot(\nabla_s^2(|\kappa|^2\kappa)+|\kappa|^4\kappa)+\lambda_1\cdot(\nabla_s^2\kappa+|\kappa|^2\kappa)\\&\quad+\lambda_2\cdot\left(\nabla_s^3(T\times\kappa)+\kappa\langle\kappa,\nabla_s(T\times\kappa)\rangle\right).\end{aligned}$$

For $m\geq 1$, (3.4) can be inductively derived by using (2.9),

$$\begin{aligned}&\nabla_t\phi_m+\nabla_s^4\phi_m\\&=\nabla_s[P_3^{m+1}(\kappa)+\sigma\cdot\left(P_3^{m+1}(\kappa)+P_5^{m-1}(\kappa)\right)+\lambda_1\cdot\left(\nabla_s^{m+1}\kappa+P_3^{m-1}(\kappa)\right)\\&\qquad+\lambda_2\cdot(\nabla_s^{m+2}(T\times\kappa)+\nabla_s^m(T\times\kappa)*P_2^0(\kappa)+\cdots+\nabla_s^1(T\times\kappa)*P_2^{m-1}(\kappa))]\\&\quad+\langle\kappa,\phi_{m-1}\rangle\cdot\nabla_s[-\nabla_s^2\kappa+\sigma|\kappa|^2\kappa+\lambda_1\kappa+\lambda_2\nabla_s(T\times\kappa)]\\&\quad-\langle\nabla_s[-\nabla_s^2\kappa+\sigma|\kappa|^2\kappa+\lambda_1\kappa+\lambda_2\nabla_s(T\times\kappa)],\phi_{m-1}\rangle\cdot\kappa\\&\quad+\langle\kappa,-\nabla_s^2\kappa+\sigma|\kappa|^2\kappa+\lambda_1\kappa+\lambda_2\nabla_s(T\times\kappa)\rangle\cdot\phi_m\\&=P_3^{m+2}(\kappa)+\sigma\cdot\left(P_3^{m+2}(\kappa)+P_5^m(\kappa)\right)+\lambda_1\cdot\left(\nabla_s^{m+2}\kappa+P_3^m(\kappa)\right)\\&\quad+\lambda_2\cdot(\nabla_s^{m+3}(T\times\kappa)+\nabla_s^{m+1}(T\times\kappa)*P_2^0(\kappa)+\cdots+\nabla_s^1(T\times\kappa)*P_2^m(\kappa)).\end{aligned}$$ ∎

By (2.8) and (3.4), we have

$$\begin{aligned}&\tfrac{d}{dt}\tfrac{1}{2}\int_I|\nabla_s^m\kappa|^2\,ds+\int_I|\nabla_s^{m+2}\kappa|^2\,ds+\lambda_1(t)\int_I|\nabla_s^{m+1}\kappa|^2\,ds\\&=\lambda_1(t)\int_I\langle\nabla_s^m\kappa,P_3^m(\kappa)\rangle\,ds+\int_I\langle\nabla_s^m\kappa,P_3^{m+2}(\kappa)+P_5^m(\kappa)\rangle\,ds\\&\quad+\lambda_2(t)\int_I\langle\nabla_s^m\kappa,\nabla_s^{m+3}(T\times\kappa)+\nabla_s^{m+1}(T\times\kappa)*P_2^0(\kappa)\\&\qquad+\cdots+\nabla_s^1(T\times\kappa)*P_2^m(\kappa)\rangle\,ds.\end{aligned}\tag{3.5}$$

Notice that estimating terms in (3.5) is the key argument of this paper. One can verify from Lemma 3.1 that

$$\begin{aligned}\tfrac{d}{dt}\mathcal{F}[f_t] &= 2\alpha\tfrac{d}{dt}\mathcal{K}[f]+\tfrac{2\beta}{L}(\mathcal{T}[f]+\triangle\Psi)\tfrac{d}{dt}\mathcal{T}[f]+\lambda_1(t)\cdot\tfrac{d}{dt}\mathcal{L}[f]\\&=\int_I\langle 2\alpha(\nabla_s^2\kappa+\frac{|\kappa|^2}{2}\kappa)-\lambda_2(t)\nabla_s(T\times\kappa)-\lambda_1(t)\kappa,\partial_t f\rangle\,ds\\&=-\int_I|2\alpha(-\nabla_s^2\kappa-\frac{|\kappa|^2}{2}\kappa)+\lambda_2(t)\nabla_s(T\times\kappa)+\lambda_1(t)\kappa|^2\,ds\\&\leq 0.\end{aligned}\tag{3.6}$$

Note that $\lambda_1(t)$ is chosen to fulfill $\mathcal{L}[f_t] \equiv L$. From (3.6), $\mathcal{F}[f_t]$ is nonincreasing as t is increasing. Thus,

$$\begin{aligned}\tfrac{\beta}{L}\left(\mathcal{T}[f_t]+\triangle\Psi\right)^2 &\le 2\alpha\mathcal{K}[f_t]+\tfrac{\beta}{L}\left(\mathcal{T}[f_t]+\triangle\Psi\right)^2\\ &= \mathcal{G}_{\triangle\Psi,L}[f_t]=\mathcal{F}[f_t]\le\mathcal{F}[f_0]\\ &= 2\alpha\mathcal{K}[f_0]+\tfrac{\beta}{L}\left(\mathcal{T}[f_0]+\triangle\Psi\right)^2.\end{aligned}$$

Therefore,

$$|\lambda_2(t)| = \frac{2\beta}{L}\left|\mathcal{T}[f]+\triangle\Psi\right| \le C\left(f_0,\triangle\Psi,\alpha,\beta,L\right) \tag{3.7}$$

is uniformly bounded. Furthermore, by (3.6),

$$\|\kappa\|_{L^2}^2 = 2\mathcal{K}[f_t] \le C\left(f_0,\alpha\right). \tag{3.8}$$

Thus $\|\kappa\|_{L^2}^2$ is uniformly bounded for any $t \ge 0$.

By applying (2.12), (3.7), (3.8), and Lemma 3.2, the sum of the last two terms in (3.5) satisfies the inequality

$$\begin{aligned}&\int_I \left\langle \nabla_s^m\kappa, P_3^{m+2}(\kappa)+P_5^m(\kappa)\right\rangle ds\\ &+\lambda_2(t)\int_I \left\langle \nabla_s^m\kappa, \nabla_s^{m+3}(T\times\kappa)+\nabla_s^{m+1}(T\times\kappa)*P_2^0(\kappa)\right.\\ &\qquad\left.+\cdots+\nabla_s^1(T\times\kappa)*P_2^m(\kappa)\right\rangle ds\\ &\le C\left(f_0,\triangle\Psi,\alpha,\beta,L\right)\cdot\left(\varepsilon\int_I\left|\nabla_s^{m+2}\kappa\right|^2 ds+C\left(f_0,m,\varepsilon\right)\right).\end{aligned} \tag{3.9}$$

Now we estimate the term involving $\lambda_1(t)$ on the right-hand side of (3.5). Since $\kappa=\partial_s^2 f$, by applying the Poincaré inequality to $\partial_s f$, we have the estimate

$$L\|\kappa\|_{L^2}^2 \ge 4\pi^2.$$

Thus, by applying (2.11) to the right-hand side of (1.6) involving $\lambda_1(t)$, we have the estimates

$$\begin{aligned}&|\lambda_1(t)|\\ &\le C\left(f_0,\triangle\Psi,\alpha,\beta,L\right)\cdot\int_I\left(\left|P_2^2(\kappa)\right|+\left|P_4^0(\kappa)\right|+\left|P_2^1(\kappa)\right|\right)ds\\ &\le C\cdot\left(\|\kappa\|_{m+2,2}^{\frac{2}{m+2}}\cdot\|\kappa\|_2^{2-\frac{2}{m+2}}+\|\kappa\|_{m+2,2}^{\frac{1}{m+2}}\cdot\|\kappa\|_2^{4-\frac{2}{m+2}}+\|\kappa\|_{m+2,2}^{\frac{1}{m+2}}\cdot\|\kappa\|_2^{2-\frac{1}{m+2}}\right)\end{aligned}$$

and

$$\left|\int_I\left\langle\nabla_s^m\kappa,P_3^m(\kappa)\right\rangle ds\right| \le \int_I\left|P_4^{2m}(\kappa)\right| ds \le c(m,L)\cdot\|\kappa\|_{m+2,2}^{2-\frac{3}{m+2}}\cdot\|\kappa\|_2^{2+\frac{3}{m+2}}.$$

Therefore,

$$\begin{aligned}&\left|\lambda_1(t)\int_I\left\langle\nabla_s^m\kappa,P_3^m(\kappa)\right\rangle ds\right|\\ &\le C\left(f_0,\triangle\Psi,\alpha,\beta,L,m\right)\cdot\left(\|\kappa\|_{m+2,2}^{2-\frac{1}{m+2}}+\|\kappa\|_{m+2,2}^{2-\frac{2}{m+2}}\right)\\ &\le \varepsilon\int_I\left|\nabla_s^{m+2}\kappa\right|^2 ds+C\left(f_0,\triangle\Psi,\alpha,\beta,L,m,\varepsilon\right),\end{aligned} \tag{3.10}$$

where the last inequality comes from applying Young's inequality and the inequality

$$\|\kappa\|_{k,2}^2 \leq c\,(k)\left(\left\|\nabla_s^k\kappa\right\|_2^2 + \|\kappa\|_2^2\right)$$

(can be yielded by a standard interpolation inequality; see [2]).

The remaining term in (3.5) to be estimated is $\lambda_1\,(t)\cdot\int_I \left|\nabla_s^{m+1}\kappa\right|^2\,ds$, which is the borderline case as applying the above estimates. In other words, the interpolation technique fails now. Instead, we use the observation that the total torsion is invariant under the rescaling; therefore, the rescaling argument in [4] still works. More precisely, it can be verified that as we rescale f by $f^{(\rho)} = p + \rho\,(f-p)$, we have the properties $\mathcal{K}\left[f^{(\rho)}\right] = \frac{1}{\rho}\mathcal{K}\,[f]$, $\mathcal{T}\left[f^{(\rho)}\right] = \mathcal{T}\,[f]$ and $\mathcal{L}\left[f^{(\rho)}\right] = \rho\mathcal{L}\,[f]$. Taking the derivative of $\mathcal{F}\left[f^{(\rho)}\right]$ at $\rho = 1$ and using (1.5), we have

$$2\alpha\mathcal{K}\,[f] - \lambda_1\mathcal{L}\,[f] = -\frac{d}{d\rho}\mathcal{F}\left[f^{(\rho)}\right]\lfloor_{\rho=1} = \int_I \left\langle \partial_t f, f-p\right\rangle\,ds.$$

Thus, as long as $p = p\,(t)$ is properly chosen, e.g., $p = L^{-1}\int_I f\,ds$, and by the energy identity

$$\frac{d}{dt}\mathcal{F}\,[f_t] = -\int_I \left|\partial_t f\right|^2\,ds, \tag{3.11}$$

one has the inequality

$$-\lambda_1\,(t) \leq L^{1/2}\left\|\partial_t f\right\|_{L^2},$$

which implies the estimate

$$\int_0^t \left(\lambda_1^-\,(\tau)\right)^2\,d\tau \leq C\,(f_0, \triangle\Psi, \alpha, \beta, L),$$

where $\lambda_1^-\,(t) = -\min\{0, \lambda_1\,(t)\}$. By applying integration by parts and the Hölder inequality, we have

$$-\lambda_1\int_I \left|\nabla_s^{m+1}\kappa\right|^2\,ds \leq \varepsilon\cdot\int_I \left|\nabla_s^{m+2}\kappa\right|^2\,ds + c\,(\varepsilon)\cdot\left(\lambda_1^-\right)^2\cdot\int_I \left|\nabla_s^m\kappa\right|^2\,ds. \tag{3.12}$$

Note that by applying the Poincaré inequality twice, we have

$$\int_I \left|\nabla_s^{m+2}\kappa\right|^2\,ds \geq \left(\frac{2\pi}{L}\right)^4\int_I \left|\nabla_s^m\kappa\right|^2\,ds. \tag{3.13}$$

Now, by (3.5), (3.9), (3.10), (3.12), (3.13), and a small enough number $\varepsilon = \varepsilon\,(f_0, \triangle\Psi, \alpha, \beta, L, m) > 0$, we have

$$\frac{d}{dt}\int_I \left|\nabla_s^m\kappa\right|^2\,ds + C_1\cdot\int_I \left|\nabla_s^m\kappa\right|^2\,ds \leq C_2\cdot\left(1 + \left(\lambda_1^-\,(t)^2\right)\cdot\int_I \left|\nabla_s^m\kappa\right|^2\,ds\right), \tag{3.14}$$

where we let $C_i = C_i\,(f_0, \triangle\Psi, \alpha, \beta, L, m) > 0$ for all $i \in \mathbb{Z}$, from now on. Let

$$u_m\,(t) := \exp\,(C_1\cdot t)\cdot\int_I \left|\nabla_s^m\kappa\right|^2\,ds.$$

By applying the Gronwall inequality to (3.14), we have

$$u_m(t) \le e^{a(t)} \cdot \left(u_m(0) + C_2 \cdot \int_0^t e^{C_1 \cdot \tau}\, d\tau \right),$$

where

$$a(t) = \int_0^t C_2 \cdot (\lambda_1^-(\tau))^2\, d\tau \le C(f_0, \triangle\Psi, \alpha, \beta, L, m).$$

Therefore, we obtain

$$\begin{aligned} \|\nabla_s^m \kappa\|_{L^2}^2(t) &\le C(f_0, \triangle\Psi, \alpha, \beta, L, m) \cdot (1 + e^{-C_1 \cdot t} \cdot \|\nabla_s^m \kappa\|_{L^2}^2(0)) \\ &\le C(f_0, \triangle\Psi, \alpha, \beta, L, m) \end{aligned} \tag{3.15}$$

for all $m \ge 0$. In addition, from the definition of λ_1 in (1.6), we conclude that $|\lambda_1| \le C(f_0, \triangle\Psi, \alpha, \beta, L)$. Notice that one has the estimate

$$\left\|\partial_s^{m-1}\kappa\right\|_{L^\infty} \le c \cdot \|\partial_s^m \kappa\|_{L^1} \quad \forall\ m \ge 1. \tag{3.16}$$

Now, by applying the induction argument on m, and using Lemmas 2.4 and 2.5, (3.15), (3.16), and the Hölder inequality, we derive the inequalities

$$\|\nabla_s^m \kappa\|_{L^\infty} + \|\partial_s^m \kappa\|_{L^\infty} \le C(f_0, \triangle\Psi, \alpha, \beta, L, m)\ \forall\ m \ge 0. \tag{3.17}$$

On the asymptotic behavior of the flow, we choose a subsequence of curves $f(t,\cdot)$ which converges smoothly to a curve f_∞, after reparametrizations of arclength and translations. Lemma 3.3 and (3.17) imply

$$\|\nabla_t(\nabla_s^m \kappa)\|_{L^\infty} \le C(f_0, \triangle\Psi, \alpha, \beta, L, m)\ \forall\ m \ge 0. \tag{3.18}$$

From (3.17) and (3.18), one sees that for $u(t) := \int_I |\partial_t f|^2\, ds$, the inequality

$$|u'(t)| \le C(f_0, \triangle\Psi, \alpha, \beta, L)$$

holds. On the other hand, the energy identity, (3.11), implies $u(t) \in L^1([0,\infty))$. Therefore, $u(t) \to 0$ as $t \to \infty$. In other words, f_∞ is independent of t and thus, by (1.5), is an equilibrium of $\mathcal{F}$. Now, by Theorem 1.1, the proof is finished. ∎

4. Numerical algorithm. We base our numerical treatment on the algorithm proposed in [4] and implement the new nonlinear term $\lambda_2 \nabla_s (T \times \kappa)$ explicitly in time.

First, observe that the divergence form of the main part in the evolution equation admits a weak formulation of the flow. In fact, we have

$$\nabla_s^2 \kappa + \tfrac{1}{2}|\kappa|^2 \kappa - \lambda_2 \nabla_s (T \times \kappa) = \partial_s \left(\partial_s \kappa + \tfrac{3}{2}|\kappa|^2 T - \lambda_2 T \times \kappa \right).$$

Second, the common way of avoiding higher order elements for the discretization is to rewrite the equation as a second-order system for position vector f and the mean curvature vector κ:

$$\partial_t f + \partial_s \left(\partial_s \kappa + \tfrac{3}{2}|\kappa|^2 T - \lambda_2 T \times \kappa \right) = \lambda_1 \kappa, \tag{4.1}$$

$$\partial_s^2 f = \kappa. \tag{4.2}$$

The weak form of the problem leads in one space dimension to a difference scheme. Decompose $I = \mathbb{R}/\mathbb{Z} = \cup_1^N I_j$ into intervals $I_j = [x_{j-1}, x_j)$, where x_j are the nodal points. We discretize the space $H^1(I, \mathbb{R}^3)$ by the space

$$X_h = \left\{ g \in C^0(I, \mathbb{R}^3) : g|_{I_j} \in_1 (I_j) \right\} = (\operatorname{span}\{\phi_1, \dots, \phi_N\})^n$$

of periodic piecewise affine functions spanned by the nodal basis functions $\phi_j \in X_h$ satisfying $\phi_j(x_i) = \delta_{ij}$. The discretization parameter is given by $h = \max_j h_j$, $h_j = |I_j|$. We use the pointwise interpolation $I_h g$, $g \in C^0(I, \mathbb{R}^3)$ uniquely defined by $I_h g \in X_h$ and $I_h g(x_j) = g(x_j)$ for all $j = 1, \dots, N$. A discrete (weak) solution to (4.1) is then a pair of functions $(f_h, \kappa_h) : [0, T] \to X_h \times X_h$,

$$f_h(x,t) = \sum_{j=1}^N f_j(t)\phi_j(x), \quad \kappa_h(x,t) = \sum_{j=1}^N \kappa_j(t)\phi_j(x)$$

satisfying for all $\phi_h, \psi_h \in X_h$ the weak problem

$$\begin{aligned} &\int_I \left(I_h(\partial_t f_h \phi_h)|\partial_x f_h| - \frac{\partial_x \kappa_h}{|\partial_x f_h|}\partial_x \phi_h - \frac{3}{2}|\kappa_h|^2 \frac{\partial_x f_h}{|\partial_x f_h|}\partial_x \phi_h \right) dx \\ &= \int_I \left(\lambda_1 \frac{\partial_x f_h}{|\partial_x f_h|} \times \kappa_h \partial_x \phi_h + \lambda_2 I_h(\kappa_h \phi_h)|\partial_x f_h| \right) dx = 0, \end{aligned} \tag{4.3}$$

$$-\int_I \frac{\partial_x f_h}{|\partial_x f_h|}\partial_x \psi_h dx = \int_I I_h(\kappa_h \psi_h)|\partial_x f_h| dx. \tag{4.4}$$

In the time direction we discretize semi-implicitly. In particular, the new nonlinear term $\lambda_2 \nabla_s (T \times \kappa)$ in our flow equation is treated explicitly. For functions defined on the time interval $[0, T]$ we use the notation $g^m = g(\cdot, mk)$, $kM = T$.

Algorithm. For given initial data $f_0(x)$ and nodal points of the parameterization x_j, $j = 1, \dots, N$ let $f_j^0 = f_0(x_j)$, $h_j^0 = |f_j^0 - f_{j-1}^0|$, and

$$\kappa_j^0 = \frac{2}{h_{j+1}^0(h_j^0 + h_{j+1}^0)} f_{j+1}^0 - \frac{2}{h_j^0 h_{j+1}^0} f_j^0 + \frac{2}{h_j^0(h_j^0 + h_{j+1}^0)} f_{j-1}^0,$$

where we use the extensions $f_0^0 = f_N^0$, $f_{N+1}^0 = f_1^0$, $h_0^0 = h_N^0$, $h_{N+1}^0 = h_1^0$.

For $m = 0, \dots, M-1$ we set

$$\begin{aligned} h_j^m &= |f_j^m - f_{j-1}^m|, \\ \beta_j^m &= |\kappa_{j-1}^m|^2 + \kappa_{j-1}^m \kappa_j^m + |\kappa_j^m|^2, \\ \gamma_j^m &= \frac{f_j^m - f_{j-1}^m}{h_j^m} \times \frac{\kappa_j^m + \kappa_{j-1}^m}{2}, \end{aligned}$$

and solve for $f_j^{m+1}, \kappa_j^{m+1}$ in

$$\begin{aligned}
\frac{\beta_j^m}{2h_j^m} f_{j-1}^{m+1} + \left(\frac{h_j^m + h_{j+1}^m}{2k} - \frac{\beta_j^m}{2h_j^m} - \frac{\beta_{j+1}^m}{2h_{j+1}^m} \right) f_j^{m+1} + \frac{\beta_{j+1}^m}{2h_{j+1}^m} f_{j+1}^{m+1} \\
+ \frac{1}{h_j^m} \kappa_{j-1}^{m+1} - \left(\frac{1}{h_j^m} + \frac{1}{h_{j+1}^m} + \frac{\lambda_1^m}{2}(h_j^m + h_{j+1}^m) + \lambda_2^m * \right) \kappa_j^{m+1} + \frac{1}{h_{j+1}^m} \kappa_{j+1}^{m+1} \\
= \frac{h_j^m + h_{j+1}^m}{2k} f_j^m + \lambda_2^m \left(\gamma_{j+1}^m - \gamma_j^m \right), \\
\frac{1}{h_j^m} f_{j-1}^{m+1} - \left(\frac{1}{h_j^m} + \frac{1}{h_{j+1}^m} \right) f_j^{m+1} + \frac{1}{h_{j+1}^m} f_{j+1}^{m+1} = \frac{h_j^m + h_{j+1}^m}{2} \kappa_j^m.
\end{aligned}$$

Here, the Lagrange multipliers are computed according to

$$\begin{aligned}
\lambda_2^m &= \frac{2\beta}{L^m}(\tau^m + \Delta\Psi), \\
\lambda_1^m &= -\frac{\sum_{j=1}^N \left(|\kappa_j^m - \kappa_{j-1}^m|^2 / h_j^m + (f_j^m - f_{j-1}^m) \cdot (\kappa_j^m - \kappa_{j-1}^m) \beta_j^m / 2h_j^m + \lambda_2^m \Gamma_j^m \right)}{\sum_{j=1}^N h_j^m \beta_j^m / 3},
\end{aligned}$$

where

$$\begin{aligned}
\Gamma_j^m &= \kappa_j^m \cdot (\gamma_{j+1}^m - \gamma_j^m), \\
\tau^m &= -3 \sum_{j=1}^N \Gamma_j^m / \beta_j^m, \\
L^m &= \sum_{j=1}^N h_j^m.
\end{aligned}$$

The algorithm is intrinsic in the sense that it does not explicitly depend on the grid parameter $h = \max_j h_j$. Nevertheless, during time evolution the distribution of nodes drift away from the equidistant grid. Thus, we redistribute the nodes tangentially according to arclength if the ratio $\max_j h_j / \min_j h_j$ exceeds 2.

We also mention that the linear system for $f_j^{m+1}, \kappa_j^{m+1}$ can be decoupled, giving a linear system for f_j^{m+1} alone. The tridiagonal structure of the matrices is then replaced by a five-diagonal structure, where the periodicity of the curve implies nonzero elements in the upper right and lower left corners. The implementation of a fifth-diagonal linear solver can easily be generalized to such a situation.

Computations and figures. Let us first note that numerical computations show that the flat circle is a stationary solution which continues to stay stable for small values of β. For increasing values of β the circle loses stability, and we observe nontrivial equilibria of nonzero total torsion.

Figure 4.1 shows a table of stationary states for given values of $\Delta\Psi$ in the interval $(-\pi, \pi)$. Observe that equilibria corresponding to the same absolute value of $\Delta\Psi$ bend into the opposite direction leading to a reflection symmetry w.r.t. the horizontal plane.

It is interesting to see that the issue of inflection points comes into play if $|\Delta\Psi|$ approaches the value π. Then the corresponding asymptotic stationary curve contains

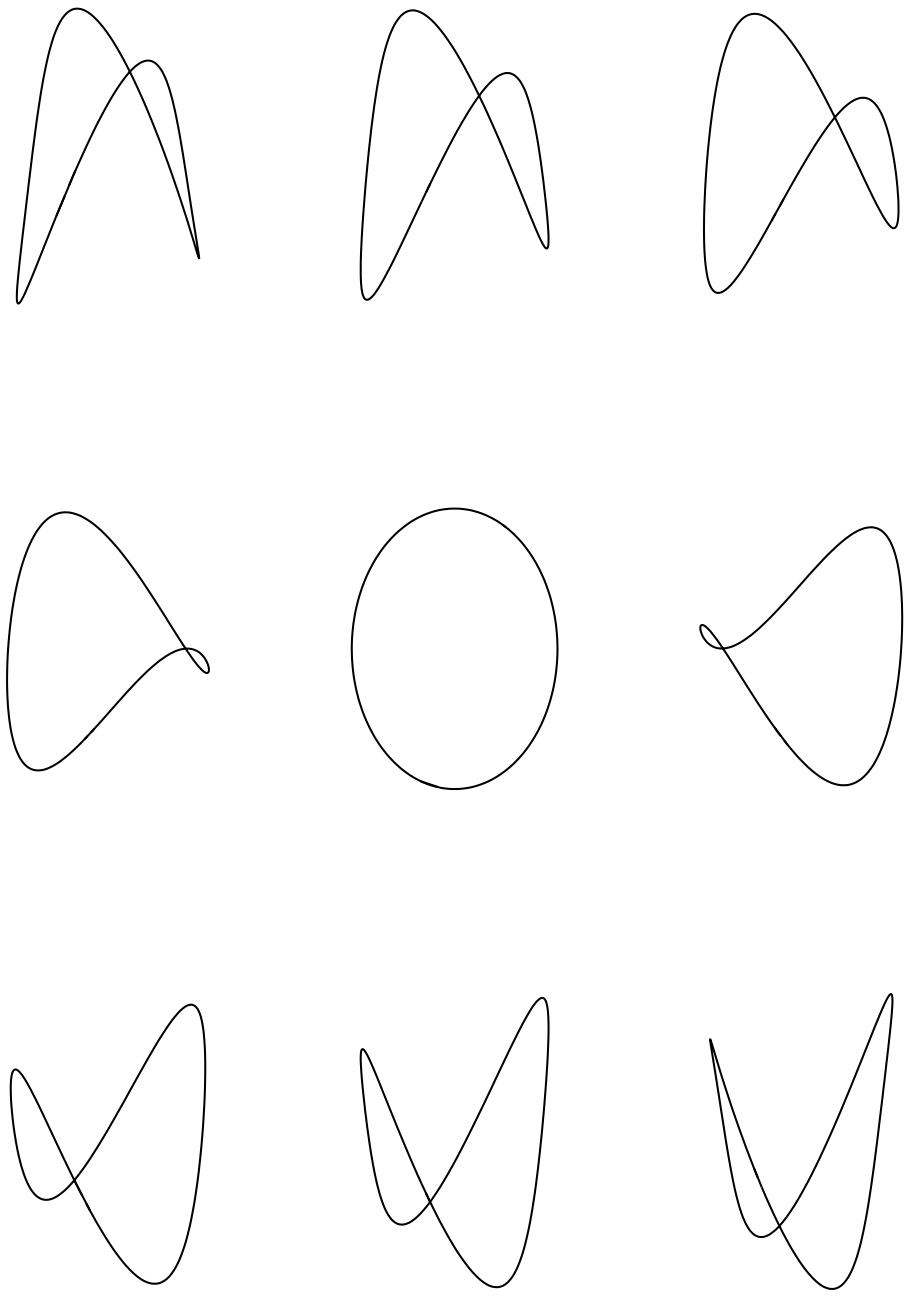


FIG. 4.1. *Stationary curves of length* 6.31 *for* $\beta = 35$ *and* $\Delta\Psi = -3.1, -2, -1, -0.5, 0, 0.5, 1, 2, 3.1$.

points having a very small magnitude of the curvature vector κ. We mentioned before that the flow equation and the computation of the torsion τ gets ill-defined in such a situation. We observe this problem also in our computations in the sense that the flow gets numerically unstable if curves with points of small $|\kappa|$ evolve.

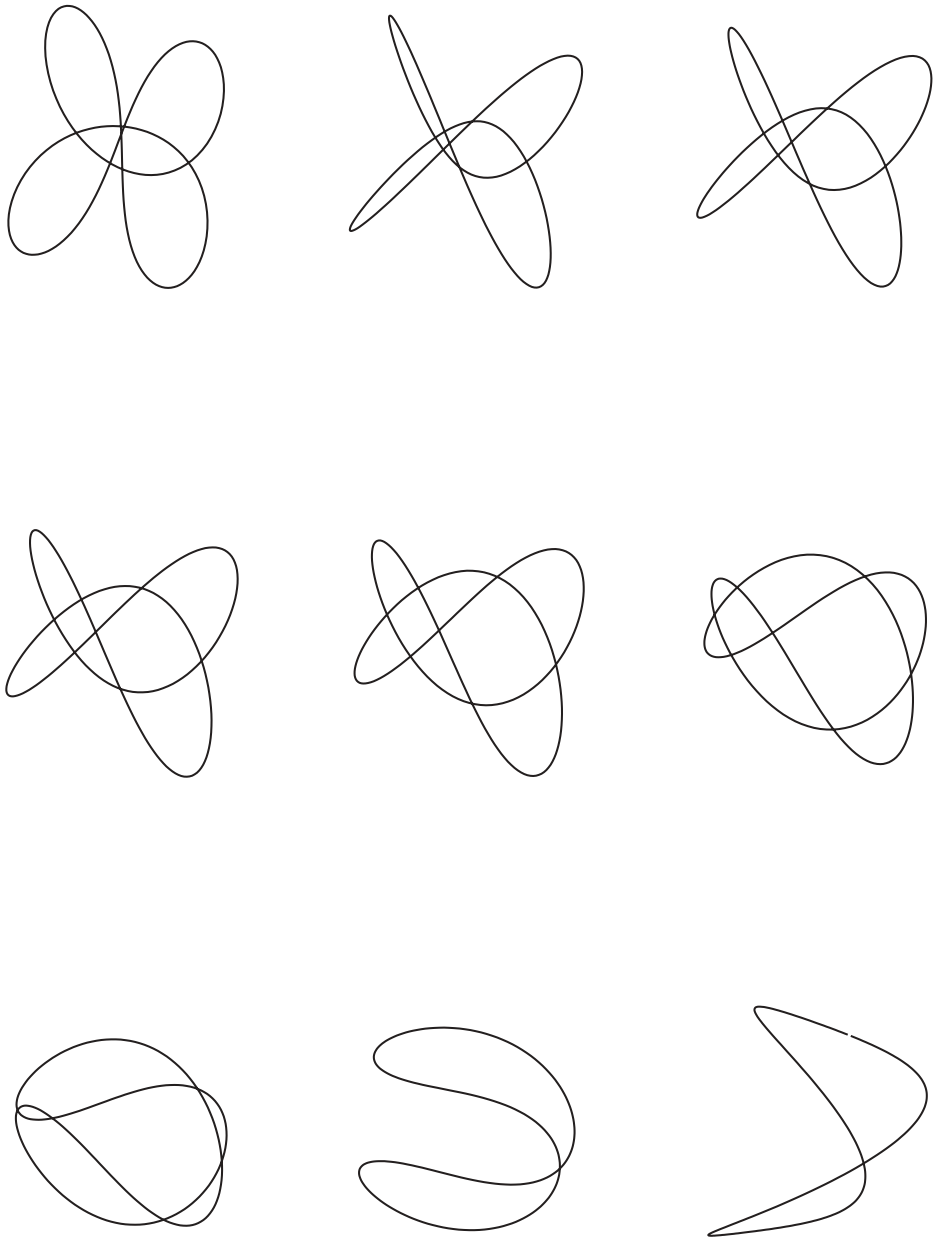


FIG. 4.2. *Evolution of a curve of length* 18.9 *under the flow with* $\beta = 60$, $\Delta\Psi = -2$ *at times* $t = 0, 1, 2, 3, 4, 6, 8, 13, 52$.

The next table in Figure 4.2 presents the evolution of a strongly bent initial curve unfolding along our flow to a stationary curve. Recalling the different lengths, the curve is similar (but not identical) to the one from Figure 4.1.

Acknowledgments. C. C. L. sincerely thanks Stefan Müller and Angela Stevens for encouragement and Dirk Drasdo for many interesting discussions on biophysics.

REFERENCES

[1] S. S. Antman, *Nonlinear Problems of Elasticity*, Appl. Math. Sci. 107, Springer-Verlag, New York, 1995.
[2] T. Aubin, *Nonlinear Analysis on Manifolds. Monge-Ampère Equations*, Grundlehren Math. Wiss. 252, Springer-Verlag, New York, 1982.
[3] R. Bishop, *There is more than one way to frame a curve*, Amer. Math. Monthly, 82 (1975), pp. 118–133.
[4] G. Dziuk, E. Kuwert, and R. Schätzle, *Evolution of elastic curves in $\mathbb{R}^n$: Existence and computation*, SIAM J. Math. Anal., 33 (2002), pp. 1228–1245.
[5] T. Ivey and D. Singer, *Knot types, homotopies and stability of closed elastic rods*, Proc. London Math. Soc. (3), 79 (1999), pp. 429–450.
[6] N. Koiso, *On the motion of a curve towards elastica*, in Actes de la Table Ronde de Géométrie Différentielle (Luminy, 1992), Sémin. Congr. 1, Soc. Math. France, Paris, 1996, pp. 403–436.
[7] J. Langer and D. Singer, *Lagrangian aspects of the Kirchhoff elastic rod*, SIAM Rev., 38 (1996), pp. 605–618.
[8] A. Polden, *Curves and Surfaces of Least Total Curvature and Fourth-Order Flows*, Ph.D. dissertation, Universität Tübingen, Tübingen, Germany, 1996.
[9] Y. Wen, *Curve straightening flow deforms closed plane curves with nonzero rotation number to circles*, J. Differential Equations, 120 (1995), pp. 89–107.

SIAM J. APPL. MATH.
Vol. 65, No. 3, pp. 737–753

BIFURCATIONS OF A RATIO-DEPENDENT PREDATOR-PREY SYSTEM WITH CONSTANT RATE HARVESTING*

DONGMEI XIAO† AND LESLIE STEPHEN JENNINGS‡

Abstract. The ratio-dependent predator-prey model exhibits rich interesting dynamics due to the singularity of the origin. The objective of this paper is to study the dynamical properties of the ratio-dependent predator-prey model with nonzero constant rate harvesting. For this model, the origin is not an equilibrium. It is shown that numerous kinds of bifurcation occur for the model, such as the saddle-node bifurcation, the subcritical and supercritical Hopf bifurcations, Bogdanov–Takens bifurcation, the homoclinic bifurcation, and the heteroclinic bifurcation, as the values of parameters of the model vary. Hence, there are different parameter values for which the model has a limit cycle, or a homoclinic loop, or a heteroclinic orbit, or a separatrix connecting a saddle and a saddle-node. These results reveal far richer dynamics compared to the model with no harvesting.

Key words. ratio-dependent predator-prey system, constant rate harvesting, bifurcation, extinction, coexistence

AMS subject classifications. Primary, 34C25, 92A15; Secondary, 58F14

DOI. 10.1137/S0036139903428719

1. Introduction.

In population dynamics, both ecologists and mathematicians are interested in the following Michaelis–Menten-type predator-prey model, the so-called ratio-dependent predator-prey model:

$$\begin{aligned} \dot{x} &= rx\left(1-\frac{x}{K}\right)-\frac{cxy}{my+x}, \\ \dot{y} &= y\left(-D+\frac{fx}{my+x}\right), \end{aligned} \tag{1.1}$$

where $x(t)$ and $y(t)$ represent population densities of prey and predator at time t, respectively. r, K, c, m, D, and f are positive constants. The prey grows with intrinsic growth rate r and carrying capacity K in the absence of predation. D, c, m, and f stand for the predator death rate, capturing rate, half saturation constant, and conversion rate, respectively. The reason for the model is that numerous field and laboratory experiments and observations (Abrams and Ginzburg [1], Arditi and Berryman [3], Arditi and Ginzburg [4], Arditi, Ginzburg, and Akcakaya [5], Akcakaya, Arditi, and Ginzburg [2], Cosner et al. [14], and Gutierrez [20]) showed that functional and numerical responses over typical ecological timescales ought to depend on the densities of both prey and predators, especially when predators must search for food (and therefore must share or compete for food). The suitable functional response is a *ratio-dependent* response function, which, roughly, the per capita predator growth rate should be a function of the ratio of prey to predator abundance. For more biological background concerning the model, the reader can refer to the above references

*Received by the editors May 25, 2003; accepted for publication (in revised form) March 20, 2004; published electronically February 25, 2005.
http://www.siam.org/journals/siap/65-3/42871.html
†Department of Mathematics, Shanghai Jiao Tong University, Shanghai 200030, China (xiaodm@sjtu.edu.cn). The research of this author was supported by the National Natural Science Foundations of China grant 10231020.
‡Centre for Applied Dynamics and Optimization, The University of Western Australia, Nedlands, Perth, WA6907 Australia (les@maths.uwa.edu.au).

and the extensive references cited. The dynamics of system (1.1) has been studied extensively (Berezovskaya, Karev, and Arditi [7], Hsu, Hwang, and Kuang [21], Jost, Arino, and Arditi [22], Kuang [23], Kuang and Beretta [24], Xiao and Ruan [29], and references therein). Research on the ratio-dependent predator-prey model (1.1) revealed rich interesting dynamics such as deterministic extinction, existence of multiple attractors, and existence of a stable limit cycle. It was shown in [7], [21], and [29] that system (1.1) has very complicated dynamics close to the origin: there exist numerous kinds of topological structures in a neighborhood of the origin, including parabolic orbits, elliptic orbits, hyperbolic orbits, and any combination thereof, depending on the different values of parameters. Thus, the origin is a degenerate equilibrium with high codimension. Mathematically, from the point of view of bifurcation, it is a very interesting question what kinds of bifurcation will occur when system (1.1) is perturbed by a small constant term. From the perspective of biology, the complicated dynamics clearly indicates that for the ratio-dependent model, even if there is a positive equilibrium, both prey and predator can still go extinct for some values of parameters, and the extinction occurs in two distinct ways. One of the ways is that both species become extinct regardless of the initial densities such as the Gause's classic observation of mutual extinction. The other way is that both species will die out only if the initial prey/predator ratio is too low. Some researchers regard such interesting dynamics as "pathological behavior" and hope to remove it so as to guarantee the persistence of the system. From the point of view of human needs, the exploitation of biological resources and the harvesting of populations are commonly practiced in fishery, forestry, and wildlife management. There is a wide range of interest in the use of bioeconomic models to gain insight into the scientific management of renewable resources like fisheries and forestries. It is related to the optimal management of renewable resources (an excellent introduction to optimal management of renewable resources is given by Clark in [13]). Generally speaking, it is necessary to consider the harvesting of populations in some models. Taking into consideration the above two-fold reasons, we focus on the ratio-dependent predator-prey model with constant harvesting. For mathematical simplicity, let us first nondimensionalize system (1.1) as in Kuang and Beretta [24] with the following scaling:

$$t \to rt, \quad x \to x/K, \quad y \to my/K.$$

Then system (1.1) takes the form

$$\begin{aligned} \dot{x} &= x(1-x) - \frac{axy}{y+x}, \\ \dot{y} &= y\left(-d + \frac{bx}{y+x}\right), \end{aligned} \tag{1.2}$$

where $a = \frac{c}{mr}$, $b = \frac{f}{r}$, and $d = \frac{D}{r}$ are positive constants. For simplicity, in the following, we keep the biological implications of parameters a, b, and d the same as c, f, and D, respectively.

In this paper, we assume that the predator in the model (1.2) is not of commercial importance. The prey is continuously being harvested at a constant rate by a harvesting agency. The harvesting activity does not affect the predator population directly. It is obvious that the harvesting activity does reduce the predator population indirectly by reducing the availability of the prey to the predator.

We formulate the problem as:

$$\begin{aligned} \dot{x} &= x(1-x) - \frac{axy}{y+x} - h, \\ \dot{y} &= y\left(-d + \frac{bx}{y+x}\right), \end{aligned} \tag{1.3}$$

where h represents the rate of harvesting or removal; hence, $h > 0$.

The objective of this paper is to systematically study the dynamical properties of the ratio-dependent predator-prey model with constant harvesting. It will be better for us to determine how the constant harvesting affects the dynamics of system (1.3). The occurrence of change of structure, or bifurcation, in a system with parameters is a major way to predict global dynamics of the system. By making local calculations, we give bifurcation analysis for system (1.3) and show that system (1.3) can exhibit numerous kinds of bifurcation phenomena in terms of the original parameters in the model, including the bifurcation of cusp type of codimension 2 (i.e., Bogdanov–Takens bifurcation), the heteroclinic bifurcation, and the separatrix connecting a saddle-node and a saddle bifurcation. However, the ratio-dependent model (1.2) cannot undergo these bifurcations. From the point of view of the optimal management of renewable resources, the aim is to determine how much we can harvest without altering dangerously the harvested population. According to Clark in [13], the management of renewable resources has been based on the maximum sustainable yield (MSY), with the property that any larger harvest rate will lead to the depletion of the population (eventually to zero). If x is harvested by some process of overexploitation (i.e., $h > h_{MSY}$), then the prey species can be led to extinction. The most rapid recovery of the population is achieved by means of a moratorium on harvesting, i.e., $h = 0$. In this paper, qualitative and bifurcation analyses are combined to determine $h_{MSY} = \frac{1}{4}$ for the model (1.3). Biologically, when $h \geq \frac{1}{4}$, overharvesting of prey species occurs for model (1.3), which may lead to the collapse of the whole system. Hence, $0 < h < \frac{1}{4}$ is of interest for model (1.3). Another noteworthy prediction from model (1.3) is that prey and predator species cannot become extinct simultaneously (mutual extinction) for all values of parameters and initial values. This, however, contradicts the observation of mutual extinction for the ratio-dependent model (1.2). Thus, prey harvesting prevents mutual extinction as a possible outcome of predator-prey interaction and removes the "pathological behavior" of model (1.2).

This paper is organized as follows. In section 2, we study the existence of the equilibria and various types of dynamical behavior in the small neighborhood of the equilibrium for the model (1.3). Some proofs of various types of dynamics are placed in the appendices. We also describe the phase portraits and the biological ramifications of our results. In section 3 we consider all possible bifurcations of the model depending on all parameters. We show that the model exhibits the saddle-node bifurcation, the subcritical and supercritical Hopf bifurcations, the Bogdanov–Takens bifurcation of codimension 2, the separatrix connecting a saddle and a saddle-node bifurcation, and the heteroclinic bifurcation in terms of the original parameters. The paper ends with a brief discussion.

2. General phase portraits analysis of equilibria. In this section, we consider the ratio-dependent predator-prey system with a constant rate harvesting:

$$\begin{aligned} \dot{x} &= x(1-x) - \frac{axy}{y+x} - h \triangleq f_1(x,y), \\ \dot{y} &= y\left(-d + \frac{bx}{y+x}\right) \triangleq f_2(x,y), \end{aligned} \tag{2.1}$$

where a, h, d, and b are positive parameters. From the standpoint of biology, we are interested only in the dynamics of system (2.1) in the closed first quadrant R_+^2. Thus, we consider only the biologically meaningful initial condition

$$x(0) \geq 0, \quad y(0) \geq 0.$$

Straightforward computation shows that $f_1(x,y)$ and $f_2(x,y)$ are continuous and Lipschizian in the closed first quadrant R_+^2 if we let $f_1(0,0) = -h$, $f_2(0,0) = 0$. Hence, solution of (2.1) with nonnegative initial condition exists and is unique. It is also easy to see that the x-axis is invariant under the flow. However, this is not the case on the y-axis. All solutions touching the y-axis cross out of the first quadrant, and the origin $(0,0)$ is not an equilibrium of system (2.1). Thus, the first quadrant is no longer positively invariant under the flow generated by system (2.1), which makes the qualitative analysis of system (2.1) difficult.

First, let us begin to determine the location and number of the equilibria of system (2.1) in the first quadrant R_+^2. From system (2.1), we can see that there exists an equilibrium of system (2.1) in R_+^2 if and only if the equations

$$\begin{aligned} x(1-x) - \frac{axy}{y+x} - h &= 0, \\ y\left(-d + \frac{bx}{y+x}\right) &= 0 \end{aligned} \tag{2.2}$$

have a pair of nonnegative real solutions (x,y). It is clear that equations (2.2) have at most four pairs of nonnegative real solutions (x_i, y_i) and (x_i^*, y_i^*),

$$x_i = \frac{1 + (-1)^i\sqrt{1-4h}}{2}, \qquad y_i = 0;$$

$$x_i^* = \frac{b - a(b-d) + (-1)^i\sqrt{\Delta}}{2b}, \qquad y_i^* = \frac{b-d}{d}x_i^*,$$

where $i = 1, 2$, $\Delta = (a(b-d) - b)^2 - 4hb^2$. Therefore, we have the following simple lemma which describes the number and location of equilibria of system (2.1). The proof is omitted.

LEMMA 2.1.

(1) *System* (2.1) *has no equilibria in* R_+^2 *if* $h > \frac{1}{4}$.

(2) *System* (2.1) *has a unique equilibrium in* R_+^2, *which is* $(x_0, y_0) = (\frac{1}{2}, 0)$ *if* $h = \frac{1}{4}$.

(3) *System* (2.1) *has two equilibria in* R_+^2, *which are* (x_1, y_1) *and* (x_2, y_2) *if one of the following conditions holds:*

(3.i) $b \leq d$ *and* $0 < h < \frac{1}{4}$;

(3.ii) $b - d \geq \frac{b}{a}$ *and* $0 < h < \frac{1}{4}$;

(3.iii) $0 < b - d < \frac{b}{a}$ *and* $(\frac{a(b-d)-b}{2b})^2 < h < \frac{1}{4}$.

(4) *System* (2.1) *has three equilibria in* R^2_+, *which are* $(x^*, y^*) = (\frac{b-a(b-d)}{2b}, \frac{b-d}{d}x^*)$, (x_1, y_1), *and* (x_2, y_2) *if* $0 < b - d < \frac{b}{a}$ *and* $(\frac{a(b-d)-b}{2b})^2 = h < \frac{1}{4}$.

(5) *System* (2.1) *has four equilibria in* R^2_+, *which are* (x_1^*, y_1^*), (x_2^*, y_2^*), (x_1, y_1), *and* (x_2, y_2) *if* $0 < b - d < \frac{b}{a}$ *and* $0 < h < \min\{(\frac{a(b-d)-b}{2b})^2, \frac{1}{4}\}$.

When $h > \frac{1}{4}$, system (2.1) has no equilibria, and $\dot{x}(t) < 0$ in R^2_+. The dynamics of system (2.1) in R^2_+ is trivial and all orbits in R^2_+ will cross the y-axis and go out of R^2_+ in finite time. This implies that the prey species goes extinct, which in turn causes the extinction of the predator. Biological overharvesting occurs. When $0 < h \leq \frac{1}{4}$, system (2.1) has some equilibria, and there exist some initial values such that the population of prey in system (2.1) does not go extinct. Thus, $h_{MSY} = \frac{1}{4}$ for system (2.1) from Lemma 2.1.

Next we consider the dynamics of system (2.1) in the neighborhood of each equilibrium. The linear part of system (2.1) at these equilibria is determined by the matrix

$$Df(x, y) = \begin{pmatrix} 1 - 2x - \frac{ay^2}{(x+y)^2} & -\frac{ax^2}{(x+y)^2} \\ \frac{by^2}{(x+y)^2} & -d + \frac{bx^2}{(x+y)^2} \end{pmatrix},$$

where x and y are coordinates of these equilibria, respectively. More precisely, we have

$$Df(x_i, y_i) = \begin{pmatrix} (-1)^{i+1}\sqrt{1-4h} & -a \\ 0 & b - d \end{pmatrix},$$

and

$$Df(x_i^*, y_i^*) = \begin{pmatrix} \frac{a(b-d)d+(-1)^{i+1}\sqrt{\Delta}}{b^2} & -\frac{ad^2}{b^2} \\ \frac{(b-d)^2}{b} & \frac{d(b-d)}{b} \end{pmatrix}.$$

The dynamics of system (2.1) in the neighborhood of an equilibrium comes directly from the property of eigenvalues of the matrix $Df(x, y)$ at the equilibrium.

When $h = \frac{1}{4}$, system (2.1) has a unique equilibrium $(x_0, y_0) = (\frac{1}{2}, 0)$. The linear part of system (2.1) at (x_0, y_0) is determined by the matrix

$$Df(x_0, y_0) = \begin{pmatrix} 0 & -a \\ 0 & b - d \end{pmatrix}.$$

Using classical qualitative methods, it is straightforward to show that the equilibrium (x_0, y_0) is nonhyperbolic, i.e., it is degenerate. More precisely, we have the following theorem. The proof is placed in Appendix A.

THEOREM 2.2. *When* $h = \frac{1}{4}$, *system* (2.1) *has only a unique equilibrium* $(x_0, y_0) = (\frac{1}{2}, 0)$, *and*

(i) (x_0, y_0) *is a saddle-node of codimension* 1 *if* $b \neq d$;

(ii) (x_0, y_0) *is a degenerate saddle-node of codimension* 4 *if* $b = d$ *and* $2a - 5b \neq 0$.

Their phase portraits are shown in Figure 2.1.

Notice that the codimension of a degenerate equilibrium of vector field $v(x)$ is the codimension of the bifurcation at the degenerate equilibrium, which is the minimum number of the parameters necessary for vector field $v(x, \lambda)$ with parameter perturbation to be a universal unfolding of $v(x)$. However, it is very difficult to determine a universal unfolding of $v(x)$. Therefore, the following view is usually used to determine codimension of the bifurcation: suppose that S is the set of all structurally stable

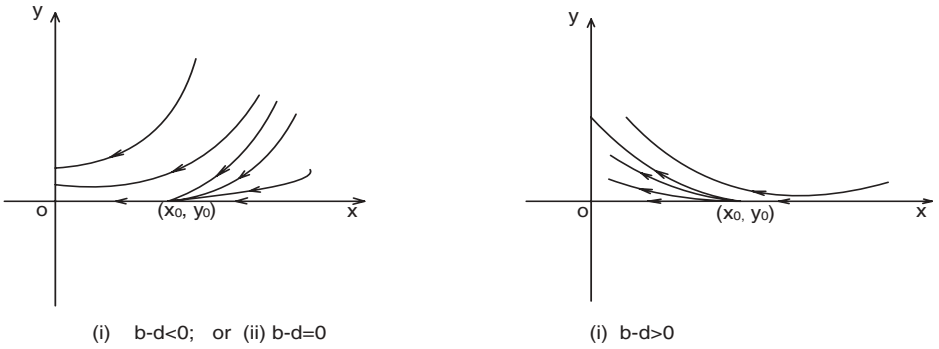


FIG. 2.1. *The phase portrait of system* (2.1) *with one equilibrium.*

vector fields in $C^1(E)$ (the set of all differential vector fields on E), and vector field $v(x)$ is in the complement of S. Then $v(x)$ belongs to the bifurcation set in $C^1(E)$ that is locally isomorphic to a manifold M in vector space or Banach space and the codimension of the bifurcation that occurs at $v(x)$ is equal to the codimension of the manifold M (cf. [19], [26]). The codimension in Theorem 2.3 is the codimension of the manifold.

In Theorem 2.2, $\dot{x} = -(x - \frac{1}{2})^2 - \frac{axy}{x+y} < 0$ in R^2_+, which implies that the prey species may go extinct as time increases for some initial values. And when $b \leq d$, predator death rate d is greater than the predator conversion rate b (i.e., the conversion efficiency of prey to predator), $\dot{y} < 0$. Hence, the predator species goes extinct as time increases for some initial values. But both cannot go to extinction simultaneously.

For sustainable development of resource, we assume that the harvesting agency has an obligation to the society and ecology of preserving the prey species. Hence, the harvesting rate h must satisfy $0 < h < \frac{1}{4}$. From Lemma 2.1 and using routine qualitative analysis, we have the following.

THEOREM 2.3. *If system* (2.1) *has only two equilibria, then the dynamics of system* (2.1) *is trivial in* R^2_+*, with no limit cycles in* R^2_+*. Each orbit of system* (2.1) *in* R^2_+ *goes to either one equilibrium on the* x*-axis or out of* R^2_+*. More precisely,*

(a) *if* $b < d$ *and* $0 < h < \frac{1}{4}$*, then system* (2.1) *has two equilibria* $(x_1, 0)$*, a hyperbolic saddle, and* $(x_2, 0)$*, a hyperbolic stable node;*

(b) *if* $b = d$ *and* $0 < h < \frac{1}{4}$*, then system* (2.1) *has two equilibria* $(x_1, 0)$ *and* $(x_2, 0)$*, both saddle-nodes;*

(c) *if* $0 < b - d < \frac{b}{a}$ *and* $(\frac{a(b-d)-b}{2b})^2 < h < \frac{1}{4}$ *(or* $b - d > \frac{b}{a}$ *and* $0 < h < \frac{1}{4}$*), then system* (2.1) *has two equilibria* $(x_1, 0)$*, a hyperbolic unstable node, and* $(x_2, 0)$*, a hyperbolic saddle.*

Their phase portraits are shown in Figure 2.2.

Biologically, system (2.1) is not persistent under the conditions of Theorem 2.3. The predator species goes extinct for some initial data or the prey species goes extinct for other initial data in R^2_+. Hence, for the persistence of the ecosystem, the equilibrium of the greatest interest would be an equilibrium interior to the first quadrant.

THEOREM 2.4. *Let* $h_0 = (\frac{a(b-d)-b}{2b})^2$*. If* $0 < b - d < \frac{b}{a}$ *and* $h = h_0 < \frac{1}{4}$*, then system* (2.1) *has three equilibria and no closed orbits in* R^2_+*. Moreover,*

(I) *if* $a \neq b$*, then system* (2.1) *has three equilibria in* R^2_+*, which are the saddle-node* $(x^*, y^*) = (\frac{b-a(b-d)}{2b}, \frac{b-d}{d}x^*) = (\sqrt{h_0}, \frac{b-d}{d}\sqrt{h_0})$*, the hyperbolic*

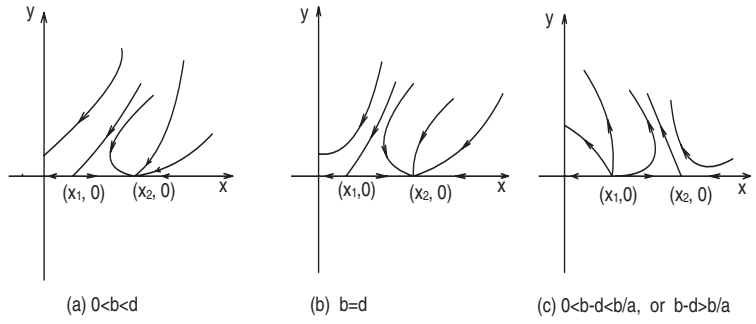


FIG. 2.2. *The phase portraits of system* (2.1) *with two equilibria.*

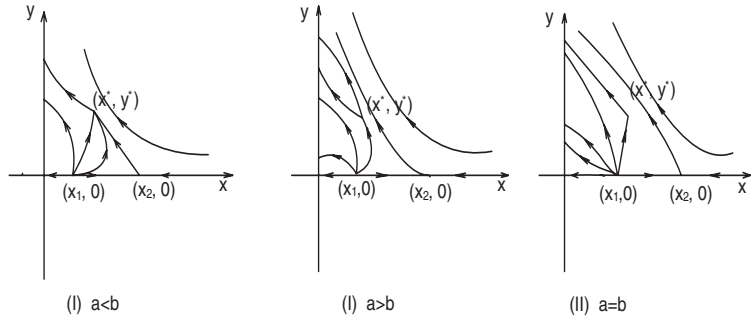


FIG. 2.3. *The phase portraits of system* (2.1) *with three equilibria.*

unstable node $(x_1, y_1) = (\frac{1-\sqrt{1-4h_0}}{2}, 0)$*, and the hyperbolic saddle* $(x_2, y_2) = (\frac{1+\sqrt{1-4h_0}}{2}, 0)$*;*

(II) *if* $a = b$*, then system* (2.1) *has three equilibria in* R^2_+*, which are the cusp* $(x^*, y^*) = (\sqrt{h_0}, \frac{b-d}{d}\sqrt{h_0})$ *of codimension* 2*, the hyperbolic unstable node* $(x_1, y_1) = (\frac{1-\sqrt{1-4h_0}}{2}, 0)$*, and the hyperbolic saddle* $(x_2, y_2) = (\frac{1+\sqrt{1-4h_0}}{2}, 0)$*.*

Their phase portraits are shown in Figure 2.3.

The proof of Theorem 2.4 is given in Appendix B. The outcome indicates that the predator species increases for all nonzero initial populations of predator except y^* when harvesting rate $h = h_0$ and the predator death rate is less than the predator conversion rate (i.e., $b > d$); however, the coexistence of predator species and prey species is possible only for some initial data and the condition that the predator capturing rate a is less than the predator conversion rate b (i.e., $a < b$). If $a > b$, the prey species goes extinct for almost all initial data in R^2_+.

If $0 < b - d < \frac{b}{a}$ and $0 < h < \min\{h_0, \frac{1}{4}\}$, then system (2.1) has two positive equilibria from Lemma 2.1. When system (2.1) has two positive equilibria, the global dynamics of system (2.1) is pretty complicated and many kinds of bifurcations will occur. We leave these to be discussed in the next section. Now we give the local dynamics of system (2.1) at the equilibria (x_1^*, y_1^*) and (x_2^*, y_2^*).

THEOREM 2.5. *If* $0 < b - d < \frac{b}{a}$ *and* $0 < h < \min\{h_0, \frac{1}{4}\}$*, then system* (2.1) *has four equilibria in* R^2_+ *as shown in Lemma* 2.1*. Moreover, the equilibria* $(x_2, 0)$ *and* (x_1^*, y_1^*) *are hyperbolic saddles, the equilibrium* $(x_1, 0)$ *is a hyperbolic unstable node, and the equilibrium* (x_2^*, y_2^*) *has the following three possibilities:*

(i) (x_2^*, y_2^*) *is a hyperbolic stable focus (or node) if either*

$$0 < b - d \le 1 < \frac{b}{a}$$

or, letting $h^* = \frac{(a(b-d)-b)^2 - (d(b-d)(a-b)/b)^2}{4b^2}$,

$$0 < b - d < \frac{b}{a} < 1, \qquad 0 < h < \min\left\{h_0,\ \frac{1}{4},\ h^*\right\};$$

(ii) (x_2^*, y_2^*) *is a weak focus or center if*

$$0 < b - d < \frac{b}{a} < 1, \qquad 0 < h = h^* < \min\left\{h_0,\ \frac{1}{4}\right\};$$

(iii) (x_2^*, y_2^*) *is a hyperbolic unstable focus (or node) if*

$$0 < b - d < \frac{b}{a} < 1, \qquad 0 < h^* < h < \min\left\{h_0,\ \frac{1}{4}\right\}.$$

3. The bifurcations of the model (2.1). In this section, we investigate the bifurcations that take place in system (2.1).

3.1. Saddle-node bifurcations. From Lemma 2.1 and Theorem 2.3, we have that

$$SN_1 = \left\{(a, b, d, h) :\ h = \frac{1}{4},\ b - d \neq 0,\ a > 0,\ b > 0,\ d > 0\right\}$$

is a *saddle-node bifurcation surface*. When the parameters pass from one side of the surface to the other side, the number of equilibria of system (2.1) changes from zero to two, and the two equilibria which are boundary equilibria are the hyperbolic saddle and node. This is the first saddle-node bifurcation surface of the model. The biological interpretation for the first saddle-node bifurcation is that $h_{MSY} = \frac{1}{4}$, the prey species is driven to extinction, and the system collapses for $h > \frac{1}{4}$, but the prey species do not go to extinction for some initial data when $0 < h < \frac{1}{4}$. On the other hand, from Theorems 2.4 and 2.5, we know that the surface

$$SN_2 = \left\{(a, b, d, h) :\ 0 < b - d < \frac{b}{a},\ h = h_0 < \frac{1}{4},\ a > 0,\ b > 0,\ d > 0\right\}$$

is a *saddle-node bifurcation surface*, which is the second saddle-node bifurcation the model undergoes. The saddle-node bifurcation yields two positive equilibria. This implies that there exists a critical harvest rate h_0 such that the predator species goes either extinct or out of R_+^2 in finite time when the harvest rate h is greater than h_0, and coexistence for model (2.1) is certain in the form of a positive equilibrium for certain choices of initial values when $0 < h < h_0$ and $0 < b - d < \frac{b}{a}$.

3.2. Hopf bifurcations. From the term (ii) of Theorem 2.5, we know that the positive equilibrium (x_2^*, y_2^*) of system (2.1) is a center-type nonhyperbolic equilibrium when $0 < b - d < \frac{b}{a} < 1$, $0 < h = h^* < \min\{h_0, \frac{1}{4}\}$. Hence, system (2.1) may undergo Hopf bifurcation. To determine the stability of the equilibrium and direction of Hopf bifurcation in this case, we must compute the Liapunov coefficients of the equilibrium.

We first translate the positive equilibrium (x_2^*, y_2^*) of system (2.1) to the origin. Then, system (2.1) in a neighborhood U of the origin can be written as

$$\begin{aligned}\dot{x} &= a_{10}x + a_{01}y + a_{20}x^2 + a_{11}xy + a_{02}y^2 \\ &\quad + a_{30}x^3 + a_{21}x^2y + a_{12}xy^2 + a_{03}y^3 + O_1(|(x,y)|^4), \\ \dot{y} &= b_{10}x + b_{01}y + b_{20}x^2 + b_{11}xy + b_{02}y^2 \\ &\quad + b_{30}x^3 + b_{21}x^2y + b_{12}xy^2 + b_{03}y^3 + O_2(|(x,y)|^4),\end{aligned} \tag{3.1}$$

where a_{ij} and b_{ij} are the coefficients of the power series expansions of $f_1(x,y)$ and $f_2(x,y)$ at (x_2^*, y_2^*) in U, respectively, $i,j = 0,1,2,3$. $O_k(|(x,y)|^4)$ is the same order infinity, $k = 1,2$. Hence, using the formula of the first Liapunov number σ for the focus at the origin of (3.1) in [26, p. 344], we have, after a tedious computation using Mathematica,

$$\sigma = \frac{3\pi d(b-d)Q}{4ab^9x_2^2\sqrt{\Delta_1^3}},$$

where $\Delta_1 = \frac{(a-b)(b-d)^2d^2}{b^3}$,

$$\begin{aligned}Q &= a^5(b-d)^8 + 2b^5(b-d)^3d^3 + 2a^4b(b-d)^5(-2b^2 + b(3+b)d - 2(1+b)d^2 + d^3) \\ &\quad + 2a^2b^3(b-d)(b^4(-2+d) + 18bd^4 - 9d^5 + 3b^3d(1+d) - b^2d^2(2+15d)) \\ &\quad + a^3b^2(b-d)^3(b(-2+d)d^2(-4+3d) + b^2d(-10 + (10-3d)d)) \\ &\quad + b^3(6 + (-4+d)d) - d^3(6 + (-4+d)d)) \\ &\quad + ab^4(44b^3d^3 - 79b^2d^4 + 62bd^5 - 18d^6 + b^4(1-9d^2)).\end{aligned}$$

By numerical calculation, we know there exist parameter values $(a,b,d) = (1, 0.5, 0.25)$, which satisfies the conditions of the term (ii) of Theorem 2.5, such that $\sigma = 465.961$. On the other hand, there exist parameter values $(a,b,d) = (4, 2, 1.85)$, which also satisfies the condition of the term (ii) of Theorem 2.5, such that $\sigma = -788.639$. Therefore, there exists an open set V_1 in the parameter space (a,b,d), such that $\sigma > 0$ and $0 < b-d < \frac{b}{a} < 1$, $0 < h^* < \min\{h_0, \frac{1}{4}\}$, i.e.,

$$V_1 = \left\{(a,b,d) : \ 0 < b-d < \frac{b}{a} < 1,\ 0 < h^* < \min\left\{h_0,\ \frac{1}{4}\right\},\ \text{and } \sigma > 0\right\}.$$

And there exists another open set V_2 in the parameter space (a,b,d), such that $\sigma < 0$ and $0 < b-d < \frac{b}{a} < 1$, $0 < h^* < \min\{h_0, \frac{1}{4}\}$, i.e.,

$$V_2 = \left\{(a,b,d) : \ 0 < b-d < \frac{b}{a} < 1,\ 0 < h^* < \min\left\{h_0,\ \frac{1}{4}\right\},\ \text{and } \sigma < 0\right\}.$$

This implies the following.

THEOREM 3.1.

(a) *If $h = h^*$, and if the parameter (a,b,d) is in V_1, then the equilibrium $(0,0)$ of system* (3.1) *is a weak focus of multiplicity* 1 and is unstable.

(b) If $h = h^*$, and if the parameter (a,b,d) is in V_2, then the equilibrium $(0,0)$ of system (3.1) is a weak focus of multiplicity 1 and is stable.

From Theorems 2.5 and 3.1, we know that the equilibrium (x_2^*, y_2^*) is a hyperbolic unstable focus if $(a,b,d) \in V_1$ and $h \geq h^*$, but the equilibrium (x_2^*, y_2^*) is a hyperbolic

stable focus if $(a,b,d) \in V_1$ and $0 < h < h^*$. Hence, when parameters pass from one side of the following surface to the other side, system (2.1) can undergo a subcritical Hopf bifurcation. An unstable limit cycle appears in the small neighborhood of (x_2^*, y_2^*) when $(a,b,d) \in V_1$ and $0 < h < h^*$. Thus, for some initial values, both species coexist for model (2.1) in the form of a positive equilibrium (x_2^*, y_2^*), and for other initial values, both species coexist for model (2.1) in the form of an oscillatory solution, which is unstable, when $(a,b,d) \in V_1$ and $0 < h < h^*$. The surface

$$H_b = \{(a,b,d,h) : \ h = h^*, \ (a,b,d) \in V_1\}$$

is called the *subcritical Hopf bifurcation surface* of system (2.1).

On the other hand, the equilibrium (x_2^*, y_2^*) is a hyperbolic stable focus if $(a,b,d) \in V_2$ and $0 < h \le h^*$, and the equilibrium (x_2^*, y_2^*) is a hyperbolic unstable focus if $(a,b,d) \in V_2$ and $h > h^*$. Hence, when parameters pass from one side of the following surface to the other side, system (2.1) can undergo a supercritical Hopf bifurcation. A stable limit cycle appears in the small neighborhood of (x_2^*, y_2^*) when $(a,b,d) \in V_2$ and $h > h^*$. Hence, the predator and the prey coexist for model (2.1) in the form of an oscillatory solution, which is stable, for some initial values when $(a,b,d) \in V_2$ and $h > h^*$. The surface

$$H_p = \{(a,b,d,h) : \ h = h^*, \ (a,b,d) \in V_2\}$$

is called the *supercritical Hopf bifurcation surface* of system (2.1).

Summarizing the above, we have the following.

THEOREM 3.2.

(i) *System* (2.1) *has at least one unstable limit cycle if* $0 < h < h^*$ *and* $(a,b,d) \in V_1$.

(ii) *System* (2.1) *has at least one stable limit cycle if* $h^* < h < \frac{1}{4}$ *and* $(a,b,d) \in V_2$.

Remark 3.1. Since there exist some parameter values such that $\sigma = 0$, system (2.1) maybe undergo *degenerate Hopf bifurcation* for some parameter values (cf. [6], [11], [12], and [26]).

3.3. The cusp bifurcation of codimension 2 (i.e., the Bogdanov–Takens bifurcation). From the item (II) of Theorem 2.4, we know that system (2.1) has a cusp (x^*, y^*) of codimension 2 when $h = h_0$, $a = b$, and $0 < b - d < \frac{b}{a}$. We now discuss if there exist the original parameters in system (2.1) such that system (2.1) exhibits Bogdanov–Takens bifurcation. We will show that a and h can be chosen as bifurcation parameters and system (2.1) can exhibit Bogdanov–Takens bifurcation.

Consider the system

$$\begin{aligned} \dot{u} &= u(1-u) - \frac{(b+\lambda_1)uv}{1+v} - (h_0 + \lambda_2), \\ \dot{v} &= v\left(-d + \frac{b}{1+v}\right) - v(1-u) + \frac{(b+\lambda_1)v^2}{1+v} + \frac{(h_0+\lambda_2)v}{u}, \end{aligned} \tag{3.2}$$

where λ_1 and λ_2 are very small parameters. When $\lambda_1 = \lambda_2 = 0$, system (3.2) has one positive equilibrium (u^*, v^*), which is a cusp of codimension 2.

Let $u_1 = u - u^*$, $v_1 = v - v^*$. Then system (3.2) becomes

$$\begin{aligned}\dot{u}_1 &= -\lambda_2 - \frac{\lambda_1 u^* v^*}{1+v^*} - \frac{(b+\lambda_1)u^* d^2}{b^2} v_1 - u_1^2 - \frac{(b+\lambda_1)d^2}{b^2} u_1 v_1 \\ &\quad + \frac{(b+\lambda_1)u^* d^3}{b^3} v_1^2 + O_1((u_1, v_1)^3), \\ \dot{v}_1 &= \frac{\lambda_1 v^{*2}}{1+v^*} + \frac{\lambda_2 v^*}{u^*} + \frac{d\lambda_1(b-d)}{b^2} v_1 + \frac{(b-d)}{du^*} u_1^2 + \frac{\lambda_1 d^3}{b^3} v_1^2 + O_2((u_1, v_1)^3).\end{aligned} \tag{3.3}$$

Next, we reduce system (3.3) to the normal form in successive steps. These steps are reminiscent of those performed in [28]. For simplicity, we omit the laborious steps and write down the normal form directly

$$\begin{aligned}\dot{u}_2 &= v_2, \\ \dot{v}_2 &= \mu_1(\lambda_1, \lambda_2) + \mu_2(\lambda_1, \lambda_2) v_2 + \frac{u^*(b-d)d^7}{2b^4} u_2^2 + \frac{2u^* d^2}{b} u_2 v_2 + O((\lambda, u_1, v_1)^3),\end{aligned} \tag{3.4}$$

where

$$\begin{aligned}\mu_1(\lambda_1, \lambda_2) &= \left(\frac{v^{*2}}{1+v^*} - \frac{d^4(b-d)}{4b^3} + \frac{d^9}{16b^8} + \frac{d^4 v^{*2}}{2(1+v^*)b^2} \right) \lambda_1 \\ &\quad + \left(\frac{v^*}{u^*} + \frac{v^* d^4}{2u^* b^2} \right) \lambda_2 + O_1((\lambda)^2), \\ \mu_2(\lambda_1, \lambda_2) &= \left(\frac{d(b-d)}{b^2} - \frac{d^6}{2b^4} + \frac{v^*}{1+v^*} + \frac{(b-d)d^5}{2b^4} - \frac{d^{10}}{4b^6} \right) \lambda_1 + \frac{1}{u^*} \lambda_2 + O_2((\lambda)^2).\end{aligned}$$

If the above parameter transformation from (λ_1, λ_2) to (μ_1, μ_2) is not singular in a small neighborhood of $(\lambda_1, \lambda_2) = (0, 0)$, then by the Bogdanov–Takens theory, system (3.4) is strongly topologically equivalent to

$$\begin{aligned}\dot{u}_3 &= v_3, \\ \dot{v}_3 &= \mu_1 + \mu_2 v_3 + u_3^2 + u_3 v_3.\end{aligned} \tag{3.5}$$

By numerical computation, we know that there exist some values of parameter (b, d) such that this parameter transformation is not singular in a small neighborhood of $(\lambda_1, \lambda_2) = (0, 0)$. Therefore, we have the following.

THEOREM 3.3. *When $0 < |h - h_0| \ll 1$, $0 < |a - b| \ll 1$, and $0 < b - d < \frac{b}{a}$, system (2.1) undergoes the cusp bifurcation of codimension 2 (i.e., the Bogadnov–Takens bifurcation). Hence, there exist values of the parameters (h, a, b, d) such that system (2.1) has a unique unstable limit cycle for some parameter values, and system (2.1) has an unstable homoclinic loop for other parameter values.*

We observe that in the model (2.1), the harvesting rate h plays the key role in determining the dynamics of (2.1). When the harvesting rate h tends to the critical harvesting rate h_0 and the predator capturing rate a tends to the predator conversion rate b, Theorem 3.3 says that if the predator death rate d satisfies $0 < b - d < \frac{b}{a}$, then there exist some values of parameters such that the prey and predator coexist in the form of a positive equilibrium or a periodic orbit with a finite period for different initial values, respectively. And there exist other values of parameters such that the prey and predator coexist in the form of a positive equilibrium for all initial values lying inside the homoclinic loop, and the prey and predator coexist in the form of a periodic orbit with infinite period for all initial values on the homoclinic loop.

3.4. The heteroclinic bifurcation and separatrix connecting a saddle-node and a saddle bifurcation. From Theorem 2.4, there may exist a separatrix connecting the saddle-node (x^*, y^*) and the saddle $(x_2, 0)$ when $a < b$, $0 < b - d < \frac{b}{a}$, and $h = h_0$. By Theorem 2.5, the saddle-node (x^*, y^*) separates into two hyperbolic equilibria, (x_1^*, y_1^*) and (x_2^*, y_2^*). Thus, when $a < b$, $0 < b - d < \frac{b}{a}$, and $0 < h < h_0$, the separatrix connecting is broken and the *separatrix connecting a saddle-node and a saddle bifurcation* occurs. There are only three possibilities for the separatrix of saddle $(x_2, 0)$ of system (2.1): (i) the separatrix of saddle $(x_2, 0)$ tends to the stable focus (x_2^*, y_2^*); (ii) there exists a heteroclinic orbit connecting saddles $(x_2, 0)$ and (x_1^*, y_1^*); (iii) the separatrix of saddle $(x_2, 0)$ goes out of the first quadrant. If possibility (ii) occurs, then system (2.1) undergoes the *heteroclinic bifurcation*. Because of technical problems, we cannot determine the exact bifurcation points.

The existence of the heteroclinic bifurcation and separatrix connecting a saddle-node and a saddle bifurcation implies that the prey and predator can coexist in the form of two positive equilibria for some initial values.

4. Discussion. Our systematic work on system (2.1) reveals that the ratio-dependent model with a constant rate harvesting is more interesting and richer in dynamics compared to the ratio-dependent model. It has been shown that the nonzero constant prey harvesting rate prevents mutual extinction as a possible outcome of predator-prey interaction. Biologically, there is still a lot of work to do in this area. For example, it would be interesting to see what the behavior of model (1.2) would be when the harvesting constant is in the predator equation. Ideally, we would be interested in studying model (1.2) with both predator and prey harvesting constants since we usually harvest, or would like to harvest, both populations. In [8], [9], and [10], Brauer and Soudack noticed some different types of dynamics whether the harvesting was in the prey or in the predator equation for a class of predator-prey system. Mathematically, we would like to point out here that our analysis of model (2.1) is a first look at the local bifurcations of the degenerate saddle-node of codimension 4, but it is far from complete. Many questions on the dynamics of the unfoldings of the saddle-node singularity of codimension 4 remain untouched. To the best of our knowledge, there are four kinds of Bogadnov–Takens-type singularity (cusp, saddle, focus, elliptic) of codimension 3, and the unfoldings of these singularities of codimension 3 are extremely complicated (cf. [19], [27], [12] and therein). The dynamical features of codimension 4 are so rich that a full description of the topology of the unfolding seems hopeless. Luckily, however, we see that there have been some perfect works that deal with the unfoldings of singularity of saddle (or cusp or elliptic) of codimension 4 (cf. [15], [16], [17], [18], [25]). In those papers, delicate results such as the limit cycles, their number, and bifurcation patterns are discussed. And it is interesting that the bifurcation of these singularities of codimension 4 can undergo the bifurcation of only the same type as singularity of codimension 3. Different from the known results of codimension 4, the bifurcation of the saddle-node singularity of codimension 4 can undergo the bifurcation of different-type singularity of codimension 3 from the following arguments. The canonical family of the saddle-node singularity of codimension 4 is

$$\begin{aligned} &\dot{x} = y, \\ &\dot{y} = \epsilon x^4 + \mu_3 x^2 + \mu_2 x + \mu_1 + y(\mu_4 + bx + cx^3), \end{aligned} \tag{4.1}$$

where $\epsilon = \pm 1$, $c, b > 0$ are parameters, and $0 < |\mu_i| \ll 1$, $i = 1, 2, \ldots, 4$, are small parameters.

The number of equilibrium points of system (4.1) depends on the number of real roots of

$$\epsilon x^4 + \mu_3 x^2 + \mu_2 x + \mu_1 = 0. \tag{4.2}$$

If (4.2) has a multiple real root with multiplicity 3, then system (4.1) has a singularity of codimension 3. The singularity can be saddle, focus, or elliptic depending on the different values of parameters (μ_1, μ_2, μ_3) and b, respectively. Thus, the unfoldings of the saddle-node singularity of codimension 4 should undergo the saddle, focus, and elliptic bifurcations of codimension 3 generally when the values of parameters vary. Its detailed qualitative and general dynamical picture of the unfoldings of the saddle-node singularity of codimension 4 remain to be seen. Naturally, these interesting topics should be pursued in the future.

It is a pity that system (2.1) does not have the singularity of codimension 3 for all possible values of original parameters in R_+^2. Thus, system (2.1) cannot be a universal unfolding of the saddle-node singularity of codimension 4. Hence, system (2.1) is only an unfolding of the saddle-node singularity of codimension 4, which leads to some bifurcations and loses some bifurcations for all possible values of parameters; for example, system (2.1) may undergo the separatrix connecting a saddle-node and a saddle bifurcation and the heteroclinic bifurcation according to the results of the saddle bifurcation of codimension 3 in [19] and [27], but system (2.1) cannot undergo the bifurcation of heteroclinic connections of two hyperbolic saddle which are produced by the saddle bifurcation of codimension 3.

Appendix A. Proof of Theorem 2.2. Since we consider the topological structure of orbits of system (2.1) in a small neighborhood of (x_0, y_0), we make the following change of variable:

$$u = x, \qquad v = \frac{y}{x}.$$

There exists a neighborhood $N_{(x_0,y_0)}$ of (x_0, y_0) such that system (2.1) in $N_{(x_0,y_0)}$ is differential homeomorphic to the following system in a neighborhood $\bar{N}_{(u_0,v_0)}$ of (u_0, v_0), where $u_0 = \frac{1}{2}$, $v_0 = 0$,

$$\begin{aligned} \dot{u} &= u(1-u) - \frac{auv}{1+v} - \frac{1}{4}, \\ \dot{v} &= v\left(-d + \frac{b}{1+v}\right) - v(1-u) + \frac{av^2}{1+v} + \frac{v}{4u}. \end{aligned} \tag{A.1}$$

Moving the equilibrium $(\frac{1}{2}, 0)$ to the origin, system (A.1) becomes

$$\begin{aligned} \dot{u} &= -\frac{1}{2}av - u^2 - auv + \frac{1}{2}av^2 + auv^2 - \frac{1}{2}av^3 - auv^3 + \frac{1}{2}av^4 + O((u,v)^5), \\ \dot{v} &= (b-d)v + (a-b)v^2 + 2u^2v + (b-a)v^3 + (a-b)v^4 - 4u^3v + O((u,v)^5). \end{aligned} \tag{A.2}$$

The conclusion (i) comes from straightforward analysis of system (A.2), as used in [30].

When $b - d = 0$, we use the procedure, used in [28], to reduce system (A.2) via the normal form method. Let

$$w_1 = -\frac{2}{a}u, \quad z_1 = v.$$

We have

$$\begin{aligned}\dot{w}_1 &= z_1 + \frac{a}{2}w_1^2 - aw_1z_1 - z_1^2 + aw_1z_1^2 + z_1^3 - aw_1z_1^3 - z_1^4 + O((w_1,z_1)^5),\\ \dot{z}_1 &= (a-b)z_1^2 + \frac{a^2}{2}w_1^2z_1 - (a-b)z_1^3 + \frac{1}{2}a^3w_1^3z_1 + (a-b)z_1^4 + O((w_1,z_1)^5).\end{aligned} \tag{A.3}$$

Consider the following C^∞ changes of coordinates in a small neighborhood of $(0,0)$ step by step

$$w_2 = w_1 + \frac{b}{2}w_1^2 + w_1z_1, \quad z_2 = z_1 + \frac{a}{2}w_1^2 - (a-b)w_1z_1;$$

$$w_3 = w_2 + \frac{a+ab+b^2}{6}w_2^3 - \frac{a-2b}{2}w_2^2z_2, \quad z_3 = z_2 + \frac{ab}{2}w_2^3 - \frac{ab-2a-b^2}{2}w_2^2z_2;$$

and

$$\begin{aligned}w_4 &= w_3,\\ z_4 &= z_3 + \left(\frac{1}{2}a^2b - \frac{1}{4}a^2 - \frac{3}{4}ab^2\right)w_3^4 + \left(\frac{1}{2}ab - \frac{1}{2}a^2 - a^2b + 2ab^2\right)w_3^3z_3\\ &\quad + \left(-a + 2a^2 - \frac{7}{2}ab + 5b^2\right)w_3^2z_3^2 + (4b-2a)w_3z_3^3 + O((w_3,z_3)^5).\end{aligned}$$

Then system (A.3) can be transformed into

$$\begin{aligned}\dot{w}_4 &= z_4,\\ \dot{z}_4 &= aw_4z_4 + \frac{1}{2}abw_4^2z_4 + \left(\frac{a^3}{2} - \frac{5}{4}a^2b\right)w_4^4\\ &\quad + \left(\frac{10}{3}a^2b - \frac{8}{3}ab^2 - \frac{7}{6}a^2 - a^3\right)w_4^3z_4 + \left(-a^2 + 3ab - a^2b + \frac{3}{2}ab^2 + \frac{5}{2}b^3\right)w_4^2z_4^2\\ &\quad + (4b-2a)z_4^4 + O((w_4,z_4)^5).\end{aligned} \tag{A.4}$$

The equilibrium $(0,0)$ of system (A.4) is a degenerate saddle-node of codimension 4 if $2a - 5b \neq 0$. The proof is completed.

Appendix B. Proof of Theorem 2.4. The existence of three equilibria comes from Lemma 2.1. Straightforward computing of the eigenvalues of the linear matrix at the equilibrium (x_1, y_1) and (x_2, y_2), respectively, reveals that (x_1, y_1) is a hyperbolic unstable node, and (x_2, y_2) is a hyperbolic saddle in both cases. Next, we determine only the various types of dynamical behavior of the equilibrium (x^*, y^*).

For simplicity, to reduce the normal form of system (2.1) at (x^*, y^*), we make the following change of variable

$$u = x, \quad v = \frac{y}{x}.$$

This transformation of variable is a differential homeomorphism in the interior of R_+^2. Thus, there exists a neighborhood $B_{(x^*,y^*)}$ of (x^*, y^*) such that system (2.1) in $B_{(x^*,y^*)}$ is differential homeomorphic to the following system in a neighborhood $\bar{B}_{(u^*,v^*)}$ of

(u^*, v^*), where $u^* = \frac{b-a(b-d)}{2b}$, $v^* = \frac{b-d}{d}$,

$$\begin{aligned} \dot{u} &= u(1-u) - \frac{auv}{1+v} - h_0 \triangleq \bar{f}_1(u,v), \\ \dot{v} &= v\left(-d + \frac{b}{1+v}\right) - v(1-u) + \frac{av^2}{1+v} + \frac{h_0 v}{u} \triangleq \bar{f}_2(u,v). \end{aligned} \tag{B.1}$$

It is clear that the functions $\bar{f}_1(u,v)$ and $\bar{f}_2(u,v)$ are analytic functions in $\bar{B}_{(u^*,v^*)}$. Let

$$u_1 = u - u^*, \qquad v_1 = v - v^*.$$

Then system (B.1) can be transformed into

$$\begin{aligned} \dot{u}_1 &= -\frac{au^*d^2}{b^2}v_1 - u_1^2 - \frac{ad^2}{b^2}u_1v_1 + \frac{au^*d^3}{b^3}v_1^2 + g_1(u_1,v_1), \\ \dot{v}_1 &= \frac{d(a-b)(b-d)}{b^2}v_1 + \frac{b-d}{du^*}u_1^2 + \frac{(a-b)d^3}{b^3}v_1^2 + g_2(u_1,v_1), \end{aligned} \tag{B.2}$$

where $g_i(u_1,v_1)$ is an analytic function with at least powers $u_1^j v_1^k$, $j+k \geq 3$, $i = 1,2$.

In the case (I), i.e., $a \neq b$, straightforward analysis, as used in [30], shows that (x^*, y^*) is a saddle-node and there exist three eigendirections of (x^*, y^*): $\theta_1 = \operatorname{arctg}(\frac{-(a-b)(b-d)}{adu^*})$, $\theta_2 = \pi$, and $\theta_3 = \pi + \operatorname{arctg}(\frac{-(a-b)(b-d)}{adu^*})$. Furthermore, when $a > b$, there exists a unique orbit in R_+^2 convergent to (x^*, y^*), and all other orbits in R_+^2 go out of R_+^2 by crossing the y-axis. Hence, there does not exist a separatrix connecting the saddle-node (x^*, y^*) and the saddle $(x_2, 0)$. When $a < b$, there exist many orbits in R_+^2 convergent to (x^*, y^*), and the other orbits in R_+^2 go out of R_+^2 by crossing the y-axis. In this case, there may exist a separatrix connecting the saddle-node (x^*, y^*) and the saddle $(x_2, 0)$. The phase portraits are shown in Figure 2.3 (both parts (I)).

In the case (II), i.e., $a = b$, let

$$u_2 = -\frac{b}{d^2u^*}u_1, \quad v_2 = v_1.$$

Then system (B.2) becomes

$$\begin{aligned} \dot{u}_2 &= v_2 + \frac{u^*d^2}{b}u_2^2 - \frac{d^2}{b}u_2v_2 - \frac{d}{b}v_2^2 + \bar{g}_1(u_2,v_2), \\ \dot{v}_2 &= \frac{u^*d^3(b-d)}{b^2}u_2^2 + \bar{g}_2(u_2,v_2), \end{aligned} \tag{B.3}$$

where $\bar{g}_i(u_2,v_2)$ is an analytic function with at least powers $u_2^j v_2^k$, $j+k \geq 3$, $i = 1,2$.

We make the following near-identity changes of variables of system (B.3) in a small neighborhood of $(0,0)$,

$$u_3 = u_2 + \frac{d^2}{2b}u_2^2 + \frac{d}{b}u_2v_2, \qquad v_3 = v_2 + \frac{u^*d^2}{b}u_2^2,$$

and

$$u_4 = u_3, \quad v_4 = v_3 + O_1(|(u_3,v_3)|^3).$$

Then by two steps, system (B.3) becomes the normal form of the cusp of codimension 2

$$\begin{aligned} \dot{u}_4 &= v_4, \\ \dot{v}_4 &= \frac{u^* d^3 (b-d)}{b^2} u_4^2 + \frac{2u^* d^2}{b} u_4 v_4 + O(|(u_4, v_4)|^3), \end{aligned} \tag{B.4}$$

where $O(|(u_4, v_4)|^3)$ is of the same order infinity.

Hence, the equilibrium $(0, 0)$ of system (B.4) is a cusp of codimension 2, i.e., the equilibrium (x^*, y^*) is a cusp of codimension 2, or a Bogdanov–Takens singularity. The phase portrait is shown in Figure 2.3 (II).

Since the closed orbit must include the equilibria in its interior and since the total sum of index of these equilibria equals one, system (2.1) does not have a closed orbit in R_+^2 under the conditions of (I) and (II). This completes the proof of the theorem.

Acknowledgment. The authors are greatly indebted to the anonymous referees for their careful reading and helpful comments, especially the referee who gave constructive suggestions for revision of the manuscript.

REFERENCES

[1] P.A. Abrams and L.R. Ginzburg, *The nature of predation: Prey dependent, ratio-dependent or neither?*, TREE, 15 (2000), pp. 337–341.

[2] H.R. Akcakaya, R. Arditi, and L.R. Ginzburg, *Ratio-dependent prediction: An abstraction that works*, Ecology, 79 (1995), pp. 995–1004.

[3] R. Arditi and A.A. Berryman, *The biological paradox*, Trends in Ecology and Evolution, 6 (1991), p. 32.

[4] R. Arditi and L.R. Ginzburg, *Coupling in predator-prey dynamics: Ratio-dependence*, J. Theor. Biol., 139 (1989), pp. 311–326.

[5] R. Arditi, L.R. Ginzburg, and H.R. Akcakaya, *Variation in plankton densities among lakes: A case for ratio-dependent models*, American Naturalist, 138 (1991), pp. 1287–1296.

[6] A. Andronov, E.A. Leontovich, I.I. Gordon, and A.G. Maier, *Theory of Bifurcations of Dynamical Systems on a Plane*, Halstead Press, New York, 1973.

[7] F. Berezovskaya, G. Karev, and R. Arditi, *Parametric analysis of the ratio-dependent predator-prey model*, J. Math. Biol., 43 (2001), pp. 221–246.

[8] F. Brauer and A.C. Soudack, *Stability regions and transition phenomena for harvested predator-prey systems*, J. Math. Biol., 7 (1979), pp. 319–337.

[9] F. Brauer and A.C. Soudack, *Stability regions in predator-prey systems with constant rate prey harvesting*, J. Math. Biol., 8 (1979), pp. 55–71.

[10] F. Brauer and A.C. Soudack, *Coexistence properties of some predator-prey systems under constant rate harvesting and stocking*, J. Math. Biol., 12 (1981), pp. 101–114.

[11] S.-N. Chow and J.K. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, New York, Heidelberg, Berlin, 1982.

[12] S.-N. Chow, C.Z. Li, and D. Wang, *Normal Forms and Bifurcation of Planar Vector Fields*, Cambridge University Press, Cambridge, UK, 1994.

[13] C.W. Clark, *Mathematical Bioeconomics, The Optimal Management of Renewable Resources*, 2nd ed., John Wiley & Sons, New York, Toronto, 1990.

[14] C. Cosner, D.L. DeAngelis, J.S. Ault, and D.B. Olson, *Effects of spatial grouping on the functional response of predators*, Theor. Pop. Biol., 56 (1999), pp. 65–75.

[15] F. Dumortier, P. Fiddelaers, and C. Li, *Generic unfolding of the nilpotent saddle of codimension four*, in Global Analysis of Dynamical Systems, Institute of Physics, Bristol, UK, 2001, pp. 131–166.

[16] F. Dumortier and C. Li, *Perturbations from an elliptic Hamiltonian of degree four*, in Differential Equations and Computational Simulations (Chengdu, 1999), P.W. Bates, K. Lu, and D. Xu, eds., World Scientific, River Edge, NJ, 2000, pp. 66–70.

[17] F. Dumortier and C. Li, *Perturbations from an elliptic Hamiltonian of degree four.* I. *Saddle loop and two saddle cycle*, J. Differential Equations, 176 (2001), pp. 114–157.

[18] F. Dumortier and C. Li, *Perturbations from an elliptic Hamiltonian of degree four.* II. *Cuspidal loop*, J. Differential Equations, 175 (2001), pp. 209–243.

[19] F. Dumortier, R. Roussarie, and J. Sotomayor, *Generic Three-Parameter Families of Planar Vector Fields, Unfoldings of Saddle, Focus and Elliptic Singularities with Nilpotent Linear Parts*, Lecture Notes in Math. 1480, Springer-Verlag, Berlin, 1991.

[20] A.P. Gutierrez, *The physiological basis of ratio-dependent predator-prey theory: A metabolic pool model of Nicholson's blowflies as an example*, Ecology, 73 (1992), pp. 1552–1563.

[21] S.B. Hsu, T.W. Hwang, and Y. Kuang, *Global analysis of the Michaelis-Menten type ratio-dependent predator-prey system*, J. Math. Biol., 42 (2001), pp. 489–506.

[22] C. Jost, O. Arino, and R. Arditi, *About deterministic extinction in ratio-dependent predator-prey models*, Bull. Math. Biol., 61 (1999), pp. 19–32.

[23] Y. Kuang, *Rich dynamics of Gause-type ratio-dependent predator-prey system*, in Differential Equations with Applications to Biology, Fields Inst. Commun. 21, S. Ruan, G.S.K. Wolkowicz, and J. Wu, eds., AMS, Providence, RI, 1999, pp. 325–337.

[24] Y. Kuang and E. Beretta, *Global qualitative analysis of a ratio-dependent predator-prey system*, J. Math. Biol., 36 (1998), pp. 389–406.

[25] C. Li and C. Rousseau, *A system with three limit cycles appearing in a Hopf bifurcation and dying in a homoclinic bifurcation: The cusp of order* 4, J. Differential Equations, 79 (1989), pp. 132–167.

[26] L. Perko, *Differential Equations and Dynamical Systems*, 2nd ed., Springer-Verlag, New York, 1996.

[27] D. Xiao, *Bifurcations of saddle singularity of codimension three of a planar vector field with nilpotent linear part*, Sci. Sinica A, 23 (1993), pp. 252–260.

[28] D. Xiao and S. Ruan, *Bogdanov-Takens bifurcations in predator-prey systems with constant rate harvesting*, in Differential Equations with Applications to Biology, Fields Inst. Commun. 21, S. Ruan, G.S.K. Wolkowicz, and J. Wu, eds., AMS, Providence, RI, 1999, pp. 493–506.

[29] D. Xiao and S. Ruan, *Global dynamics of a ratio-dependent predator-prey system*, J. Math. Biol., 43 (2001), pp. 268–290.

[30] Z. Zhang, T. Ding, W. Huang, and Z. Dong, *Qualitative Theory of Differential Equations*, Transl. Math. Monogr. 101, AMS, Providence, RI, 1992.

SIAM J. APPL. MATH.
Vol. 65, No. 3, pp. 754–766

GEOMETRIC SINGULAR PERTURBATION APPROACH TO STEADY-STATE POISSON–NERNST–PLANCK SYSTEMS*

WEISHI LIU†

Abstract. Boundary value problems of a one-dimensional steady-state Poisson–Nernst–Planck (PNP) system for ion flow through a narrow membrane channel are studied. By assuming the ratio of the Debye length to a characteristic length to be small, the PNP system can be viewed as a singularly perturbed problem with multiple time scales and is analyzed using the newly developed geometric singular perturbation theory. Within the framework of dynamical systems, the global behavior is first studied in terms of limiting fast and slow systems. It is rather surprising that a *complete* set of integrals is discovered for the (nonlinear) limiting fast system. This allows a detailed description of the boundary layers for the problem. The slow system itself turns out to be a singularly perturbed one, too, which indicates that the singularly perturbed PNP system has three different time scales. A singular orbit (zeroth order approximation) of the boundary value problem is identified based on the dynamics of limiting fast and slow systems. An application of the geometric singular perturbation theory gives rise to the existence and (local) uniqueness of the boundary value problem.

Key words. singular perturbation, boundary layers, exchange lemma

AMS subject classifications. 34A26, 34B16, 34D15, 37D10, 92C35

DOI. 10.1137/S0036139903420931

1. Introduction. Poisson–Nernst–Planck (PNP) systems serve as basic electrodiffusion equations modeling, for example, ion flow through membrane channels, and transport of holes and electrons in semiconductors (see [1, 2, 11, 14] and references therein). In the context of ion flow through a membrane channel, the flow of ions is driven by their concentration gradients and by the electric field modeled together by the Nernst–Planck equations, and the electric field is in turn governed by the ion concentrations through the Poisson equation. To motivate the one-dimensional PNP system to be studied, we give a brief account of the modeling. We will be interested in flow of two types of ions through a narrow membrane channel. For practical purposes, the narrow membrane channel through which ions flow is tubelike with a small aspect ratio and, in this regard, it is natural to approximate the channel as a one-dimensional object (see, e.g., [1, 2]). Now consider flow of two types of ions, S_1 and S_2, with valences $\alpha > 0$ and $-\beta < 0$, passing through an ion channel viewed as a line segment. Let x be the coordinate along the channel normalized from $x = 0$ to $x = 1$. Denote the concentrations of S_1 and S_2 at location x and at time t by $c_1(t, x)$ and $c_2(t, x)$. Then the electric potential $\phi(t, x)$ in the channel at time t is determined by the Poisson equation

$$\frac{\partial^2 \phi}{\partial x^2} = -\frac{1}{\epsilon^2}(\alpha c_1 - \beta c_2),$$

where the parameter ϵ^2 is related to the ratio of the Debye length to a characteristic length scale. The flux densities, $\bar{J}_1$ and $\bar{J}_2$, of the two ions contributed from the concentration gradients of the two ions and the electric field satisfy the Nernst–Planck

*Received by the editors January 8, 2003; accepted for publication (in revised form) December 1, 2003; published electronically February 25, 2005.
http://www.siam.org/journals/siap/65-3/42093.html
†Department of Mathematics, University of Kansas, Lawrence, KS 66045 (wliu@math.ku.edu).
This work was partially supported by NSF grant DMS-0071931.

equations

$$D_1\left(\frac{\partial c_1}{\partial x}+\alpha c_1\frac{\partial \phi}{\partial x}\right)=-\bar{J}_1,\quad D_1\left(\frac{\partial c_2}{\partial x}-\beta c_2\frac{\partial \phi}{\partial x}\right)=-\bar{J}_2,$$

where D_1 and D_2 are the diffusion constants of ions S_1 and S_2 relative to the membrane channel, together with the conservation of mass

$$\frac{\partial c_1}{\partial t}+\frac{\partial \bar{J}_1}{\partial x}=0,\quad \frac{\partial c_2}{\partial t}+\frac{\partial \bar{J}_2}{\partial x}=0.$$

Combining the above equations, we obtain the one-dimensional PNP system as a simplified model for flow of two ions through a narrow membrane channel:

$$\begin{aligned}&\epsilon^2\frac{\partial^2\phi}{\partial x^2}=-(\alpha c_1-\beta c_2),\quad \frac{\partial c_1}{\partial t}+\frac{\partial \bar{J}_1}{\partial x}=0,\quad \frac{\partial c_2}{\partial t}+\frac{\partial \bar{J}_2}{\partial x}=0,\\ &D_1\left(\frac{\partial c_1}{\partial x}+\alpha c_1\frac{\partial \phi}{\partial x}\right)=-\bar{J}_1,\quad D_1\left(\frac{\partial c_2}{\partial x}-\beta c_2\frac{\partial \phi}{\partial x}\right)=-\bar{J}_2.\end{aligned}\tag{1}$$

To understand the asymptotic behavior that is most relevant from a physical point of view, the first step is to study the steady-state problem. On one hand, steady-state solutions are among those that are responsible for the global structure of the full system and, on the other hand, they often represent asymptotic states of solutions of general initial conditions. In this work, we study boundary value problems of the one-dimensional steady-state PNP system. The corresponding system is

$$\begin{aligned}&\epsilon^2\frac{d^2\phi}{dx^2}=-(\alpha c_1-\beta c_2),\quad \frac{dJ_1}{dx}=0,\quad \frac{dJ_2}{dx}=0,\\ &\frac{dc_1}{dx}+\alpha c_1\frac{d\phi}{dx}=-J_1,\quad \frac{dc_2}{dx}-\beta c_2\frac{d\phi}{dx}=-J_2,\end{aligned}\tag{2}$$

where $J_1=\bar{J}_1/D_1$ and $J_2=\bar{J}_2/D_2$, and the boundary conditions are

$$\begin{aligned}&\phi(0)=v_0,\quad c_1(0)=L_1,\quad c_2(0)=L_2,\\ &\phi(1)=0,\quad c_1(1)=R_1,\quad c_2(1)=R_2.\end{aligned}\tag{3}$$

Many mathematical works have been done on the existence, uniqueness, and qualitative properties of boundary value problems even for high dimensional systems, and algorithms have been developed toward numerical approximations (see, e.g., [5, 6, 13, 7]). Under the assumption that $\epsilon \ll 1$, the problem can be viewed as a singularly perturbed system. Typical solutions of singularly perturbed systems exhibit different time scales; for example, boundary and internal layers (inner solutions) evolve at fast pace and regular layers (outer solutions) vary slowly. For the boundary value problems (2) and (3), there are two boundary layers, one at each end. Physically, near boundaries $x = 0$ and $x = 1$, the potential function $\phi(x)$ and the concentration functions $c_1(x)$ and $c_2(x)$ exhibit a large gradient or a sharp change. In [2], for $\alpha = \beta = 1$, the boundary value problem was studied using the method of matched asymptotic expansions as well as numerical simulations, which provide a good quantitative and qualitative understanding of the problem.

We also treat the problem as a singularly perturbed one by assuming $\epsilon \ll 1$ but for general α and β. Our approach uses the newly developed geometric singular perturbation theory (see, e.g., [4, 8, 10, 12]). The basic ideas behind this theory for boundary value problems are

(i) to derive, based on different time scales of the system, various limiting systems for $\epsilon = 0$ and examine their dynamical structures;
(ii) to construct a singular orbit (zeroth order approximation) consisting of orbits of limiting systems, which include boundary layers, regular layers, and, sometimes, internal layers;
(iii) to show that there are true solutions near the singular orbit for $\epsilon > 0$.

Since limiting systems essentially have lower order than the full system, it is often easier to study which make (i) useful. Understanding the dynamics of limiting subsystems allows one to carry out (ii). The most difficult part is the task (iii). It requires us to investigate the interaction between the fast and slow dynamics. A successful type of results is called the exchange lemma (see, e.g., [8, 10, 15, 12]). Its objective is to track the smooth configuration of an invariant manifold as it passes regions overlapping different time scales. For boundary value problems, two invariant manifolds, say, M_L and M_R, will be tracked: M_L will be the trace of one boundary under the flow, and M_R will be the trace of the other boundary. The existence of a solution for $\epsilon > 0$ is then reduced to the nontrivial intersection of M_L and M_R. This is where the exchange lemma comes in to play the crucial role. This approach provides not only a construction of a limiting solution but also a direct verification of the validity of the limiting solution.

The rest of the paper is organized as follows. Section 2 contains three subsections. In section 2.1, the PNP system (2) is rewritten as a singularly perturbed system of first order equations, and the boundary value problem is converted to a *connecting problem*. Two systems, slow and fast systems, with different scales are first identified according to different time scales, and some general aspects of dynamical system theory are laid out for the boundary value problem. The boundary layer behavior governed by the limiting fast system is studied in section 2.2. It is rather surprising that a complete set of integrals is discovered for the nonlinear limiting fast system which allows a detailed study of the boundary layer behavior. (The physical meanings of the integrals remain unclear.) The regular layers governed by the slow flow are analyzed in section 2.3. It turns out that the slow system itself is a singularly perturbed one which is examined using again the geometric singular perturbation theory. In section 3, we construct a singular orbit of the boundary value problem and apply the exchange lemma to show the existence and uniqueness of a solution near the singular orbit. A derivation of the integrals of the fast system is given in section 4 as an appendix.

2. A dynamical system framework.

2.1. A basis of geometric singular perturbation theory. We will recast the singularly perturbed PNP system into a system of first order equations. This singularly perturbed system corresponds to the slow scale which is suitable for understanding dynamics within the membrane channel. A fast scale system can be derived through a change of scale of the independent variable x, which can be used to capture the sharp boundary behavior. Slow and fast systems of the singularly perturbed PNP system are equivalent for $\epsilon \neq 0$, but their limits are not: they provide complementary limiting information for the full system. We begin with a dynamical system formulation of the singularly perturbed PNP system (2).

Denote derivatives with respect to x by overdot symbols and introduce

$$u = \epsilon\dot{\phi}, \quad v = \beta c_2 - \alpha c_1, \quad w = \alpha^2 c_1 + \beta^2 c_2, \quad \text{and} \quad \tau = x.$$

System (2) becomes

$$
\begin{aligned}
&\epsilon\dot{\phi} = u, \quad \epsilon\dot{u} = v, \quad \epsilon\dot{v} = uw - \epsilon(\beta J_2 - \alpha J_1),\\
&\epsilon\dot{w} = \alpha\beta uv + (\beta - \alpha)uw - \epsilon(\alpha^2 J_1 + \beta^2 J_2),\\
&\dot{J}_1 = 0, \quad \dot{J}_2 = 0, \quad \dot{\tau} = 1.
\end{aligned} \tag{4}
$$

System (4) will be treated as a dynamical system with the phase space $\mathbb{R}^7$, and the independent variable x will be viewed as time. The boundary condition (3) becomes

$$
\begin{aligned}
&\phi(0) = v_0, \quad v(0) = \beta L_2 - \alpha L_1, \quad w(0) = \alpha^2 L_1 + \beta^2 L_2, \quad \tau(0) = 0,\\
&\phi(1) = 0, \quad v(1) = \beta R_2 - \alpha R_1, \quad w(1) = \alpha^2 R_1 + \beta^2 R_2, \quad \tau(1) = 1.
\end{aligned} \tag{5}
$$

Formulation of high order equations into dynamical systems of first order equations is not unique. For the boundary value problem considered in this paper, two issues need particular attention. One is toward the derivative of $\phi(x)$. Since $\phi(x)$ is expected to have large derivatives near the boundaries, the introduction of $u = \epsilon\dot{\phi}$ seems natural. The introduction of a new variable $\tau = x$ is a special treatment for boundary value problems. The small price paid is the addition of an extra dimension with trivial dynamics to the phase space. The apparent advantage is that, to find a solution of the boundary value problem, one needs only an orbit from one boundary to the other without worrying how much time it takes the orbit to move from one side to the other: it is automatically 1 since, as a component of the orbit, $\tau = x$ will vary from 0 to 1. The change of variables from c_1 and c_2 to v and w is motivated purely from the analysis point of view.

Observe that by setting $\epsilon = 0$ in system (4), we get $u = v = 0$. The set $\mathcal{Z}_0 = \{u = v = 0\}$ is called *the slow manifold* which supports the regular layer of the boundary value problem. The regular layer will not satisfy all conditions in (5) if $\beta L_2 - \alpha L_1 \neq 0$ or $\beta R_2 - \alpha R_1 \neq 0$, and this defect has to be remedied by boundary layers. To examine boundary layer behavior, we will now derive a system, the fast system, with a time scale different from that of (4). This will be achieved through the following rescaling of time (independent variable) for dependent variables:

$$
\begin{gathered}
\Phi(\xi) = \phi(\epsilon\xi),\ U(\xi) = u(\epsilon\xi),\ V(\xi) = v(\epsilon\xi),\ W(\xi) = w(\epsilon\xi),\\
I_i(\xi) = J_i(\epsilon\xi), \ \text{ and } \ T(\xi) = \tau(\epsilon\xi).
\end{gathered}
$$

Note that capital letters for same dependent variables are used to indicate merely different time scales. In terms of ξ, we obtain *the fast system* of (4):

$$
\begin{aligned}
&\Phi' = U, \quad U' = V, \quad V' = UW - \epsilon(\beta I_2 - \alpha I_1),\\
&W' = \alpha\beta UV + (\beta - \alpha)UW - \epsilon(\alpha^2 I_1 + \beta^2 I_2),\\
&I_1' = 0, \quad I_2' = 0, \quad T' = \epsilon,
\end{aligned} \tag{6}
$$

where the prime symbol denotes the derivative with respect to the variable ξ. The limiting fast system at $\epsilon = 0$ is

$$
\begin{aligned}
&\Phi' = U, \quad U' = V, \quad V' = UW, \quad W' = \alpha\beta UV + (\beta - \alpha)UW,\\
&I_1' = 0, \quad I_2' = 0, \quad T' = 0.
\end{aligned} \tag{7}
$$

The slow manifold $\mathcal{Z}_0$ is precisely the set of equilibria of (7).

Now let B_L and B_R be the subsets of $\mathbb{R}^7$ defined, respectively, by

$$
\begin{aligned}
B_L =&\{\phi = v_0,\ v = \beta L_2 - \alpha L_1,\ w = \alpha^2 L_1 + \beta^2 L_2,\ \tau = 0\},\\
B_R =&\{\phi = 0,\ v = \beta R_2 - \alpha R_1,\ w = \alpha^2 R_1 + \beta^2 R_2,\ \tau = 1\}.
\end{aligned} \tag{8}
$$

The boundary value problem is then equivalent to the following *connecting problem*: finding a solution of (4) from B_L to B_R.

For $\epsilon > 0$, let M_L^ϵ be the union of all forward orbits of (4) starting from B_L and let M_R^ϵ be the union of all backward orbits starting from B_R. To obtain the existence and (local) uniqueness of a solution for the connecting problem, it thus suffices to show M_L^ϵ and M_R^ϵ intersect transversally. The intersection is exactly the orbit of a solution of the boundary value problem, and the transversality implies the local uniqueness. The strategy is to obtain a singular orbit and track the evolution of M_L^ϵ and M_R^ϵ along the singular orbit. As discussed in the introduction, a singular orbit will be a union of orbits of subsystems of (4) with different time scales.

The boundary layers will be two orbits of (7): one from B_L to $\mathcal{Z}_0$ in forward time along the stable manifold of $\mathcal{Z}_0$ and the other from B_R to $\mathcal{Z}_0$ in backward time along the unstable manifold of $\mathcal{Z}_0$. The two boundary layers will be connected by a regular layer on $\mathcal{Z}_0$, which is an orbit of a limiting system of (4). The next two subsections are devoted to the study of boundary layers and regular layers.

2.2. Fast dynamics and boundary layers. We start with the study of boundary layers governed by system (7). This system has many invariant structures that are useful for characterizing the global dynamics.

The slow manifold $\mathcal{Z}_0 = \{U = V = 0\}$ consisting entirely of equilibria of system (7) is a five-dimensional manifold of the phase space $\mathbb{R}^7$. For each equilibrium $z = (\Phi, 0, 0, W, I_1, I_2, T) \in \mathcal{Z}_0$, the linearization of system (7) has five zero eigenvalues corresponding to the dimension of $\mathcal{Z}_0$, and two eigenvalues in directions normal to $\mathcal{Z}_0$. The latter two eigenvalues and their associated eigenvectors are given by

$$
\lambda_\pm = \pm\sqrt{W} \ \text{ and } \ n_\pm = \Big((\pm\sqrt{W})^{-1}, 1, \pm\sqrt{W}, \pm(\beta-\alpha)\sqrt{W}, 0, 0, 0\Big)^\tau. \tag{9}
$$

Thus, every equilibrium has a one-dimensional stable manifold and a one-dimensional unstable manifold. The global configurations of the stable and unstable manifolds will be needed for the boundary layer behavior. For any constants I_1^*, I_2^*, and T^*, the set $\mathcal{N} = \{I_1 = I_1^*, I_2 = I_2^*, T = T^*\}$ is a four-dimensional invariant subspace of the phase space $\mathbb{R}^7$.

Surprisingly, system (7) possesses a complete set of integrals with which the dynamics can be fully analyzed; in particular, the stable and unstable manifolds can be characterized and the behavior of boundary layers can be described in detail.

PROPOSITION 2.1. (i) *System* (7) *has a complete set of six integrals given by*

$$
\begin{aligned}
&H_1 = W - (\beta-\alpha)V - \frac{\alpha\beta}{2}U^2,\ H_2 = \Phi - \frac{\ln|W+\alpha V|}{\beta},\\
&H_3 = |W+\alpha V|^\alpha |W-\beta V|^\beta,\ H_4 = I_1,\ H_5 = I_2,\ \ \textit{and}\ \ H_6 = T,
\end{aligned}
$$

where the argument of H_i's is $(\Phi, U, V, W, I_1, I_2, T)$.

(ii) *The stable and unstable manifolds $W^s(\mathcal{Z}_0)$ and $W^u(\mathcal{Z}_0)$ of $\mathcal{Z}_0$ are characterized as follows:*

$$
W^s(\mathcal{Z}_0) = \cup\{W^s(z^*) :\ z^* \in \mathcal{Z}_0\} \ \ \textit{and} \ \ W^u(\mathcal{Z}_0) = \cup\{W^u(z^*) :\ z^* \in \mathcal{Z}_0\}
$$

and, for $z^*=(\Phi^*,0,0,W^*,I_1^*,I_2^*,T^*)\in\mathcal{Z}_0$, *a point* $z=(\Phi,U,V,W,I_1,I_2,T)\in W^s(z^*)\cup W^u(z^*)$ *if and only if*

$$H_1(z)=W^*,\ H_2(z)=\Phi^*-\frac{\ln W^*}{\beta},\ H_3(z)=(W^*)^{\alpha+\beta},\ I_i=I_i^*,\ T=T^*.$$

(iii) *The stable manifold* $W^s(\mathcal{Z}_0)$ *intersects* B_L *transversally at points with*

$$U=-\text{sgn }(\beta L_2-\alpha L_1)\sqrt{\frac{2\alpha\beta(L_1+L_2)-2(\alpha+\beta)(\alpha L_1)^{\frac{\beta}{\alpha+\beta}}(\beta L_2)^{\frac{\alpha}{\alpha+\beta}}}{\alpha\beta}} \tag{10}$$

and arbitrary I_1 *and* I_2, *where* sgn *is the sign function. The unstable manifold* $W^u(\mathcal{Z}_0)$ *intersects* B_R *transversally at points with*

$$U=\text{sgn }(\beta R_2-\alpha R_1)\sqrt{\frac{2\alpha\beta(R_1+R_2)-2(\alpha+\beta)(\alpha R_1)^{\frac{\beta}{\alpha+\beta}}(\beta R_2)^{\frac{\alpha}{\alpha+\beta}}}{\alpha\beta}} \tag{11}$$

and arbitrary I_1 *and* I_2. *Let* $N_L=B_L\cap W^s(\mathcal{Z}_0)$ *and* $N_R=B_R\cap W^u(\mathcal{Z}_0)$. *Then,*

$$\omega(N_L)=\left\{\left(v_0+\frac{1}{\alpha+\beta}\ln\frac{\alpha L_1}{\beta L_2},0,0,(\alpha+\beta)(\alpha L_1)^{\frac{\beta}{\alpha+\beta}}(\beta L_2)^{\frac{\alpha}{\alpha+\beta}},I_1,I_2,0\right)\right\},$$

$$\alpha(N_R)=\left\{\left(\frac{1}{\alpha+\beta}\ln\frac{\alpha R_1}{\beta R_2},0,0,(\alpha+\beta)(\alpha R_1)^{\frac{\beta}{\alpha+\beta}}(\beta R_2)^{\frac{\alpha}{\alpha+\beta}},I_1,I_2,1\right)\right\}$$

for all I_1 *and* I_2.

Proof. The statement (i) can be verified directly (see section 4 for a derivation of H_3). The statement (ii) is a simple consequence of (i) together with the fact that $\Phi(\xi)\to\Phi^*$, $W(\xi)\to W^*$, $U(\xi)\to 0$, and $V(\xi)\to 0$ as $\xi\to\infty$ for the stable manifold and as $\xi\to-\infty$ for the unstable manifold.

For the statement (iii), we present only the proof regarding the intersection of $W^s(\mathcal{Z}_0)$ and B_L. Suppose

$$z^0=(\Phi^0,U^0,V^0,W^0,I_1^0,I_2^0,0)=(v_0,U^0,\beta L_2-\alpha L_1,\alpha^2L_1+\beta^2L_2,I_1^0,I_2^0,0)$$

is a point in $B_L\cap W^s(\mathcal{Z}_0)$. Then, using the integrals H_1, H_2, and H_3, the solution $z(\xi)=(\Phi(\xi),U(\xi),V(\xi),W(\xi),I_1^0,I_2^0,0)$ of system (7) with initial condition $z(0)=z^0$ satisfies

$$\begin{aligned}
H_1(z(\xi))&=W(\xi)-(\beta-\alpha)V(\xi)-\frac{\alpha\beta}{2}U^2(\xi)=A,\\
H_2(z(\xi))&=\Phi(\xi)-\frac{\ln|W(\xi)+\alpha V(\xi)|}{\beta}=B,\\
H_3(z(\xi))&=|W(\xi)+\alpha V(\xi)|^\alpha|W(\xi)-\beta V(\xi)|^\beta=C
\end{aligned}$$

for some constants A, B, and C, and for all ξ. Since $U(\xi)\to 0$ and $V(\xi)\to 0$ as $\xi\to+\infty$, $W(+\infty)=A$ from $H_1(z(\xi))=A$, and hence, $C=A^{\alpha+\beta}$ from $H_3(z(\xi))=C$. Now using the equations $H_3(z(0))=C=A^{\alpha+\beta}$ and $H_2(z(0))=B$, we have

$$A=(\alpha+\beta)(\alpha L_1)^{\frac{\beta}{\alpha+\beta}}(\beta L_2)^{\frac{\alpha}{\alpha+\beta}},\quad B=v_0-\frac{\ln\left((\alpha+\beta)\beta L_2\right)}{\beta}.$$

Then, from $H_1(z(0)) = A$ and $H_2(z(\infty)) = B$, one has

$$U^0 = -\text{sgn }(V^0)\sqrt{\frac{2(\alpha\beta(L_1+L_2)-A)}{\alpha\beta}} \text{ and } \Phi(+\infty) = v_0 + \frac{1}{\alpha+\beta}\ln\frac{\alpha L_1}{\beta L_2}.$$

The choice of the sign for U^0 comes from the consideration that the stable eigenvector n_- in (9) has U and V components with opposite signs. Thus, B_L and $W^s(\mathcal{Z}_0)$ intersect at the points with $U = U^0$ given above, and all I_1 and I_2. If $N_L = B_L \cap W^s(\mathcal{Z}_0)$, then $\omega(N_L) = \{(\Phi(+\infty), 0, 0, W(+\infty), I_1, I_2, 0)\}$. The above formulas for $\Phi(+\infty)$ and $W(+\infty) = A$ give the desired characterization of $\omega(N_L)$. Lastly, since the stable manifold is completely characterized, one can compute its tangent space at each intersection point to verify the transversality of the intersection. It is slightly complicated but straightforward. We will omit the detail here. □

Part (iii) of this result implies that the boundary layer on the left end will be an orbit of (7) from $(v_0, U_L, \beta L_2 - \alpha L_1, \alpha^2 L_1 + \beta^2 L_2, I_1, I_2, 0) \in B_L$ to the point

$$z_L = \left(v_0 + \frac{1}{\alpha+\beta}\ln\frac{\alpha L_1}{\beta L_2}, 0, 0, (\alpha+\beta)(\alpha L_1)^{\frac{\beta}{\alpha+\beta}}(\beta L_2)^{\frac{\alpha}{\alpha+\beta}}, I_1, I_2, 0\right) \in \mathcal{Z}_0,$$

where U_L is given by the display (10) and I_1 and I_2 are arbitrary at this moment, and that on the right end will be a backward orbit of (7) from the point $(0, U_R, \beta R_2 - \alpha R_1, \alpha^2 R_1 + \beta^2 R_2, I_1, I_2, 1) \in B_R$ to the point

$$z_R = \left(\frac{1}{\alpha+\beta}\ln\frac{\alpha R_1}{\beta R_2}, 0, 0, (\alpha+\beta)(\alpha R_1)^{\frac{\beta}{\alpha+\beta}}(\beta R_2)^{\frac{\alpha}{\alpha+\beta}}, I_1, I_2, 1\right) \in \mathcal{Z}_0,$$

where U_R is given by the display (11) and I_1 and I_2 are arbitrary at this moment. It turns out that there is a unique pair of numbers I_1 and I_2 so that the corresponding points z_L and z_R can be connected by a regular layer solution on $\mathcal{Z}_0$. The regular orbit together with the two boundary layer orbits provides the singular orbit.

Remark 2.1. The integrals H_2 and H_3 imply that

$$\tilde{H}_2 = \Phi + \frac{\ln|W - \beta V|}{\alpha}$$

is also an integral which can be viewed as the symmetric part to H_2.

To find the explicit expressions of the boundary layers from B_L and B_R to $\mathcal{Z}_0$, there are certain technical difficulties. But for some special cases, for example, $\alpha = \beta$, or $\alpha = 2$ and $\beta = 1$, or $\alpha = 1$ and $\beta = 2$, the difficulty can be overcome. In particular, our results for the case $\alpha = \beta = 1$ agree with those in [2], and we provide the detail below for demonstration.

COROLLARY 2.2. *If $\alpha = \beta = 1$, then the expressions of the solutions from B_L and B_R to $\mathcal{Z}_0$ can be explicitly given.*

Proof. We will derive the solution from B_L to $\mathcal{Z}_0$ for general α and β first. Let $r = W + \alpha V$ and $s = W - \beta V$. Then, $r^\alpha s^\beta = A^{\alpha+\beta}$, where A is as in Proposition 2.1, $W = (\beta r + \alpha s)/(\alpha+\beta)$, and $V = (r-s)/(\alpha+\beta)$. Using the equations in (7), one gets

$$r' = \pm\sqrt{\frac{2\beta}{\alpha(\alpha+\beta)}}r\sqrt{\alpha r + \beta A^{\frac{\alpha+\beta}{\beta}}r^{-\frac{\alpha}{\beta}} - (\alpha+\beta)A}.$$

The technical difficulty mentioned above for general α and β is the integration of this equation. Once r is found, the rest can be explicitly solved. The equation can be integrated for the cases mentioned above. We now carry out the rest of the analysis for $\alpha = \beta = 1$.

Without loss of generality, we assume $L_2 > L_1$. Then, $A = 2\sqrt{L_1 L_2}$ and

$$r' = -\sqrt{r}(r - 2\sqrt{L_1 L_2}).$$

Solving the equation and using $r(0) = W(0) + V(0) = 2L_2$, one gets

$$r = \frac{A(1 + ce^{-\sqrt{A}\xi})^2}{(1 - ce^{-\sqrt{A}\xi})^2}, \quad \text{where} \quad c = \frac{L_2^{1/4} - L_1^{1/4}}{L_2^{1/4} + L_1^{1/4}}.$$

Thus,

$$s = \frac{A^2}{r} = \frac{A(1 - ce^{-\sqrt{A}\xi})^2}{(1 + ce^{-\sqrt{A}\xi})^2}, \; W = \frac{r+s}{2} = A\left(1 + \frac{8c^2 e^{-2\sqrt{A}\xi}}{(1 - c^2 e^{-2\sqrt{A}\xi})^2}\right),$$

$$V = \frac{r-s}{2} = \frac{4Ace^{-\sqrt{A}\xi}(1 + c^2 e^{-2\sqrt{A}\xi})}{(1 - c^2 e^{-2\sqrt{A}\xi})^2}, \; U = -\sqrt{2W - 2A} = -\frac{4\sqrt{A}ce^{-\sqrt{A}\xi}}{1 - c^2 e^{-2\sqrt{A}\xi}},$$

$$\Phi = B + \ln(W + V) = v_0 + \frac{1}{2}\ln\frac{L_1}{L_2} + 2\ln\left|\frac{1 + ce^{-\sqrt{A}\xi}}{1 - ce^{-\sqrt{A}\xi}}\right|.$$

The expression for Φ is obtained by either using the integral H_2 and the solutions for V and W or by directly integrating $\Phi' = U$ from U. □

2.3. Slow dynamics and regular layers. We now examine the slow flow in the vicinity of the slow manifold $\mathcal{Z}_0 = \{u = v = 0\}$ for regular layers. If we take $\epsilon = 0$ in system (4), we get $u = v = 0$ and

$$\dot{J}_1 = 0, \; \dot{J}_2 = 0, \; \dot{\tau} = 1.$$

The information on ϕ and w is lost. This indicates that the slow flow in the vicinity of $\mathcal{Z}_0$ is itself a singular perturbation problem. To see this, we zoom into an $O(\epsilon)$-neighborhood of $\mathcal{Z}_0$ by blowing up the u and v coordinates; that is, we make a scaling $u = \epsilon p$ and $v = \epsilon q$. System (4) becomes

$$\begin{aligned} &\dot{\phi} = p, \quad \epsilon\dot{p} = q, \quad \epsilon\dot{q} = pw - (\beta J_2 - \alpha J_1), \\ &\dot{w} = \epsilon\alpha\beta pq + (\beta - \alpha)pw - (\alpha^2 J_1 + \beta^2 J_2), \\ &\dot{J}_1 = 0, \quad \dot{J}_2 = 0, \quad \dot{\tau} = 1, \end{aligned} \tag{12}$$

which is indeed a singular perturbation problem. When $\epsilon = 0$, the system reduces to

$$\begin{aligned} &\dot{\phi} = p, \quad 0 = q, \quad 0 = pw - (\beta J_2 - \alpha J_1), \\ &\dot{w} = (\beta - \alpha)pw - (\alpha^2 J_1 + \beta^2 J_2), \\ &\dot{J}_1 = 0, \quad \dot{J}_2 = 0, \quad \dot{\tau} = 1. \end{aligned} \tag{13}$$

The dynamics of ϕ and w survives in this limiting process. For this system, the slow manifold is

$$\mathcal{S}_0 = \left\{ p = \frac{\beta J_2 - \alpha J_1}{w},\ q = 0 \right\}.$$

The corresponding fast system obtained by the scaling of time

$$\Phi(\xi) = \phi(\epsilon\xi),\ P(\xi) = p(\epsilon\xi),\ Q(\xi) = q(\epsilon\xi),\ \text{ and }\ W(\xi) = w(\epsilon\xi)$$

is

$$\begin{aligned} &\Phi' = \epsilon P, \quad P' = Q, \quad Q' = PW - (\beta I_2 - \alpha I_1), \\ &W' = \epsilon^2 \alpha\beta PQ + \epsilon(\beta - \alpha)PW - \epsilon(\alpha^2 I_1 + \beta^2 I_2), \\ &I_1' = 0, \quad I_2' = 0, \quad T' = 0. \end{aligned} \tag{14}$$

The limiting system of (14) when $\epsilon = 0$ is

$$\begin{aligned} &\Phi' = 0, \quad P' = Q, \quad Q' = PW - (\beta I_2 - \alpha I_1), \\ &W' = 0, \quad I_1' = 0, \quad I_2' = 0, \quad T' = 0. \end{aligned} \tag{15}$$

The slow manifold $\mathcal{S}_0$ is the set of equilibria of (15). The eigenvalues normal to $\mathcal{S}_0$ are $\lambda_\pm(p) = \pm\sqrt{W}$. In particular, the slow manifold $\mathcal{S}_0$ is normally hyperbolic, and hence, it persists for system (14) for $\epsilon > 0$ small (see [4]).

The limiting slow dynamic on $\mathcal{S}_0$ is governed by system (13), which reads

$$\dot\phi = \frac{\beta J_2 - \alpha J_1}{w}, \quad \dot w = -\alpha\beta(J_1 + J_2), \quad \dot J_i = 0, \quad \dot\tau = 1.$$

The general solution is characterized as follows: J_1 and J_2 are arbitrary constants, and

$$\begin{aligned} &\tau(x) = \tau_0 + x, \quad w(x) = \alpha_0 - \alpha\beta(J_1 + J_2)x, \\ &\phi(x) = \phi_0 - \frac{\beta J_2 - \alpha J_1}{\alpha\beta(J_1 + J_2)} \ln\left(1 - \frac{\alpha\beta(J_1 + J_2)}{\alpha_0} x\right), \end{aligned} \tag{16}$$

where $\tau_0 = \tau(0)$, $\phi(0) = \phi_0$, and $w(0) = \alpha_0$. Note that if $J_1 + J_2 = 0$, then $w(x) = \alpha_0$ and $\phi(x) = \phi_0 + (\beta J_2 - \alpha J_1)x/\alpha_0$. The latter is the limit of $\phi(x)$ in (16) as $J_1 + J_2 \to 0$. We thus use the unified formula (16) even if $J_1 + J_2 = 0$.

To identify the slow portion of the singular orbit on $\mathcal{S}_0$, we need to examine the ω-limit (resp., the α-limit) set of $M_L^\epsilon \cap W^s(\mathcal{S}_0)$ (resp., $M_R^\epsilon \cap W^u(\mathcal{S}_0)$) as $\epsilon \to 0$. To do this, we fix an $O(1)$-neighborhood of $\mathcal{S}_0$. In terms of U and V, this neighborhood is of order $O(\epsilon)$. For $\epsilon > 0$ small, the time taken in terms of ξ for M_L^ϵ and M_R^ϵ to evolve to any $O(\epsilon)$-neighborhood of $\{U = V = 0\}$ is of order $O(\epsilon|\ln\epsilon|)$. Thus, the λ-lemma (see [3]) implies that M_L^ϵ (resp., M_R^ϵ) is C^1 $O(\epsilon)$-close to M_L^0 (resp., M_R^0) in any $O(\epsilon)$-neighborhood of $\{U = V = 0\}$. Therefore, in an $O(1)$-neighborhood of $\mathcal{S}_0$ in terms of P and Q, M_L^ϵ (resp., M_R^ϵ) intersects $W^s(\mathcal{S}_0)$ (resp., $W^u(\mathcal{S}_0)$) transversally. And, by abusing the notation, if $N_L = M_L^0 \cap W^s(\mathcal{S}_0)$ and $N_R = M_R^0 \cap W^u(\mathcal{S}_0)$, then $\omega(N_L)$ and $\alpha(N_R)$ have the same descriptions as those in Proposition 2.1 with $U = V = 0$ replaced by $P = (\beta I_2 - \alpha I_1)/W$ and $Q = 0$.

The slow orbit should be one given by (16) that connects $\omega(N_L)$ and $\alpha(N_R)$. Let $\bar M_L$ (resp., $\bar M_R$) be the forward (resp., backward) image of $\omega(N_L)$ (resp., $\alpha(N_R)$) under the slow flow (13).

PROPOSITION 2.3. *$\bar{M}_L$ and $\bar{M}_R$ intersect transversally along the unique orbit given by* (16) *from $x = 0$ to $x = 1$ with*

$$\tau_0 = 0,\ \alpha_0 = (\alpha+\beta)(\alpha L_1)^{\frac{\beta}{\alpha+\beta}}(\beta L_2)^{\frac{\alpha}{\alpha+\beta}},\ \phi_0 = v_0 + \frac{1}{\alpha+\beta}\ln\frac{\alpha L_1}{\beta L_2},$$

$$J_1 = \frac{\left(\ln\frac{R_1}{L_1} - \alpha v_0\right)\left((\alpha L_1)^{\frac{\beta}{\alpha+\beta}}(\beta L_2)^{\frac{\alpha}{\alpha+\beta}} - (\alpha R_1)^{\frac{\beta}{\alpha+\beta}}(\beta R_2)^{\frac{\alpha}{\alpha+\beta}}\right)}{\frac{\alpha\beta}{\alpha+\beta}\ln\frac{R_1}{L_1} + \frac{\alpha^2}{\alpha+\beta}\ln\frac{R_2}{L_2}},$$

$$J_2 = \frac{\left(\ln\frac{R_2}{L_2} + \beta v_0\right)\left((\alpha L_1)^{\frac{\beta}{\alpha+\beta}}(\beta L_2)^{\frac{\alpha}{\alpha+\beta}} - (\alpha R_1)^{\frac{\beta}{\alpha+\beta}}(\beta R_2)^{\frac{\alpha}{\alpha+\beta}}\right)}{\frac{\beta^2}{\alpha+\beta}\ln\frac{R_1}{L_1} + \frac{\alpha\beta}{\alpha+\beta}\ln\frac{R_2}{L_2}}.$$

Proof. We show first that $\bar{M}_L$ and $\bar{M}_R$ intersect along the orbit with the above characterization. In view of (16) and the descriptions for $\omega(N_L)$ and $\alpha(N_R)$ in Proposition 2.1, the intersection is uniquely determined by

$$\tau_0 = 0,\ \alpha_0 = w(0) = (\alpha+\beta)(\alpha L_1)^{\frac{\beta}{\alpha+\beta}}(\beta L_2)^{\frac{\alpha}{\alpha+\beta}},$$

$$w(1) = (\alpha+\beta)(\alpha R_1)^{\frac{\beta}{\alpha+\beta}}(\beta R_2)^{\frac{\alpha}{\alpha+\beta}},$$

$$\phi_0 = \Phi(0) = v_0 + \frac{1}{\alpha+\beta}\ln\frac{\alpha L_1}{\beta L_2},\ \Phi(1) = \frac{1}{\alpha+\beta}\ln\frac{\alpha R_1}{\beta R_2}.$$

Substituting into (16) gives

$$J_1 + J_2 = \frac{\alpha+\beta}{\alpha\beta}\left((\alpha L_1)^{\frac{\beta}{\alpha+\beta}}(\beta L_2)^{\frac{\alpha}{\alpha+\beta}} - (\alpha R_1)^{\frac{\beta}{\alpha+\beta}}(\beta R_2)^{\frac{\alpha}{\alpha+\beta}}\right),$$

$$\beta J_2 - \alpha J_1 = \frac{(\alpha+\beta)\left((\alpha L_1)^{\frac{\beta}{\alpha+\beta}}(\beta L_2)^{\frac{\alpha}{\alpha+\beta}} - (\alpha R_1)^{\frac{\beta}{\alpha+\beta}}(\beta R_2)^{\frac{\alpha}{\alpha+\beta}}\right)}{\frac{\beta}{\alpha+\beta}\ln\frac{R_1}{L_1} + \frac{\alpha}{\alpha+\beta}\ln\frac{R_2}{L_2}} \times\left(v_0 + \frac{1}{\alpha+\beta}\ln\frac{L_1R_2}{L_2R_1}\right),$$

which in turn yields the expressions for J_1 and J_2. To see the transversality of the intersection, it suffices to show that $\omega(N_L)\cdot 1$ (the image of $\omega(N_L)$ under the time one map of the flow of system (13)) is transversal to $\alpha(N_R)$ on $\mathcal{S}_0 \cap \{\tau = 1\}$. If we use (ϕ, w, J_1, J_2) as a coordinate system on $\mathcal{S}_0 \cap \{\tau = 1\}$, then the set $\omega(N_L)\cdot 1$ is given by $\{(\phi(J_1,J_2), w(J_1,J_2), J_1, J_2)\}$ with

$$\phi(J_1,J_2) = v_0 + \frac{1}{\alpha+\beta}\ln\frac{\alpha L_1}{\beta L_2} - \frac{\beta J_2 - \alpha J_1}{\alpha\beta(J_1+J_2)}\ln\left(1 - \frac{\alpha\beta(J_1+J_2)}{\alpha_0}\right),$$
$$w(J_1,J_2) = (\alpha+\beta)(\alpha L_1)^{\frac{\beta}{\alpha+\beta}}(\beta L_2)^{\frac{\alpha}{\alpha+\beta}} - \alpha\beta(J_1+J_2).$$

Thus, the tangent space to $\omega(N_L)\cdot 1$ restricted on $\mathcal{S}_0 \cap \{\tau = 1\}$ is spanned by $(\phi_{J_1}, w_{J_1}, 1, 0) = (\phi_{J_1}, -\alpha\beta, 1, 0)$ and $(\phi_{J_2}, w_{J_2}, 0, 1) = (\phi_{J_2}, -\alpha\beta, 0, 1)$. In view of the display in Proposition 2.1, the tangent space to $\alpha(N_R)$ restricted on $\mathcal{S}_0 \cap \{\tau = 1\}$ is spanned by $(0,0,1,0)$ and $(0,0,0,1)$. Note that $\mathcal{S}_0 \cap \{\tau = 1\}$ is four-dimensional.

Thus, it suffices to show that the above four vectors are linearly independent or, equivalently, $\phi_{J_1} \neq \phi_{J_2}$. The latter can be verified by a direct computation. Indeed, if $J_1 + J_2 \neq 0$ at the intersection points, then

$$\phi_{J_1} - \phi_{J_2} = \frac{\alpha + \beta}{\alpha\beta(J_1 + J_2)} \ln\left(1 - \frac{\alpha\beta(J_1 + J_2)}{\alpha_0}\right) \neq 0;$$

if $J_1 + J_2 = 0$ at the intersection points, then $\phi(J_1, J_2) = \phi_0 + (\beta J_2 - \alpha J_1)/\alpha_0$ and hence $\phi_{J_1} - \phi_{J_2} = -(\alpha + \beta)/\alpha_0 \neq 0$. □

3. Main result. Based on the study of the limiting behavior of boundary layers and regular layers in the previous section, we can easily construct a singular orbit (zeroth order approximation) of the boundary value problem. To show that there indeed exists a true solution near the singular orbit, we apply the exchange lemma to show M_L^ϵ and M_R^ϵ intersect around the singular orbit.

We now state the existence and uniqueness result of the boundary value problem, which also provides the description of a singular orbit.

THEOREM 3.1. *Assume that $\alpha L_1 \neq \beta L_2$ and $\alpha R_1 \neq \beta R_2$. For $\epsilon > 0$ small, the connecting problem* (4), (8) *has a unique solution near a singular orbit. The singular orbit is the union of two fast orbits of system* (7) *and one slow orbit of system* (13)*; more precisely, with both $I_1 = J_1$ and $I_2 = J_2$ given in Proposition* 2.3*,*

(i) *the fast orbit representing the limiting boundary layer at $x = 0$ lies on $B_L \cap W^s(\mathcal{Z}_0)$ from B_L to $\omega(N_L) \subset \mathcal{Z}_0$, whose starting point has the U-component given by* (10) *in Proposition* 2.1*;*

(ii) *the fast orbit representing the limiting boundary layer at $x = 1$ lies on $B_R \cap W^u(\mathcal{Z}_0)$ from B_R to $\alpha(N_R) \subset \mathcal{Z}_0$, whose starting point has the U-component given by* (11) *in Proposition* 2.1*;*

(iii) *the slow orbit on $\mathcal{S}_0$ connecting the two boundary layers from $x = 0$ to $x = 1$ is displayed in* (16) *together with the quantities in Proposition* 2.3*.*

Proof. The singular orbit which has been studied in sections 2.2 and 2.3 is summarized in (i), (ii), and (iii) of this theorem. It remains to show the existence and uniqueness of a solution near the singular orbit for $\epsilon > 0$. Recall that M_L^ϵ (resp., M_R^ϵ) is the union of all forward (resp., backward) orbits starting from B_L (resp., B_R). It suffices to show that, for $\epsilon > 0$ small, M_L^ϵ and M_R^ϵ intersect transversally with each other around the singular orbit. We note that the assumptions $\alpha L_1 \neq \beta L_2$ and $\alpha R_1 \neq \beta R_2$ imply that the vector field of (4) is not tangent to B_L and B_R and hence, M_L^ϵ and M_R^ϵ are smooth invariant manifolds.

For $\epsilon > 0$ small, the evolutions of M_L^ϵ and M_R^ϵ from B_L and B_R, respectively, to an ϵ-neighborhood of $\mathcal{Z}_0$ along the two boundary layers are governed by system (6). Since, for system (7), M_L^0 and M_R^0 intersect $W^s(\mathcal{Z}_0)$ and $W^u(\mathcal{Z}_0)$ transversally, we have that M_L^ϵ and M_R^ϵ intersect $W^s(\mathcal{Z}_0)$ and $W^u(\mathcal{Z}_0)$ transversally. As discussed in section 2.3, in terms of the blow-up coordinates, M_L^ϵ and M_R^ϵ intersect $W^s(\mathcal{S}_0)$ and $W^u(\mathcal{S}_0)$ transversally for system (14). And, if we denote $N_L = M_L^0 \cap W^s(\mathcal{S}_0)$ and $N_R = M_R^0 \cap W^u(\mathcal{S}_0)$, then the vector field on $\mathcal{S}_0$ is not tangent to $\omega(N_L)$ and $\alpha(N_R)$. Furthermore, the traces $\bar{M}_L$ and $\bar{M}_R$ of $\omega(N_L)$ and $\alpha(N_R)$, respectively, under the slow flow on $\mathcal{S}_0$ intersect transversally. All conditions for the exchange lemma (see [15] and also [10, 8, 9]) are satisfied, and hence, M_L^ϵ and M_R^ϵ intersect transversally. The intersection has dimension

$$\dim M_L^\epsilon + \dim M_R^\epsilon - 7 = 4 + 4 - 7 = 1,$$

which is the orbit of the unique solution for the connecting problem near the singular orbit. □

Remark 3.1. We have considered the situation that $\alpha L_1 \neq \beta L_2$ and $\alpha R_1 \neq \beta R_2$. In the case that $\alpha L_1 = \beta L_2$ or $\alpha R_1 = \beta R_2$, then B_L or B_R is on the slow manifold $\mathcal{S}_0$ and hence there is no boundary layer at $x = 0$ or $x = 1$.

4. Appendix. A derivation of the integral H_3 in Proposition 2.1. The complete set of six integrals of system (7) in Proposition 2.1 is crucial in the quantitative investigation of the boundary layers of the boundary value problem. The integrals H_1 and H_2 are relatively easy to guess. The integral H_3, although easily verified, is discovered through several observations. It may have some general interest, and we provide a formal derivation below.

We divide the W-equation by the V-equation from system (7) to get

$$\frac{dW}{dV} = \frac{\alpha\beta V}{W} + (\beta - \alpha),$$

which is a homogeneous equation of order zero. This leads to the substitution $W = yV$. From $dW = Vdy + ydV$ and the above equation one gets

$$\left(\frac{\alpha\beta V}{yV} + (\beta - \alpha)\right) dV = Vdy + ydV \quad \text{or} \quad -\frac{dV}{V} = \frac{ydy}{y^2 - (\beta - \alpha)y - \alpha\beta}.$$

Integrating both sides, we have, for some constant C,

$$-\ln V + C = \frac{\alpha}{\alpha + \beta} \ln|y + \alpha| + \frac{\beta}{\alpha + \beta} \ln|y - \beta|,$$

or, for some constant D,

$$V = \frac{D}{|y + \alpha|^{\frac{\alpha}{\alpha+\beta}} |y - \beta|^{\frac{\beta}{\alpha+\beta}}}, \quad W = \frac{Dy}{|y + \alpha|^{\frac{\alpha}{\alpha+\beta}} |y - \beta|^{\frac{\beta}{\alpha+\beta}}}.$$

Substitute $y = W/V$ to get

$$|W + \alpha V|^{\alpha} |W - \beta V|^{\beta} = D^{\alpha+\beta}.$$

This completes the derivation of the integral H_3.

Acknowledgments. This work was initiated from the discussions in the seminar on mathematical physiology at the University of Kansas. The author thanks all participants for their interest in this work. The author also thanks the referees for their valuable comments and suggestions on the original manuscript.

REFERENCES

[1] V. Barcilon, D.-P. Chen, and R. S. Eisenberg, *Ion flow through narrow membrane channels: Part* II, SIAM J. Appl. Math., 52 (1992), pp. 1405–1425.

[2] V. Barcilon, D.-P. Chen, R. S. Eisenberg, and J. W. Jerome, *Qualitative properties of steady-state Poisson–Nernst–Planck systems: Perturbation and simulation study*, SIAM J. Appl. Math., 57 (1997), pp. 631–648.

[3] B. Deng, *The Sil'nikov problem, exponential expansion, strong λ-lemma, C^1 linearization and homoclinic bifurcation*, J. Differential Equations, 79 (1989), pp. 189–231.

[4] N. Fenichel, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98.

[5] M. H. Holmes, *Nonlinear ionic diffusion through charged polymeric gels*, SIAM J. Appl. Math., 50 (1990), pp. 839–852.
[6] J. W. Jerome, *Consistency of semiconductor modeling: An existence/stability analysis for the stationary Van Roosbroeck system*, SIAM J. Appl. Math., 45 (1985), pp. 565–590.
[7] J. W. Jerome and T. Kerkhoven, *A finite element approximation theory for the drift diffusion semiconductor model*, SIAM J. Numer. Anal., 28 (1991), pp. 403–422.
[8] C. K. R. T. Jones, *Geometric singular perturbation theory*, in Dynamical Systems (Montecatini Terme, 1994), Lect. Notes in Math. 1609, Springer-Verlag, Berlin, 1995, pp. 44–118.
[9] C. K. R. T. Jones, T. J. Kaper, and N. Kopell, *Tracking invariant manifolds up to exponentially small errors*, SIAM J. Math. Anal., 27 (1996), pp. 558–577.
[10] C. K. R. T. Jones and N. Kopell, *Tracking invariant manifolds with differential forms in singularly perturbed systems*, J. Differential Equations, 108 (1994), pp. 64–88.
[11] J. Keener and J. Sneyd, *Mathematical Physiology*, Interdiscip. Appl. Math. 8, Springer-Verlag, New York, 1998.
[12] W. Liu, *Exchange lemmas for singular perturbations with certain turning points*, J. Differential Equations, 167 (2000), pp. 134-180.
[13] J.-H. Park and J. W. Jerome, *Qualitative properties of steady-state Poisson–Nernst–Planck systems: Mathematical study*, SIAM J. Appl. Math., 57 (1997), pp. 609–630.
[14] I. Rubinstein, *Electro-Diffusion of Ions*, SIAM Stud. Appl. Math. 11, SIAM, Philadelphia, PA, 1990.
[15] S.-K. Tin, N. Kopell, and C. K. R. T. Jones, *Invariant manifolds and singularly perturbed boundary value problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1558–1576.

SIAM J. APPL. MATH.
Vol. 65, No. 3, pp. 767–789

THE DETERMINATION OF THE SURFACE CONDUCTIVITY OF A PARTIALLY COATED DIELECTRIC*

FIORALBA CAKONI†, DAVID COLTON†, AND PETER MONK†

Abstract. A variational method is given for determining the essential supremum of the surface conductivity of a partially coated anisotropic dielectric medium from a knowledge of the far field pattern of the time-harmonic electric field at fixed frequency corresponding to an incident plane wave. It is assumed that the shape of the scatterer has been determined (e.g., by solving the far field equation and using the linear sampling method). Numerical examples are given for the scalar case with constant surface conductivity.

Key words. inverse scattering problem, interior transmission problem, electromagnetic waves, mixed boundary value problems

AMS subject classifications. 35P25, 35R30, 78A45

DOI. 10.1137/040604224

1. Introduction.

In a previous paper in this journal [5], we considered the problem of determining the surface impedance of a perfect conductor that is partially coated with a dielectric from a knowledge of the far field pattern of the scattered electromagnetic wave corresponding to an incident time-harmonic plane wave at fixed frequency. Such problems are the simplest model for detecting hostile objects that have been partially coated with a dielectric in order to avoid detection by using such a coating to reduce the radar cross section of the scattered wave. In [5] it was shown that the solution of the far field equation that determines the shape of the scatterer by means of the linear sampling method [11] can also be used in conjunction with a variational method to determine the essential supremum of the surface impedance of the coated portion of the boundary, and numerical examples were given showing the viability of our method.

In this paper we consider the problem complementary to the one described above; i.e., we now wish to detect a benign object that has been partially coated by a thin conducting material in order to make it appear hostile [6]. An example of this is a wooden decoy in the shape of a tank that is partially coated by metallic paint. The problem is again to determine a coefficient (the surface conductivity) in the boundary condition from a knowledge of the far field pattern of the scattered electromagnetic wave corresponding to an incident time-harmonic plane wave. (The shape of the scatterer can again be determined by the linear sampling method.) However, the problem of determining the surface conductivity is considerably more complicated than the problem of determining the surface impedance of a coated perfect conductor since we now have a mixed boundary value problem for a penetrable obstacle. In particular, we now must consider an interior transmission problem with mixed boundary conditions, and the well-posedness of such problems is unknown.

The plan of our paper is as follows. After formulating the mathematical model for the scattering of time harmonic electromagnetic waves by an anisotropic medium

*Received by the editors February 17, 2004; accepted for publication (in revised form) July 7, 2004; published electronically February 25, 2005. This work was supported in part by Air Force Office of Scientific Research grants F49620-02-1-0071 and F49620-02-1-0353.
http://www.siam.org/journals/siap/65-3/60422.html
†Department of Mathematical Sciences, University of Delaware, Newark, DE 19716 (cakoni@math.udel.edu, colton@math.udel.edu, monk@math.udel.edu).

that is partially coated by a thin conducting layer, we consider the scalar case corresponding to the scattering of electromagnetic waves by an infinite cylinder. We first show that in this case both the direct scattering problem and the interior transmission problem are well posed. We then use these results to derive a variational formula for the determination of the essential supremum of the surface conductivity of the coated portion of the boundary from a knowledge of the far field pattern of the scattered time-harmonic magnetic field. Finally, we derive an analogous formula for determining the surface conductivity in the case of Maxwell's equations in $\mathbb{R}^3$ under the assumption that the interior transmission problem in this case is well posed. We conclude by presenting some numerical examples for the scalar case with constant surface conductivity.

2. Formulation of the direct and inverse scattering problem. We consider the scattering of time-harmonic electromagnetic waves with frequency ω from an infinitely long cylindrical anisotropic dielectric partially coated with a very thin layer of a highly conductive material. We assume that the electric permittivity ϵ_0 and magnetic permeability μ_0 of the exterior dielectric background medium are positive constants, whereas the scatterer has the same magnetic permeability μ_0 as the exterior medium but the electric permittivity ϵ and the conductivity σ are real 3×3 matrix valued functions. After an appropriate scaling [12], the total electric and magnetic fields E, H satisfy the time-harmonic homogeneous Maxwell equations in the exterior of the scatterer,

$$\begin{cases} \nabla \times E - ikH = 0, \\ \nabla \times H + ikE = 0, \end{cases} \tag{2.1}$$

and the interior electric and magnetic fields E_0, H_0 solve the following equations in the interior of the scattering object:

$$\begin{cases} \nabla \times E_0 - ikH_0 = 0, \\ \nabla \times H_0 + ikN(x)E_0 = 0, \end{cases} \tag{2.2}$$

where $k^2 = \epsilon_0\mu_0\omega^2$ and the index of refraction is given by $N(x) = \frac{1}{\epsilon_0}\big(\epsilon(x) + i\frac{\sigma(x)}{\omega}\big)$.

Let the real valued function $\eta > 0$ defined on the coated portion of the boundary of the scatterer describe the physical properties of the highly conductive coating (see [1]). As shown in [8], the tangential component of the electric field is continuous across the boundary

$$\nu \times E - \nu \times E_0 = 0, \tag{2.3}$$

while the tangential component of the magnetic field is continuous only on the uncoated part of the boundary

$$\nu \times H - \nu \times H_0 = 0 \tag{2.4}$$

and satisfies the following relation on the coated part of the boundary:

$$\nu \times H - \nu \times H_0 = \eta(x)(\nu \times E) \times \nu. \tag{2.5}$$

The exterior field E, H is given by

$$E = E^i + E^s, \qquad H = H^i + H^s, \tag{2.6}$$

where E^s, H^s is the scattered field satisfying the Silver–Müller radiation condition at infinity [12] and E^i, H^i is the given incident field.

Now we assume that the scatterer is an infinitely long cylinder with axis in the z-direction and that the incident electromagnetic field is a plane wave propagating in the direction perpendicular to the cylinder. Let the bounded domain $D \subset \mathbb{R}^2$ with Lipschitz boundary Γ be the cross section of the cylinder such that the exterior domain $D_e := \mathbb{R}^2 \setminus \overline{D}$ is connected. We denote by ν the outward unit normal to Γ defined almost everywhere on Γ. The boundary $\Gamma = \Gamma_1 \cup \Pi \cup \Gamma_2$ is split into two open disjoint parts Γ_1 and Γ_2 having Π as their possible common boundary in Γ. Here Γ_1 corresponds to the uncoated part and Γ_2 corresponds to the coated part. We assume that the dielectric is orthotropic; i.e., the matrix N is of the form

$$N = \begin{pmatrix} n_{11} & n_{12} & 0 \\ n_{21} & n_{22} & 0 \\ 0 & 0 & n_{33} \end{pmatrix},$$

and the functions N and η do not depend on z. If we consider incident waves such that the electric field is polarized perpendicular to the z-axis, then the magnetic fields have a component in only the z-direction, i.e.,

$$H^i = (0,\ 0,\ u^i), \quad H_0 = (0,\ 0,\ v), \quad H^s = (0,\ 0,\ u^s).$$

Assuming that N^{-1} exists and expressing the electric fields in terms of magnetic fields, (2.1)–(2.6) now lead to the following transmission problem for v and u:

$$\begin{aligned}
&\text{(i)} && \nabla\cdot A\nabla v + k^2 v = 0 && \text{in } D,\\
&\text{(ii)} && \Delta u + k^2 u = 0 && \text{in } D_e,\\
&\text{(iii)} && v - u = 0 && \text{on } \Gamma_1,\\
&\text{(iv)} && v - u = -i\eta \frac{\partial u}{\partial \nu} && \text{on } \Gamma_2,\\
&\text{(v)} && \frac{\partial v}{\partial \nu_A} - \frac{\partial u}{\partial \nu} = 0 && \text{on } \Gamma,\\
&\text{(vi)} && u = u^s + u^i, &&\\
&\text{(vii)} && \lim_{r\to\infty} \sqrt{r}\left(\frac{\partial u^s}{\partial r} - iku^s\right) = 0, &&
\end{aligned} \tag{2.7}$$

$r = |x|$, where u^s is the scattered field and u^i is the given incident field. In the case of plane waves the incident field is given by $u^i := e^{ikx\cdot d}$, $d \in \Omega := \{x : |x| = 1\}$. Moreover,

$$\frac{\partial v}{\partial \nu_A}(x) := \nu(x) \cdot A(x)\nabla v(x), \qquad x \in \Gamma,$$

$$A = \frac{1}{n_{11}n_{22} - n_{12}n_{21}} \begin{pmatrix} n_{11} & n_{21} \\ n_{12} & n_{22} \end{pmatrix},$$

and the radiation condition (2.7)(vii) holds uniformly with respect to $\hat{x} = x/|x|$. Note that A is not the inverse of a 2×2 submatrix N but rather comes from substituting $H_0 = (0, 0, v)$ into (2.2).

In the following we assume that A is a 2×2 matrix valued function whose entries are continuously differentiable functions in $\overline{D}$ such that A is symmetric, $\mathcal{R}e(\,\bar{\xi} \cdot A\,\xi\,) \geq \gamma|\xi|^2$, and $\mathcal{I}m(\,\bar{\xi} \cdot A\,\xi\,) \leq 0$ for all $\xi \in \mathbb{C}^2$ and $x \in \overline{D}$, where γ is a positive constant. Note that, due to the symmetry of A, we have $\mathcal{R}e(\,\bar{\xi} \cdot A\,\xi\,) = \bar{\xi} \cdot \mathcal{R}e(A)\,\xi$ and $\mathcal{I}m(\,\bar{\xi} \cdot A\,\xi\,) = \bar{\xi} \cdot \mathcal{I}m(A)\,\xi$. Moreover, we require that $\eta \in L_\infty(\Gamma_2)$ and $\eta(x) \geq \eta_0 > 0$ for all $x \in \Gamma_2$.

Let $H^1(D)$ and $H^1_{\rm loc}(D_e)$ denote the usual Sobolev spaces and $H^{\frac{1}{2}}(\Gamma)$ the corresponding trace space. For $\Gamma_2 \subset \Gamma$ we define

$$H^{\frac{1}{2}}(\Gamma_2) := \{u|_{\Gamma_2} : u \in H^{\frac{1}{2}}(\Gamma)\},$$

$$\tilde{H}^{\frac{1}{2}}(\Gamma_2) := \{u \in H^{\frac{1}{2}}(\Gamma_2) : \operatorname{supp} u \subseteq \overline{\Gamma}_2\},$$

and denote by $H^{-\frac{1}{2}}(\Gamma_2)$ and $\tilde{H}^{-\frac{1}{2}}(\Gamma_2)$ the dual spaces $(\tilde{H}^{\frac{1}{2}}(\Gamma_2))'$ and $(H^{\frac{1}{2}}(\Gamma_2))'$, respectively, with L^2 as a pivot space (for details, see [16]). We recall that a function in $\tilde{H}^{\frac{1}{2}}(\Gamma_2)$ and $\tilde{H}^{-\frac{1}{2}}(\Gamma_2)$ can be extended by zero to a function in $H^{\frac{1}{2}}(\Gamma)$ and $H^{-\frac{1}{2}}(\Gamma)$, respectively. Note that for $u \in H^1(D)$ with $\Delta u \in L^2(D)$ the trace $\frac{\partial u}{\partial \nu} \in H^{-\frac{1}{2}}(\Gamma)$ is well defined.

For later use we also define the Hilbert space

$$\mathbb{H}^1(D, \Gamma_2) := \left\{ u \in H^1(D) \quad \text{such that} \quad \frac{\partial u}{\partial \nu} \in L^2(\Gamma_2) \right\}$$

equipped with the usual graph norm

$$\|u\|^2_{\mathbb{H}^1(D,\Gamma_2)} := \|u\|^2_{H^1(D)} + \left\| \frac{\partial u}{\partial \nu} \right\|^2_{L^2(\Gamma_2)}.$$

The *forward scattering problem* reads: Given D, A, η as above and the incident field $u^i \in H^1_{\rm loc}(\mathbb{R}^2)$, find $v \in H^1(D)$ and $u \in H^1_{\rm loc}(D_e)$ that satisfy (2.7), where the boundary conditions are assumed in the sense of the trace operator. In what follows, we refer to this mixed transmission problem as (MTP).

It is known [12] that solutions of the Helmholtz equation that satisfy the Sommerfeld radiation condition (2.7)(vi) have the asymptotic behavior

$$u^s(x) = \frac{e^{ikr}}{\sqrt{r}} u_\infty(\hat{x}) + O(r^{-3/2}), \quad r \to \infty, \tag{2.8}$$

where $u_\infty(\hat{x})$ is the *far field pattern* of the radiating solution u^s. In the case of incident plane waves, $u_\infty(\hat{x})$ depends on the incident direction d, which we indicate by $u_\infty(\hat{x}, d)$. The *inverse scattering problem* we are concerned with is to determine D and η from a knowledge of the far field pattern $u_\infty(\hat{x}, d)$ of the scattered field u^s for $\hat{x}$, $-d \in \Omega_0$, where Ω_0 is a subset of the unit circle Ω. Note that no a priori knowledge of the amount of coating is required.

3. The direct scattering problem. First we want to show that the mixed transmission problem (2.7) is well posed.

LEMMA 3.1. *The problem* (MTP) *has at most one solution.*

Proof. Let $v \in H^1(D)$ and $u \in H^1_{\rm loc}(D_e)$ be the solution of (2.7) corresponding to the incident wave $u^i \equiv 0$. Applying Green's formula in D and $D_e \cap B_R$, where B_R is a disk of radius R containing D, and using the transmission conditions, we have

$$\begin{aligned}
&\int_D \left(\nabla \overline{v} \cdot A \nabla v - k^2 |v|^2\right) dy + \int_{D_e \cap B_R} \left(|\nabla u|^2 - k^2 |u|^2\right) dy \\
&\quad = \int_\Gamma \overline{v} \cdot \frac{\partial v}{\partial \nu_A} \, ds - \int_\Gamma \overline{u} \cdot \frac{\partial u}{\partial \nu} \, ds + \int_{S_R} \overline{u} \cdot \frac{\partial u}{\partial \nu} \, ds \\
&\quad = i \int_{\Gamma_2} \frac{1}{\eta} |v - u|^2 \, ds + \int_{S_R} \overline{u} \cdot \frac{\partial u}{\partial \nu} \, ds.
\end{aligned}$$

Now taking the imaginary part of both sides and using the fact that $\mathcal{I}m(A) \leq 0$ is a real valued matrix and $\eta \geq \eta_0 > 0$, we obtain

$$\mathcal{I}m \int_{S_R} u \cdot \frac{\partial \overline{u}}{\partial \nu} \, ds \geq 0.$$

Finally, an application of Rellich's lemma and the unique continuation principle yield $u = v = 0$. ∎

In order to give a variational formulation of the problem (MTP) we introduce the Dirichlet-to-Neumann map $\Lambda : H^{\frac{1}{2}}(S_R) \to H^{-\frac{1}{2}}(S_R)$, which maps $h \in H^{\frac{1}{2}}(S_R)$ to $\frac{\partial \tilde{u}}{\partial \nu}$, where $\tilde{u}$ solves the exterior Dirichlet problem for the Helmholtz equation in $\mathbb{R}^2 \backslash \overline{B}_R$ with Dirichlet boundary data h. The following result is known [12], [15].

LEMMA 3.2. *There exists an operator $\Lambda_0 : H^{\frac{1}{2}}(S_R) \to H^{-\frac{1}{2}}(S_R)$ such that*

$$\int_{S_R} \overline{\varphi} \, \Lambda_0 \varphi \, ds \leq 0 \tag{3.1}$$

and $\Lambda - \Lambda_0$ is a compact operator from $H^{\frac{1}{2}}(S_R)$ to $H^{-\frac{1}{2}}(S_R)$.

Integrating by parts the equations of (MTP) with a test function φ, we can put (MTP) into the following variational form: Find $w \in H^1(B_R \backslash \overline{\Gamma}_2)$ such that

$$\begin{aligned}
&\int_D \left(\nabla \overline{\varphi} \cdot A \nabla w - k^2 \overline{\varphi} \, w\right) dy + \int_{D_e \cap B_R} \left(\nabla \overline{\varphi} \cdot \nabla w - k^2 \overline{\varphi} \, w\right) dy \\
&\quad - \int_{\Gamma_2} \frac{i}{\eta} \, [\overline{\varphi}] \cdot [w] \, ds - \int_{S_R} \overline{\varphi} \, \Lambda w \, ds = - \int_{S_R} \overline{\varphi} \, \Lambda u^i \, ds + \int_{S_R} \overline{\varphi} \frac{\partial u^i}{\partial \nu} \, ds
\end{aligned} \tag{3.2}$$

for any function $\varphi \in H^1(B_R \backslash \overline{\Gamma}_2)$, where $[u] = u^+|_{\Gamma_2} - u^-|_{\Gamma_2}$ denotes the jump of u across Γ_2. Note that for $u \in H^1(B_R \backslash \overline{\Gamma}_2)$ the jump $[u] \in \tilde{H}^{\frac{1}{2}}(\Gamma_2)$. Let us denote by $\mathcal{A}_1$ and $\mathcal{A}_2$ the following sesquilinear forms:

$$\begin{aligned}
\mathcal{A}_1(w, \varphi) := &\int_D \left(\nabla \overline{\varphi} \cdot A \nabla w + \overline{\varphi} \, w\right) dy + \int_{D_e \cap B_R} \left(\nabla \overline{\varphi} \cdot \nabla w + \overline{\varphi} \, w\right) dy \\
&- \int_{\Gamma_2} \frac{i}{\eta} \, [\overline{\varphi}] \cdot [w] \, ds - \int_{S_R} \overline{\varphi} \, \Lambda_0 w \, ds
\end{aligned} \tag{3.3}$$

and

$$\mathcal{A}_2(w, \varphi) := - \int_{B_R} (k^2 + 1) \, \overline{\varphi} \, w \, dy - \int_{S_R} \overline{\varphi} \, (\Lambda_0 - \Lambda) w \, ds,$$

respectively. Then (3.2) becomes the following: Find $w \in H^1(B_R \backslash \overline{\Gamma}_2)$ such that

$$\mathcal{A}_1(w,\varphi) + \mathcal{A}_2(w,\varphi) = L(\varphi) \qquad \forall \varphi \in H^1(B_R \backslash \overline{\Gamma}_2), \tag{3.4}$$

where $L(\varphi)$ denotes the continuous antilinear form defined by the right-hand side of (3.2). Obviously if w is a solution of (3.4), then $v := w|_D$ and $u := w|_{B_R \cap D_e}$ satisfy the differential equations and the transmission conditions of (MTP). Then using Green's formula and the radiation condition, one can extend $w = u - u^i$ to a radiating solution of the Helmholtz equation in the exterior domain D_e (see, e.g., [15]).

Next we want to show that there exists a function $w \in H^1(B_R \backslash \overline{\Gamma}_2)$ that satisfies (3.4). The uniqueness of (3.4) is equivalent to the uniqueness of a solution to (MTP) (see Lemma 3.1). Note that, due to (2.7(iv)) and (2.7(v)), if $u \in H^1(D)$ and $v \in H^1_{\rm loc}(D^e)$ solve (2.7), then $w \in H^1(B_R \backslash \overline{\Gamma}_2)$. Using the classical trace theorems and Cauchy–Schwarz inequality, the chain of continuous imbeddings

$$\tilde{H}^{\frac{1}{2}}(\Gamma_2) \subset H^{\frac{1}{2}}(\Gamma_2) \subset L^2(\Gamma_2) \subset \tilde{H}^{-\frac{1}{2}}(\Gamma_2) \subset H^{-\frac{1}{2}}(\Gamma_2),$$

and the boundedness of A and η, we obtain

$$|\mathcal{A}_1(w,\varphi)| \le C_1 \|w\|_{H^1(B_R \backslash \overline{\Gamma}_2)} \|\varphi\|_{H^1(B_R \backslash \overline{\Gamma}_2)}$$

with $C_1 > 0$ independent of w and φ. Hence $\mathcal{A}_1$ is bounded. Furthermore, from the fact that $\mathcal{R}e(A)$ is positive definite together with Lemma 3.2, we obtain the following coercivity result:

$$\mathcal{R}e\left(\mathcal{A}_1(w,w)\right) \ge C_2 \|w\|^2_{H^1(B_R \backslash \overline{\Gamma}_2)},$$

where the constant $C_2 > 0$ does not depend on w.

Next, based on the Riesz representation theorem, we define an operator $K : H^1(B_R \backslash \overline{\Gamma}_2) \to H^1(B_R \backslash \overline{\Gamma}_2)$ by

$$(Kw,\ \varphi) = \mathcal{A}_2(w,\ \varphi) \qquad \forall w, \varphi \in H^1(B_R \backslash \overline{\Gamma}_2).$$

The compact embedding of $H^1(B_R \setminus \overline{\Gamma})$ into $L^2(B_R)$ and the compactness of the operator $\Lambda - \Lambda_0$ from Lemma 3.2 imply that the operator K is compact.

The above analysis shows that the Fredholm alternative can be applied to (3.2), which, together with the uniqueness of a solution to (3.2), implies the solvability of (3.2) and therefore the solvability of (2.7). Summarizing the above analysis, we have proved the following theorem.

THEOREM 3.3. *For any incident field $u^i \in H^1_{\rm loc}(\mathbb{R}^2)$ there exists a unique solution $(v,u) \in H^1(D) \times H^1_{\rm loc}(D_e)$ of* (MTP) *which depends continuously on u^i.*

3.1. The interior transmission problem. As will be seen in what follows, an important role in solving the inverse problem of determining D and η is played by the *interior transmission problem*: Given $f \in H^{\frac{1}{2}}(\Gamma)$, $h \in H^{-\frac{1}{2}}(\Gamma)$, and $r \in L^2(\Gamma_2)$, find $v \in H^1(D)$ and $w \in \mathbb{H}^1(D,\Gamma_2)$ such that

$$\begin{aligned}
&\text{(i)} && \nabla\cdot A\nabla v + k^2 v = 0 && \text{in } D,\\
&\text{(ii)} && \Delta w + k^2 w = 0 && \text{in } D,\\
&\text{(iii)} && v - w = f|_{\Gamma_1} && \text{on } \Gamma_1,\\
&\text{(iv)} && v - w = -i\eta \frac{\partial w}{\partial \nu} + f|_{\Gamma_2} + r && \text{on } \Gamma_2,\\
&\text{(v)} && \frac{\partial v}{\partial \nu_A} - \frac{\partial w}{\partial \nu} = h && \text{on } \Gamma.
\end{aligned} \tag{3.5}$$

In the remainder of the paper we will refer to (3.5) as (IMTP). The well-posedness of the interior transmission problem in the case when $\eta \equiv 0$ and $r = 0$ is established in [7]. Here we will adapt the variational approach used in [7] to our mixed transmission case. In order to avoid repetition we will only sketch the proof, emphasizing the changes due to the boundary terms involving η. We first modify (IMTP) to

$$\begin{aligned}
&\text{(i)} && \nabla\cdot A\nabla v - m\, v = \ell_1 && \text{in } D,\\
&\text{(ii)} && \Delta w - w = \ell_2 && \text{in } D,\\
&\text{(iii)} && v - w = f|_{\Gamma_1} && \text{on } \Gamma_1,\\
&\text{(iv)} && v - w = -i\eta\,\frac{\partial w}{\partial \nu} + f|_{\Gamma_2} + r && \text{on } \Gamma_2,\\
&\text{(v)} && \frac{\partial v}{\partial \nu_A} - \frac{\partial w}{\partial \nu} = h && \text{on } \Gamma,
\end{aligned} \tag{3.6}$$

where $m > 0$, $\ell_1 \in L^2(D)$, and $\ell_2 \in L^2(D)$. We will now reformulate (3.6) as an equivalent variational problem. To this end let

$$\mathbb{W}(D) = \{\mathbf{w} \in L^2(D)^2 \,:\, \nabla\cdot \mathbf{w} \in L^2(D), \text{ and } \operatorname{curl} \mathbf{w} = 0 \text{ and } \nu\cdot\mathbf{w}|_{\Gamma_2} \in L^2(\Gamma_2)\}$$

equipped with the natural norm

$$\|\mathbf{w}\|^2_{\mathbb{W}} = \|\mathbf{w}\|^2_{L^2} + \|\nabla\cdot\mathbf{w}\|^2_{L^2} + \|\nu\cdot\mathbf{w}\|^2_{L^2},$$

and denote by $\langle\cdot\,,\cdot\rangle$ the duality pairing between $H^{\frac{1}{2}}(\Gamma)$ and $H^{-\frac{1}{2}}(\Gamma)$. We also introduce the duality identity

$$\langle \varphi,\, \boldsymbol{\psi}\cdot\nu\rangle = \int_D \varphi\, \nabla\cdot\boldsymbol{\psi}\, dx + \int_D \nabla\varphi\cdot\boldsymbol{\psi}\, dx \tag{3.7}$$

for $(\varphi,\, \boldsymbol{\psi}) \in H^1(D)\times\mathbb{W}(D)$.

By doing exactly the same as in the proof of Theorem 3.3 in [7], one can show that the modified interior transmission problem (3.6) is equivalent to the following variational problem: Find $V = (v, \mathbf{w}) \in H^1(D)\times\mathbb{W}(D)$ such that

$$\mathcal{A}(V,\Psi) = L(\Psi), \qquad \Psi \in H^1(D)\times\mathbb{W}(D), \tag{3.8}$$

where the sesquilinear form $\mathcal{A}$ defined in $(H^1(D)\times\mathbb{W}(D))^2$ is given by

$$\begin{aligned}
\mathcal{A}(V,\Psi) = &\int_D A\nabla v\cdot\nabla\bar\varphi\, dx + \int_D m\, v\,\bar\varphi\, dx + \int_D \nabla\cdot\mathbf{w}\,\nabla\cdot\bar{\boldsymbol{\psi}}\, dx + \int_D \mathbf{w}\cdot\bar{\boldsymbol{\psi}}\, dx\\
&- i\int_{\Gamma_2} \eta\,(\mathbf{w}\cdot\nu)\,(\bar{\boldsymbol{\psi}}\cdot\nu), ds - \left\langle v,\, \bar{\boldsymbol{\psi}}\cdot\nu\right\rangle - \langle\bar\varphi,\, \mathbf{w}\cdot\nu\rangle
\end{aligned} \tag{3.9}$$

and the antilinear form L is given by

$$L(\Psi) = \int_D (\ell_1\,\bar\varphi + \ell_2\,\nabla\cdot\bar{\boldsymbol{\psi}})\, dx - i\int_{\Gamma_2} \eta\, r\,(\bar{\boldsymbol{\psi}}\cdot\nu) + \langle\bar\varphi,\, h\rangle - \left\langle f,\, \bar{\boldsymbol{\psi}}\cdot\nu\right\rangle.$$

The modified interior transmission problem (3.6) has a unique solution $(v, w) \in H^1(D)\times\mathbb{H}^1(D,\Gamma_2)$ if and only if the variational problem (3.8) has a unique solution $V \in H^1(D)\times W(D)$. If (v, w) is the unique solution (3.6), then $V = (v, \nabla w)$ is a

unique solution to (3.8). Conversely if V is the unique solution to (3.8), then the unique solution (v, w) to (3.6) is such that $V = (v, \nabla w)$.

Now assume that there exists a constant $\gamma > 1$ such that $\bar{\xi} \cdot \mathcal{R}e(A)\,\xi \geq \gamma|\xi|^2$ and choose $m > 1$. Classical trace theorems and Schwarz's inequality ensure the continuity of the sesquilinear form $\mathcal{A}$ and the antilinear form L. On the other hand, by taking the real and the imaginary part of $\mathcal{A}(V, V)$, we have from the assumptions on $\mathcal{R}e(A)$, $\mathcal{I}m(A)$, and η that

$$|\mathcal{A}(V,V)| \geq \gamma\|v\|^2_{H^1(D)} + \|\mathbf{w}\|^2_{L^2(D)} + \|\nabla\cdot\mathbf{w}\|^2_{L^2(D)} - 2\mathcal{R}e(\langle\bar{v},\, \nu\cdot\mathbf{w}\rangle) + \eta_0\|\nu\cdot\mathbf{w}\|^2_{L^2(\Gamma_2)}.$$

From the duality identity (3.7) and Schwarz's inequality we have

$$2\mathcal{R}e(\langle\bar{v},\, \nu\cdot\mathbf{w}\rangle) \leq |\,\langle\bar{v},\, \mathbf{w}\rangle\,| \leq \|v\|_{H^1(D)}(\|\mathbf{w}\|^2_{L^2(D)} + \|\nabla\cdot\mathbf{w}\|^2_{L^2(D)})^{\frac{1}{2}}.$$

Hence since $\gamma > 1$, we conclude that

$$|\mathcal{A}(V,V)| \geq \frac{\gamma-1}{\gamma+1}(\|v\|^2_{H^1(D)} + \|\mathbf{w}\|^2_{L^2(D)} + \|\nabla\cdot\mathbf{w}\|^2_{L^2(D)}) + \eta_0\|\nu\cdot\mathbf{w}\|^2_{L^2(\Gamma_2)},$$

which means that $\mathcal{A}$ is coercive; i.e.,

$$|\mathcal{A}(V,V)| \geq C(\|v\|^2_{H^1(D)} + \|\mathbf{w}\|^2_{W(D)}),$$

where $C = \min((\gamma-1)/(\gamma+1),\, \eta_0)$. Therefore from the Lax–Milgram theorem we have that the variational problem (3.8) is uniquely solvable, whence the modified interior transmission problem has a unique solution (u, v) that satisfies

$$\|v\|_{H^1(D)} + \|w\|_{\mathbb{H}^1(D,\Gamma_2)} \leq C(\|f\|_{H^{\frac{1}{2}}(\Gamma)} + \|h\|_{H^{-\frac{1}{2}}(\Gamma)} + \|r\|_{L^2(\Gamma_2)}),$$

where $C > 0$ is independent on f, h, r.

THEOREM 3.4. *Assume that* $\bar{\xi} \cdot \mathcal{R}e(A)\,\xi \geq \gamma|\xi|^2$ *with* $\gamma > 1$ *and* $\eta(x) \geq \eta_0 > 0$. *Then the Fredholm alternative can be applied to the problem* (IMTP).

Proof. Let us define

$$\mathcal{Y}(D) := \big\{(v,w) \in H^1(D)\times\mathbb{H}^1(D,\Gamma_2)\ :\ \nabla\cdot A\nabla v \in L^2(D) \text{ and } \Delta w \in L^2(D)\big\}$$

and consider the operator $\mathcal{G}$ from $\mathcal{Y}(D)$ into $L^2(D)\times L^2(D)\times H^{\frac{1}{2}}(\Gamma_1)\times L^2(\Gamma_2)\times H^{-\frac{1}{2}}(\Gamma)$ defined by

$$\mathcal{G}(v,w) = \left\{\nabla\cdot A\nabla v - mv,\ \Delta w - w,\ (v-w)|_{\Gamma_1},\ \left(v - w + i\eta\,\frac{\partial w}{\partial\nu}\right)_{\Gamma_2},\ \left(\frac{\partial v}{\partial\nu_A} - \frac{\partial w}{\partial\nu}\right)_{\Gamma}\right\},$$

where $m > 1$. We have shown that the inverse of $\mathcal{G}$ exists and is continuous. Since $\mathcal{G}$ is continuous, we deduce that $\mathcal{G}$ is a bijective operator. Now consider the operator $\mathcal{T}$ from $\mathcal{Y}(D)$ into $L^2(D)\times L^2(D)\times H^{\frac{1}{2}}(\Gamma_1)\times L^2(\Gamma_2)\times H^{-\frac{1}{2}}(\Gamma)$ defined by

$$\mathcal{T}(v,w) = \{(k^2+m)v,\ (k^2+1)w,\ 0,\ 0,\ 0\}.$$

By the compact embedding of $H^1(D)$ into $L^2(D)$, the operator $\mathcal{T}$ is compact. Hence $\mathcal{G}+\mathcal{T}$ is a Fredholm operator of index one, which proves the theorem. ∎

By modifying the variational approach of [9] in a similar way, one can also prove the following result.

THEOREM 3.5. *Assume that* $\bar{\xi}\cdot\mathcal{R}e(A^{-1})\,\xi \geq \gamma|\xi|^2$ *with* $\gamma > 1$. *Then the Fredholm alternative can be applied to the problem* (IMTP).

LEMMA 3.6. *Assume that* $\bar{\xi}\cdot\mathcal{I}m(A)\,\xi < 0$ *at a point* $x_0 \in D$ *and* $\eta \geq \eta_0 > 0$ *on* Γ_2. *Then* (IMTP) *has at most one solution.*

Proof. Let us consider the homogeneous problem (i.e., $f = h = r = 0$). Applying the divergence theorem to $\overline{v}$ and $A\nabla v$, making use of the boundary conditions, and applying Green's theorem for $\overline{w}$ and w, we obtain

$$\int_D \nabla\overline{v}\cdot A\nabla v\,dy - \int_D k^2|v|^2\,dy = \int_D |\nabla w|^2\,dy - \int_D k^2|w|^2\,dy + \int_{\Gamma_2} i\eta\left|\frac{\partial w}{\partial\nu}\right|^2 ds.$$

Hence

$$\mathcal{I}m\left(\int_D \nabla\overline{v}\cdot A\nabla v\,dy\right) = 0 \qquad \text{and} \qquad \int_{\Gamma_2}\eta\left|\frac{\partial w}{\partial\nu}\right|^2 ds = 0.$$

Since $\bar{\xi}\cdot\mathcal{I}m(A)\xi < 0$ in a small ball $B_{x_0} \subset D$, from the first equality we obtain that $\nabla v = 0$ in B_{x_0}, whence $v \equiv 0$ in D since the unique continuation principle holds for (3.5)(i). From the boundary conditions and the integral representation, formula w also vanishes in D. ∎

We summarize the above analysis in the following theorem.

THEOREM 3.7. *Assume that* $\bar{\xi}\cdot\mathcal{I}m(A)\,\xi < 0$ *at a point* $x_0 \in D$ *and* $\eta \geq \eta_0 > 0$. *In addition, assume that there exists a constant* $\gamma > 1$ *such that*

$$\textit{either} \quad \bar{\xi}\cdot\mathcal{R}e(A)\,\xi \geq \gamma|\xi|^2 \quad \textit{or} \quad \bar{\xi}\cdot\mathcal{R}e(A^{-1})\,\xi \geq \gamma|\xi|^2 \qquad \forall\,\xi\in\mathbb{C}^2.$$

Then the interior transmission problem (IMTP) *has a unique solution* (v, w) *which satisfies*

$$\|v\|^2_{H^1(D)} + \|w\|^2_{\mathbb{H}^1(D,\Gamma_2)} \leq C\left(\|f\|_{H^{\frac{1}{2}}(\Gamma)} + \|h\|_{H^{-\frac{1}{2}}(\Gamma)} + \|r\|_{L^2(\Gamma_2)}\right). \tag{3.10}$$

The values of k for which (IMTP) is not uniquely solvable are called the *transmission eigenvalues*. The latter may occur, for example, if $\bar{\xi}\cdot\mathcal{I}m(A)\,\xi = 0$ in D. In this case, from the proof of Lemma 3.6 we obtain that $\frac{\partial w}{\partial\nu} = 0$ on Γ_2, whence the eigenvalues of (IMTP) form a subset of the transmission eigenvalues corresponding to the (usual) interior transmission problem discussed in [7]. Moreover, if $\Gamma_2 = \Gamma$, then the eigenvalues of (IMTP) form a subset of the Neumann eigenvalues of $-\nabla\cdot A\nabla$.

4. The inverse problem. The inverse problem that we consider here is to determine *both* the shape of the scattering object D and the surface conductivity η from a knowledge of the far field pattern $u_\infty(\hat{x}, d)$ for all incident plane waves $u^i := e^{ikx\cdot d}$, $d \in \Omega$, and all observation directions $\hat{x} \in \Omega$. (Note that it suffices to know the far field pattern corresponding to all $d \in \Omega_1 \subset \Omega$ and all $\hat{x} \in \Omega_2 \subset \Omega$; of particular interest is the case $d = -\hat{x} \in \Omega_0 \subset \Omega$.) We start the investigation of the inverse problem by stating a uniqueness theorem for determining the support D.

THEOREM 4.1. *Let the domains* D^1 *and* D^2 *with the boundaries* Γ^1 *and* Γ^2, *respectively; the matrix valued functions* A_1 *and* A_2; *and the functions* η_1 *and* η_2 *determined on the portions* $\Gamma_2^1 \subseteq \Gamma^1$ *and* $\Gamma_2^2 \subseteq \Gamma^2$, *respectively* (*either* Γ_2^1 *or* Γ_2^2 *or both can possibly be empty sets*), *satisfy the assumptions of* (*MTP*) *in section* 2.

Moreover, let us assume that either $\bar{\xi}\cdot\Re(A_1)\,\xi \geq \gamma|\xi|^2$ *or* $\bar{\xi}\cdot\Re(A_1^{-1})\,\xi \geq \gamma|\xi|^2$, *and either* $\bar{\xi}\cdot\Re(A_2)\,\xi \geq \gamma|\xi|^2$ *or* $\bar{\xi}\cdot\Re(A_2^{-1})\,\xi \geq \gamma|\xi|^2$ *for some* $\gamma > 1$. *If the far field patterns* $u_\infty^1(\hat{x},d)$ *corresponding to the data* D^1, A_1, η_1 *and* $u_\infty^2(\hat{x},d)$ *corresponding to the data* D^2, A_2, η_2 *coincide for all* $\hat{x}, d \in \Omega$, *then* $D^1 \equiv D^2$.

This theorem is proved in [8] for the case of Maxwell's equations in $\mathbb{R}^3$. In the scalar case under consideration, one can adapt the approach of Hähner in [15] to prove the above theorem. Note that the main ingredient of Hähner's approach is the well-posedness of the (modified) interior transmission problem investigated in section 3.1.

The next question to ask is the uniqueness of the surface conductivity η. From the above theorem we can now assume that D is known. Furthermore, we require that for an arbitrarily choice of Γ_2, A, and η there exist at least one incident plane wave such that the corresponding total field u satisfies $\frac{\partial u}{\partial \nu}\big|_{\Gamma_0} \neq 0$, where $\Gamma_0 \subset \Gamma$ is an arbitrary portion of Γ. In the context of our application this is a reasonable assumption since otherwise the portion of the boundary where $\frac{\partial u}{\partial \nu}\big|_{\Gamma_0} = 0$ for all incident plane waves will behave like a perfect conductor, contrary to the assumption that the metallic coating is thin enough for the incident field to penetrate into D. We can prove the following result.

THEOREM 4.2. *Assume that* $\eta \in C(\overline{\Gamma}_2)$ *and that* k *is not a Neumann eigenvalue for* $-\nabla\cdot A\nabla$. *Then, under the above assumption and for fixed* D *and* A, *the surface conductivity* η *is uniquely determined from the far field pattern* $u_\infty(\hat{x},d)$ *for all* $\hat{x}, d \in \Omega$.

Proof. Let D and A be fixed, and suppose there exist $\eta_1 \in C(\overline{\Gamma}_2^1)$ and $\eta_2 \in C(\overline{\Gamma}_2^2)$ such that the corresponding scattered fields $u^{s,1}$ and $u^{s,2}$, respectively, have the same far field patterns $u_\infty^1(\hat{x},d) = u_\infty^2(\hat{x},d)$ for all $\hat{x}, d \in \Omega$. Then from Rellich's lemma, $u^{s,1} = u^{s,2}$ in $\mathbb{R}^2\backslash D$. Hence, from the transmission condition, the difference $V = v^1 - v^2$ satisfies

$$\nabla\cdot A\nabla V + k^2 V = 0 \quad \text{in } D, \tag{4.1}$$

$$\frac{\partial V}{\partial \nu_A} = 0 \quad \text{on } \Gamma, \tag{4.2}$$

$$V = -i(\tilde{\eta}_1 - \tilde{\eta}_2)\frac{\partial u^1}{\partial \nu} \quad \text{on } \Gamma, \tag{4.3}$$

where $\tilde{\eta}_1$ and $\tilde{\eta}_2$ are the extension by zero of η_1 and η_2, respectively, to the whole of Γ and $u^1 = u^{s,1} + u^i$. Assuming that k is not a Neumann eigenvalue for $-\nabla\cdot A\nabla$ (in particular, this is the case if $\mathcal{I}m(A) < 0$ at $x_0 \in D$, (4.1)), (4.2) implies that $V = 0$ in D, and hence (4.3) becomes

$$(\tilde{\eta}_1 - \tilde{\eta}_2)\frac{\partial u^1}{\partial \nu} = 0 \quad \text{on} \quad \Gamma$$

for all incident waves. Since for a given $\Gamma_0 \subset \Gamma$ there exists at least one incident plane wave such that $\frac{\partial u^1}{\partial \nu}\big|_{\Gamma_0} \neq 0$, the continuity of η_1 and η_2 in $\overline{\Gamma}_2^1$ and $\overline{\Gamma}_2^2$, respectively, implies that $\tilde{\eta}_1 = \tilde{\eta}_2$. ∎

We now define the *far field operator* $F : L^2(\Omega) \to L^2(\Omega)$ by

$$Fg(\hat{x}) := \int_\Omega u_\infty(\hat{x},d)g(d)\,ds(d) \tag{4.4}$$

and introduce the far field equation

$$(Fg)(\hat{x}) = \gamma e^{-ik\hat{x}\cdot z}, \quad g \in L^2(\Omega), \quad z \in D, \tag{4.5}$$

where $\gamma = \frac{e^{i\pi/4}}{\sqrt{8\pi k}}$ and $\gamma e^{-ik\hat{x}\cdot z}$ is the far field pattern of the fundamental solution $\Phi(x,z) := \frac{i}{4}H_0^{(1)}(k|x-z|)$ to the Helmholtz equation in $\mathbb{R}^2$, with $H_0^{(1)}$ being a Hankel function of the first kind of order zero. A reconstruction of D can be obtained by using the linear sampling method which characterizes the support D from a solution of the far field equation (4.5) (see, e.g., [3], [7]). Assuming that D is known, our goal is to provide a formula for computing the L_∞ norm of η in terms of the solution of the far field equation (4.5).

To this end, assuming that k is not a transmission eigenvalue, for $z \in D$ we denote by v_z and w_z the unique solution of the interior transmission problem

$$
\begin{aligned}
&\nabla\cdot A\nabla v_z + k^2 v_z = 0 && \text{in } D,\\
&\Delta w_z + k^2 w_z = 0 && \text{in } D,\\
&v_z - (w_z + \Phi(\cdot,z)) = 0 && \text{on } \Gamma_1,\\
&v_z - (w_z + \Phi(\cdot,z)) = -i\eta\,\frac{\partial}{\partial\nu}(w_z + \Phi(\cdot,z)) && \text{on } \Gamma_2,\\
&\frac{\partial v_z}{\partial \nu_A} - \frac{\partial}{\partial\nu}(w_z + \Phi(\cdot,z)) = 0 && \text{on } \Gamma.
\end{aligned}
\tag{4.6}
$$

We recall that a Herglotz wave function with kernel $g \in L^2(\Omega)$ is an entire solution of the Helmholtz equation defined by

$$
v_g(x) = \int_\Omega e^{ikx\cdot d} g(d)\,ds(d), \qquad x \in \mathbb{R}^2. \tag{4.7}
$$

The following theorem holds.

THEOREM 4.3. *Assume that k is not a transmission eigenvalue. Let $\epsilon > 0$, $z \in D$, and (w_z, v_z) be the unique solution of* (4.6). *Then there exists a Herglotz wave function $v_{g_\epsilon^z}$ with kernel $g_\epsilon^z \in L^2(\Omega)$ such that*

$$
\|w_z - v_{g_\epsilon^z}\|_{\mathbb{H}^1(D,\Gamma_2)} \le \epsilon. \tag{4.8}
$$

Moreover, there exists a positive constant $c > 0$ independent of ϵ and z such that

$$
\|(Fg_\epsilon^z)(\hat{x}) - \gamma e^{-ik\hat{x}\cdot z}\|_{L^2(\Omega)} \le c\epsilon. \tag{4.9}
$$

Proof. To prove the first part of the theorem we first show that the operator $\mathcal{H} : L^2(\Omega) \to H^{\frac{1}{2}}(\Gamma_1) \times L^2(\Gamma_2)$ defined by

$$
(\mathcal{H}g)(x) := \begin{cases} \displaystyle\int_\Omega e^{-iky\cdot\hat{x}} g(\hat{x})\,ds(\hat{x}), & y \in \Gamma_1,\\ \displaystyle\frac{\partial}{\partial\nu}\int_\Omega e^{-iky\cdot\hat{x}} g(\hat{x})\,ds(\hat{x}) + i\int_\Omega e^{-iky\cdot\hat{x}} g(\hat{x})\,ds(\hat{x}), & y \in \Gamma_2, \end{cases} \tag{4.10}
$$

has dense range. To this end it suffices to show that the corresponding dual operator $\mathcal{H}^* : \tilde{H}^{-\frac{1}{2}}(\Gamma_1) \times L^2(\Gamma_2) \to L^2(\Omega)$ defined by

$$
\langle \mathcal{H}g,\, \phi\rangle_{H^{\frac{1}{2}}(\Gamma_1),\tilde{H}^{-\frac{1}{2}}(\Gamma_1)} + \langle \mathcal{H}g,\, \psi\rangle_{L^2(\Gamma_2),L^2(\Gamma_2)} = \langle g,\, \mathcal{H}^*(\phi,\,\psi)\rangle_{L^2(\Omega),L^2(\Omega)}
$$

for all $g \in L^2(\Omega)$, $\phi \in \tilde{H}^{-\frac{1}{2}}(\Gamma_1), \psi \in L^2(\Gamma_2)$ is injective. By interchanging the order of integration, one can show that

$$
\mathcal{H}^*(\phi,\,\psi)(\hat{x}) = \int_\Gamma e^{-iky\cdot\hat{x}}\tilde{\phi}(y)\,ds(y) + \int_\Gamma \frac{\partial e^{-iky\cdot\hat{x}}}{\partial\nu}\tilde{\psi}(y)\,ds(y) + i\int_\Gamma e^{-iky\cdot\hat{x}}\tilde{\psi}(y)\,ds(y),
$$

where $\tilde{\phi} \in H^{-\frac{1}{2}}(\Gamma)$ and $\tilde{\psi} \in L^2(\Gamma)$ are the extension by zero to the whole boundary Γ of ϕ and ψ, respectively. Assume that $\mathcal{H}^*(\phi, \psi) = 0$. Since $\mathcal{H}^*(\phi, \psi)$ is, up to a constant, the far field pattern of the potential

$$P(x) = \int_\Gamma \Phi(x,y)\tilde{\phi}(y)\,ds(y) + \int_\Gamma \frac{\partial \Phi(x,y)}{\partial \nu}\tilde{\psi}(y)\,ds(y) + i\int_\Gamma \Phi(x,y)\tilde{\psi}(y)\,ds(y),$$

which satisfies the Helmholtz equation in D_e, from Rellich's lemma we have that $P(x) = 0$ in D_e. As $x \to \Gamma$, the following jump relations (in the L^2 limit sense [12], [16]) hold:

$$P^+ - P^-|_{\Gamma_1} = 0, \qquad P^+ - P^-|_{\Gamma_2} = \psi,$$

$$\frac{\partial P^+}{\partial \nu} - \frac{\partial P^-}{\partial \nu}\bigg|_{\Gamma_1} = -\phi, \qquad \frac{\partial P^+}{\partial \nu} - \frac{\partial P^-}{\partial \nu}\bigg|_{\Gamma_2} = -i\psi,$$

where by the superscript $+$ and $-$ we distinguish the limits obtained by approaching the boundary Γ from D^e and D, respectively. Using the fact that $P^+ = \frac{\partial P^+}{\partial \nu} = 0$, we see that P satisfies the Helmholtz equation and

$$P^-|_{\Gamma_1} = 0, \quad \frac{\partial P^-}{\partial \nu} + iP^-\bigg|_{\Gamma_2} = 0,$$

where the equalities are understood in the L^2 limit sense. Using Green's theorem and a parallel surface argument, one can conclude as in Theorem 2.1 in [3] that $P = 0$ in D, whence from the above jump relations $\phi = \psi = 0$.

Now let $w \in \mathbb{H}^1(D, \Gamma_2)$ be a solution of the Helmholtz equation in D. From the above we can approximate $w|_{\Gamma_1} \in H^{\frac{1}{2}}(\Gamma_1)$ and $\frac{\partial w}{\partial \nu} + iw|_{\Gamma_2} \in L^2(\Gamma_2)$ by $\mathcal{H}g$. Hence using the a priori estimate for the solution of the mixed boundary value problem for the Helmholtz equation (see Theorem 2.3 in [3]),

$$\|w\|_{H^1(D)} + \left\|\frac{\partial w}{\partial \nu}\right\|_{L^2(\Gamma_2)} \le C\left(\|w\|_{H^{\frac{1}{2}}(\Gamma_1)} + \left\|\frac{\partial w}{\partial \nu} + iw\right\|_{L^2(\Gamma_2)}\right),$$

we obtain that w can be approximated by a Herglotz wave function v_g with respect to the $\mathbb{H}^1(D, \Gamma_2)$-norm, which proves the first part of the theorem. Note that, by a change of variable, v_g defined by (4.7) can be written as $\int_\Omega e^{-ikx\cdot d} g(d)ds(d)$.

Next let $z \in D$. Then $\gamma e^{-ik\hat{x}\cdot z}$ is the far field pattern of the radiating solution $\Phi(x,z)$. Let w_z and v_z be the unique solution of (4.6). Obviously v_z and $\Phi(x,z)$ satisfy (MTP) with incident field the $H^1_{\mathrm{loc}}(\mathbb{R})$-extension of w_z. The well-posedness of (MTP) (section 3) together with the classical trace theorems and the approximation of w_z by a Herglotz wave function $v_{g_\epsilon^z}$ show that for every $\epsilon > 0$

$$\|(Fg_\epsilon^z)(\hat{x}) - \gamma e^{-ik\hat{x}\cdot z}\|_{L^2(\Omega)} \le c_1\|u^s_{g_\epsilon^z} - \Phi(\cdot, z)\|_{H^1(D_e \cap B_R)} \le c\|v_{g_\epsilon^z} - w_z\|_{H^1(D)} \le c\epsilon$$

for $c_1, c > 0$, where $u^s_{g_\epsilon^z}$ is the scattered field corresponding to $v_{g_\epsilon^z}$ as the incident wave. (Note that by superposition Fg_ϵ^z coincides with $u^s_{g_\epsilon^z}$.) This ends the proof. ∎

Now let us define W_z by

$$W_z := w_z + \Phi(\cdot, z). \tag{4.11}$$

In particular, since $w_z \in \mathbb{H}^1(D, \Gamma_2)$, $\Delta w_z \in L^2(D)$, and $z \in D$, we have that $W_z|_\Gamma \in H^{\frac{1}{2}}(\Gamma)$, $\frac{\partial W_z}{\partial \nu}|_\Gamma \in H^{-\frac{1}{2}}(\Gamma)$, and $\frac{\partial W_z}{\partial \nu}|_{\Gamma_2} \in L^2(\Gamma_2)$.

LEMMA 4.4. *For every two points z_1 and z_2 in D we have that*

$$
\begin{aligned}
(4.12)\qquad & -2\int_D \nabla v_{z_1}\cdot \mathcal{I}m(A)\nabla \overline{v}_{z_2}\,dx + 2\int_{\Gamma_2}\eta(x)\frac{\partial W_{z_1}}{\partial \nu}\frac{\partial \overline{W}_{z_2}}{\partial \nu}\,ds \\
& = -4k\pi|\gamma|^2 J_0(k|z_1-z_2|) + i\left(w_{z_1}(z_2)-\overline{w}_{z_2}(z_1)\right),
\end{aligned}
$$

where w_{z_1}, W_{z_1} and w_{z_2}, W_{z_2} are defined by (4.6) *and* (4.11)*, respectively, and J_0 is a Bessel function of order zero.*

Proof. Let z_1 and z_2 be two points in D and v_{z_1}, w_{z_1}, W_{z_1} and v_{z_2}, w_{z_2}, W_{z_2} the corresponding functions defined by (4.6) and (4.11). Applying the divergence theorem to $v_{z_1}, \overline{v}_{z_2}$ and using (4.6) together with the fact that A is symmetric, we have

$$
\begin{aligned}
\int_\Gamma \left(v_{z_1}\frac{\partial \overline{v}_{z_2}}{\partial \nu_{\overline{A}}} - \overline{v}_{z_2}\frac{\partial v_{z_1}}{\partial \nu_A}\right)ds &= \int_D \left(\nabla v_{z_1}\cdot \overline{A}\nabla \overline{v}_{z_2} - \nabla \overline{v}_{z_2}\cdot A\nabla v_{z_1}\right)dx \\
+\int_D \left(v_{z_1}\nabla\cdot \overline{A}\nabla \overline{v}_{z_2} - \overline{v}_{z_2}\nabla\cdot A\nabla v_{z_1}\right)dx &= -2i\int_D \nabla v_{z_1}\cdot \mathcal{I}m(A)\nabla \overline{v}_{z_2}\,dx.
\end{aligned}
$$

On the other hand, from the boundary conditions we have

$$
\begin{aligned}
&\int_\Gamma \left(v_{z_1}\frac{\partial \overline{v}_{z_2}}{\partial \nu_{\overline{A}}} - \overline{v}_{z_2}\frac{\partial v_{z_1}}{\partial \nu_A}\right)ds \\
&\quad = \int_\Gamma \left(W_{z_1}\frac{\partial \overline{W}_{z_2}}{\partial \nu} - \overline{W}_{z_2}\frac{\partial W_{z_1}}{\partial \nu}\right)ds - 2i\int_{\Gamma_2}\eta(x)\frac{\partial W_{z_1}}{\partial \nu}\frac{\partial \overline{W}_{z_2}}{\partial \nu}\,ds.
\end{aligned}
$$

Hence

$$
\begin{aligned}
&-2i\int_D \nabla v_{z_1}\cdot \mathcal{I}m(A)\nabla \overline{v}_{z_2}\,dx + 2i\int_{\Gamma_2}\eta(x)\frac{\partial W_{z_1}}{\partial \nu}\frac{\partial \overline{W}_{z_2}}{\partial \nu}\,ds \\
&\quad = \int_\Gamma \left(W_{z_1}\frac{\partial \overline{W}_{z_2}}{\partial \nu} - \overline{W}_{z_2}\frac{\partial W_{z_1}}{\partial \nu}\right)ds = \int_\Gamma \left(\Phi(\cdot,z_1)\frac{\partial \overline{\Phi(\cdot,z_2)}}{\partial \nu} - \overline{\Phi(\cdot,z_2)}\frac{\partial \Phi(\cdot,z_1)}{\partial \nu}\right)ds \\
&\qquad + \int_\Gamma \left(w_{z_1}\frac{\partial \overline{\Phi(\cdot,z_2)}}{\partial \nu} - \overline{\Phi(\cdot,z_2)}\frac{\partial w_{z_1}}{\partial \nu}\right)ds + \int_\Gamma \left(\Phi(\cdot,z_1)\frac{\partial \overline{w}_{z_2}}{\partial \nu} - \overline{w}_{z_2}\frac{\partial \Phi(\cdot,z_1)}{\partial \nu}\right)ds.
\end{aligned}
$$

Green's theorem applied to the radiating solution $\Phi(\cdot,z)$ of the Helmholtz equation in D_e implies that [13]

$$
\begin{aligned}
&\int_\Gamma \left(\Phi(\cdot,z_1)\frac{\partial \overline{\Phi(\cdot,z_2)}}{\partial \nu} - \overline{\Phi(\cdot,z_2)}\frac{\partial \Phi(\cdot,z_1)}{\partial \nu}\right)ds = -2ik\int_\Omega \Phi_\infty(\cdot,z_1)\overline{\Phi_\infty(\cdot,z_2)}ds \\
&\quad = -2ik\int_\Omega |\gamma|^2 e^{-ik\hat{x}\cdot z_1}e^{ik\hat{x}\cdot z_2}\,ds = -4ik\pi|\gamma|^2 J_0(k|z_1-z_2|),
\end{aligned}
$$

and from the representation formula for w_{z_1} and w_{z_2} we now obtain

$$
\begin{aligned}
&-2i\int_D \nabla v_{z_1}\cdot \mathcal{I}m(A)\nabla \overline{v}_{z_2}\,dx + 2i\int_{\Gamma_2}\eta(x)\frac{\partial W_{z_1}}{\partial \nu}\frac{\partial \overline{W}_{z_2}}{\partial \nu}\,ds \\
&\quad = -4ik\pi|\gamma|^2 J_0(k|z_1-z_2|) + \overline{w}_{z_2}(z_1) - w_{z_1}(z_2).
\end{aligned}
$$

Dividing both sides of the above relation by i yields the result. ∎

In the following we consider a ball $B_r \subset D$ of radius r contained in D and define a subset of $L^2(\Gamma_2)$ by

$$\mathcal{V} := \left\{ f \in L^2(\Gamma_2) : \quad \begin{array}{l} f = \dfrac{\partial W_z}{\partial \nu}\Big|_{\Gamma_2} \text{ with } W_z = w_z + \Phi(\cdot, z), \\ z \in B_r \text{ and } w_z, v_z \text{ the solution of (4.6).} \end{array} \right\}$$

LEMMA 4.5. *Assuming that k is neither a transmission eigenvalue nor a Neumann eigenvalue for $-\nabla \cdot A\nabla$, then $\mathcal{V}$ is complete in $L^2(\Gamma_2)$.*

Proof. Let φ be a function in $L^2(\Gamma_2)$ such that for every $z \in B_r$

$$\int_{\Gamma_2} \frac{\partial W_z}{\partial \nu}\, \varphi \, ds = 0.$$

Construct $v \in H^1(D)$ and $w \in \mathbb{H}^1(D, \Gamma_2)$ as the unique solution of the interior transmission problem

$$\begin{array}{lll} \text{(i)} & \nabla \cdot A\nabla v + k^2 v = 0 & \text{in } D, \\ \text{(ii)} & \Delta w + k^2 w = 0 & \text{in } D, \\ \text{(iii)} & v - w = 0 & \text{on } \Gamma_1, \\ \text{(iv)} & v - w = -i\eta \dfrac{\partial w}{\partial \nu} + \varphi & \text{on } \Gamma_2, \\ \text{(v)} & \dfrac{\partial v}{\partial \nu_A} - \dfrac{\partial w}{\partial \nu} = 0 & \text{on } \Gamma. \end{array}$$

Then we have

$$\begin{aligned} 0 &= \int_{\Gamma_2} \frac{\partial W_z}{\partial \nu}\, \varphi \, ds = \int_{\Gamma} \frac{\partial W_z}{\partial \nu}\, (v - w)\, ds + i \int_{\Gamma_2} \eta \frac{\partial W_z}{\partial \nu} \frac{\partial w}{\partial \nu}\, ds \\ &= \int_{\Gamma} \frac{\partial W_z}{\partial \nu}\, v \, ds - \int_{\Gamma} \frac{\partial W_z}{\partial \nu}\, w \, ds + i \int_{\Gamma_2} \eta \frac{\partial W_z}{\partial \nu} \frac{\partial w}{\partial \nu}\, ds. \end{aligned} \tag{4.13}$$

Next from the equations for v_z and v, the divergence theorem, and the transmission conditions, we have

$$\begin{aligned} \int_{\Gamma} \frac{\partial W_z}{\partial \nu}\, v \, ds &= \int_{\Gamma} \frac{\partial v_z}{\partial \nu_A}\, v \, ds = \int_{\Gamma} \frac{\partial v}{\partial \nu_A}\, v_z \, ds \\ &= \int_{\Gamma} \frac{\partial w}{\partial \nu}\, W_z \, ds - i \int_{\Gamma_2} \eta \frac{\partial W_z}{\partial \nu} \frac{\partial w}{\partial \nu}\, ds. \end{aligned} \tag{4.14}$$

Finally, substituting (4.14) into (4.13) and using the integral representation formula yields

$$\begin{aligned} 0 &= \int_{\Gamma} \left(\frac{\partial w}{\partial \nu}\, W_z - \frac{\partial W_z}{\partial \nu}\, w \right) ds = \int_{\Gamma} \left(\frac{\partial w}{\partial \nu}\, w_z - \frac{\partial w_z}{\partial \nu}\, w \right) ds \\ &= \int_{\Gamma} \left(\frac{\partial w}{\partial \nu}\, \Phi(\cdot, z) - \frac{\partial \Phi(\cdot, z)}{\partial \nu}\, w \right) ds = w(z) \qquad \forall z \in B_r. \end{aligned} \tag{4.15}$$

The unique continuation principle for the Helmholtz equation now implies that $w = 0$ in D. Hence if k is not a Neumann eigenvalue corresponding to $-\nabla \cdot A\nabla$ (e.g., if

$\mathcal{I}m(A) < 0$ at a point $x_0 \in D$), then $v \equiv 0$ and therefore $\varphi = 0$, which proves the lemma. ∎

Now we are ready to prove the main result of this section.

THEOREM 4.6. *Let $\eta \in L_\infty(\Gamma_2)$ be the surface conductivity of* (MTP)*, and assume that $\mathcal{I}m(A) = 0$ in D and k is neither a transmission eigenvalue nor a Neumann eigenvalue for $-\nabla \cdot A\nabla$. Then*

$$\|\eta\|_{L_\infty(\Gamma_2)} = \sup_{\substack{z_i,\, z_j \in B_r \\ \alpha_i \in \mathbb{C}}} \frac{\sum_{i,j} \alpha_i \overline{\alpha_j} \left(-4\pi k|\gamma|^2 J_0(k|z_i - z_j|) + i w_{z_i}(z_j) - i\overline{w}_{z_j}(z_i)\right)}{2\| \sum_i \alpha_i \frac{\partial}{\partial \nu}(w_{z_i} + \Phi(\cdot; z_i))\|^2_{L^2(\Gamma_2)}}, \tag{4.16}$$

where w_z is such that (w_z, v_z) solves (4.6) *and the sums are arbitrary finite sums.*

Proof. We recall that

$$\|\eta\|_{L_\infty(\Gamma_2)} := \operatorname{ess\,sup} \eta = \sup_{f \in L^2(\Gamma_2)} \frac{1}{\|f\|^2_{L^2(\Gamma_2)}} \int_{\Gamma_2} \eta(x)|f|^2 ds.$$

The theorem then follows from Lemmas 4.4 and 4.5 by fixing first z_2 and then z_1 and considering linear combinations of $\frac{\partial W_z}{\partial \nu}$ for different $z \in B_r$. ∎

Given that D is known, w_z in the right-hand side of (4.16) still cannot be computed, since it depends on the unknown functions η and A. However, from Theorem 4.3, we can use in (4.16) an approximation to w_z given by the Herglotz wave function v_{g^z} with kernel g^z being the (regularized) solutions of the far field equation (4.5).

In the particular case where the coating is homogeneous, i.e., the surface conductivity is a positive constant $\eta > 0$, we can further simplify (4.16). In particular, fix an arbitrary point $z_0 \in B_r$ and consider $z_1 = z_2 = z_0$. Then (4.12) simply becomes

$$\eta = \frac{-2k\pi|\gamma|^2 - \mathcal{I}m\left(w_{z_0}(z_0)\right)}{\|\frac{\partial}{\partial \nu}(w_{z_0} + \Phi(\cdot; z_0))\|^2_{L^2(\Gamma_2)}}. \tag{4.17}$$

A drawback of (4.16) and (4.17) is that the extent of the coating Γ_2 is not known. Thus, in practice these expressions provide only a lower bound for $\|\eta\|_{L_\infty(\Gamma_2)}$. However, if the object is fully coated, that is, $\Gamma_2 = \Gamma$, we can compute an approximation of $\|\eta\|_{L_\infty(\Gamma_2)}$ by (4.12) and (4.17), where Γ_2 is replaced by of Γ.

5. Remarks on Maxwell's equations in $\mathbb{R}^3$. The analysis of the previous three sections for the case of scattering by an infinite cylinder can in principle be extended to the scattering of electromagnetic waves by a bounded dielectric in $\mathbb{R}^3$. In this case the direct scattering problem is given by (2.1)–(2.6), and the existence of a unique solution to this problem was established in [8]. The results for the inverse scattering problem for an infinite cylinder established in section 4 of this paper can in turn be extended to the case of Maxwell's equations in $\mathbb{R}^3$, *provided* that one can establish the existence of a unique solution to the interior transmission problem

$$\begin{aligned} &\nabla \times E_z - ikH_z = 0 \\ &\nabla \times H_z + ikE_z = 0 \end{aligned} \qquad \text{in } D, \tag{5.1}$$

$$\begin{aligned} &\nabla \times E_z^{\text{int}} - ikH_z^{\text{int}} = 0 \\ &\nabla \times H_z^{\text{int}} + ikN(x)E_z^{\text{int}} = 0 \end{aligned} \qquad \text{in } D, \tag{5.2}$$

together with the boundary relations

$$
\begin{aligned}
&\nu \times E_z^{\rm int} - \nu \times E_z = \nu \times E_e(\cdot, z, q) && \text{on } \Gamma,\\
&\nu \times H_z^{\rm int} - \nu \times H_z = \nu \times H_e(\cdot, z, q) && \text{on } \Gamma_1,\\
&\nu \times H_z^{\rm int} - \nu \times H_z = -\eta\,[\nu \times (E_z + E_e(\cdot, z, q)) \times \nu] + \nu \times H_e(\cdot, z, q) && \text{on } \Gamma_2,
\end{aligned}
\tag{5.3}
$$

where $E_e(\cdot, z, q)$, $H_e(\cdot, z, q)$ is the electric dipole defined by

$$
E_e(x, z, q) := \frac{i}{k} \nabla_x \times \nabla_x \times q\, \Phi(x, z), \qquad H_e(x, z, q) := \nabla_x \times q\, \Phi(x, z),
\tag{5.4}
$$

where $q \in \mathbb{R}^3$ is a constant vector and

$$
\Phi(x, z) := \frac{1}{4\pi} \frac{e^{ik|x-z|}}{|x - z|}.
$$

Unfortunately this result remains an open problem. (For the existence of a unique solution to a modified version of (5.1)–(5.3), see [8].)

Assuming the existence of a unique solution of (5.1)–(5.3), one can now proceed to derive the three-dimensional analogue of Theorem 4.6; i.e., if $\mathcal{I}m(N) = 0$ and k is not a transmission eigenvalue, then

$$
\|\eta\|_{L_\infty(\Gamma_2)} = \sup_{\substack{z_i \in B_r,\, q \in \mathbb{R}^3 \\ \alpha_i \in \mathbb{C}}} \frac{\sum_{i,j} \alpha_i \overline{\alpha_j} \left[-\|q\|^2 A(z_i, z_j, k, q) + q \cdot E_{z_i}(z_j) + q \cdot \overline{E}_{z_j}(z_i) \right]}{2 \| \sum_i \alpha_i\, \nu \times (E_{z_i} + E_e(\cdot, z_i, q)) \|^2_{L_t^2(\Gamma_2)}},
\tag{5.5}
$$

where $B_r \subset D$ is a ball of radius r;

$$
A(z_i, z_j, k, q) = \frac{k^2}{6\pi} \left[2 j_0(k|z_i - z_j|) + j_2(k|z_i - z_j|)(3\cos^2\phi - 1) \right],
\tag{5.6}
$$

j_0 and j_2 being spherical Bessel functions of order 0 and 2, respectively; ϕ is the angle between $(z_i - z_j)$ and q; and E_z, $E_z^{\rm int}$ is the unique solution of the interior transmission problem (5.1)–(5.3). In particular, E_z can be approximated by

$$
E_{g_z}(x) := ik \int_\Omega e^{ikx \cdot d} g_z(d) ds(d),
\tag{5.7}
$$

where $\Omega := \{x \in \mathbb{R}^3 : |x| = 1\}$ and g_z is the (regularized) solution of the *far field equation*

$$
\int_\Omega E_\infty(\hat{x}, d, g(d)) ds(d) = E_{e,\infty}(\hat{x}, z, q).
\tag{5.8}
$$

Here E_∞ is the electric far field pattern corresponding to the scattering problem (2.1)–(2.6), and $E_{e,\infty}$ is the electric far field pattern of the electric dipole (5.4). For details in the case of a perfect conductor coated by a dielectric, see section 3 of [5].

In the special case when η is a constant, (5.5) simplifies to

$$
\eta = \frac{-\frac{k^2}{6\pi}\|q\|^2 + \mathcal{R}e\,(q \cdot E_{z_0}(z_0))}{\|\nu \times (E_{z_0} + E_e(\cdot, z_0, q))\|^2_{L_t^2(\Gamma_2)}},
\tag{5.9}
$$

where z_0 is an arbitrary point in D.

6. Numerical examples. In this section we shall present some numerical tests of the inversion scheme using synthetic far field data for the Helmholtz equation. For a given scatterer, the far field data is computed by using a cubic finite element code to approximate the near field, and then employing a near to far field transformation [18]. The finite element computational domain is terminated by a rectilinear perfectly matched layer using a linear absorption function in the layer [2], [10].

Having computed approximate values of the far field pattern at N uniformly spaced points on the unit circle for N incoming waves, we have an $N \times N$ matrix $\mathcal{A}$ of approximate far field data

$$\mathcal{A}_{m,n} = u_{h,\infty}(d_m, d_n) \quad \text{where} \quad d_m = \left(\cos\left(\frac{2\pi(m-1)}{N}\right),\ \sin\left(\frac{2\pi(m-1)}{N}\right)\right)^T$$

for $1 \le m, n \le N$, where $u_{h,\infty}$ is the finite element far field pattern. To this we add further noise with parameter ϵ to obtain $\mathcal{A}_\epsilon$ using

$$(\mathcal{A}_\epsilon)_{m,n} = \mathcal{A}_{m,n}(1 + \epsilon(\xi_{1,m,n} + i\xi_{2,m,n})),$$

where $\xi_{1,m,n}$ and $\xi_{2,m,n}$ are given by a random number generator, uniformly distributed in the range $[-1, 1]$. Unless otherwise stated, $\epsilon = 0.01$ in these studies.

For a given sampling point z, the discrete far field equation is then to compute $\vec{g} = (g_1, \ldots, g_N)$ such that $\mathcal{A}_\epsilon \vec{g} = \vec{b}$, where

$$b_m = N\gamma \exp \frac{(-ik(z \cdot d_m))}{(2\pi)}, \qquad 1 \le m \le N.$$

This ill-conditioned problem is solved approximately using the Tikhonov regularization and the Morozov discrepancy principle as described, for example, in [14].

6.1. Exact knowledge of the boundary. We start as in [5], assuming an exact knowledge of the boundary in order to assess the accuracy of (4.17) without the added error of computing an approximation to the boundary of the scatterer. In this case, for a given scatterer, we compute $\vec{g}$ for $z = z_0$ using the Morozov method outlined in the previous section, and then approximate (4.17) using the trapezoidal rule with 100 integration points. After limited experiments, we choose $z_0 = (0, 0)^T$ (both upcoming examples have this point as their centroid).

To simplify the presentation, we have limited our discussion to two scatterers: an ellipse given by $x = 0.5\cos(s)$ and $y = 0.2\sin(s)$, $s \in [0, 2\pi]$, and the rectangle $[-0.5, 0.5] \times [0.4, 0.4]$. In (2.7) we choose $A = (1/4)I$. In all cases $k = 5$.

For the ellipse we consider either a fully coated or partially coated object. The partially coated boundary is shown in Figure 6.1. In Figure 6.2 we show results of the reconstruction of a range of conductivities η for the fully coated ellipse, partially coated ellipse, and fully coated rectangle. For each exact η we compute the far field data, add noise, and compute an approximation to w_{z_0}, as discussed before. Despite the noise on the data, η is well approximated in the case of the fully coated scatterers, provided that the conductivity is not too large. In all cases the approximation of η deteriorates for large conductivities, and as expected, (4.17) leads to an underestimate of η when the boundary is partially coated. These limited examples suggest that (4.17) provides a viable method for reconstructing η, provided that η is small enough and the boundary of the scatterer is known sufficiently accurately.

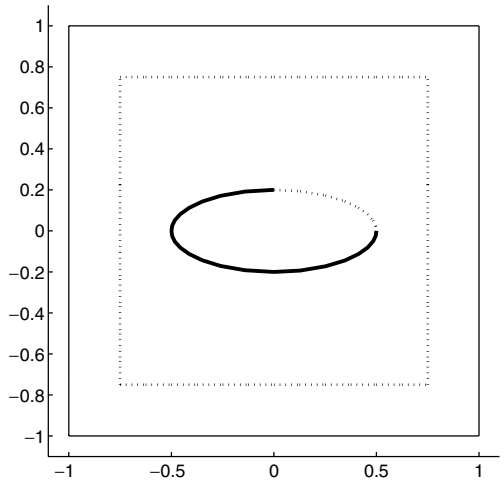
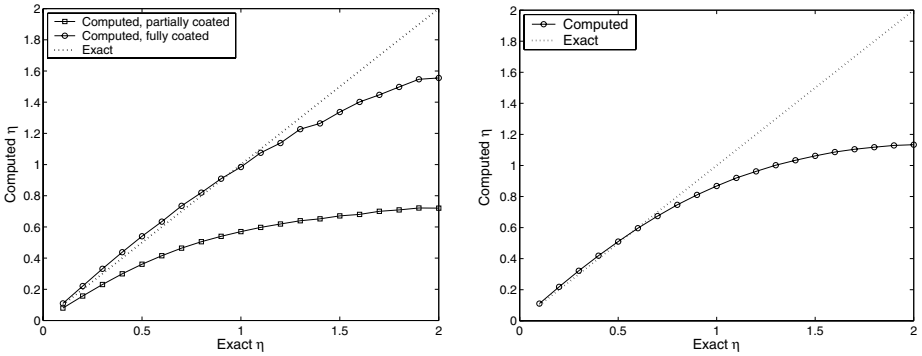


FIG. 6.1. *A diagram showing the coated portion of the partially coated ellipse as a thick line. The dotted square is the inner boundary of the PML, and the solid square is the boundary of the finite element computational domain.*



(a) (b)

FIG. 6.2. *Computation of η using the exact boundary. Panel* (a) *shows results for the fully and the partially coated ellipse. Panel* (b) *shows the corresponding results for the fully coated rectangle. Clearly in all cases the approximation of η deteriorates for large conductivities.*

6.2. The ellipse. We now wish to investigate the solution of the full inverse problem. We start by using the standard linear sampling method to approximate the boundary of the scatterer. In particular we compute $1/\|\vec{g}\|$ for z on a uniform grid in the sampling domain. In the upcoming numerical results we have arbitrarily chosen $N = 61$, and we sample on a 101×101 grid on the square $[-1, 1] \times [-1, 1]$. This procedure takes around 10 seconds in MATLAB on an Apple G5 computer, so it is not time-consuming.

Having computed $\vec{g}$ for each sample point, we have a discrete level set function $1/\|\vec{g}\|$. Choosing a contour value C then provides a reconstruction of the support of the given scatterer. We extract the edge of the reconstruction and then fit this using a trigonometric polynomial of degree M, assuming that the reconstruction is star-like with respect to the origin. (For more advanced applications it would be necessary to

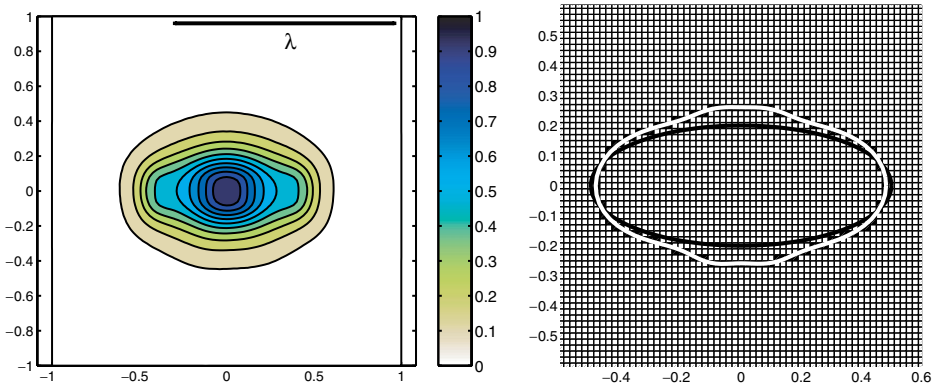


FIG. 6.3. *Illustration of the steps in the computation of η. First the standard linear sampling method is used to compute $1/\|\vec{g}\|$, as shown in panel* (a) *(the bar labeled λ shows the wavelength of the radiation). Choosing a cutoff C (in this case $C = 0.3$), the surface in panel* (a) *provides an approximation to the boundary of the scatterer shown as shaded blocks in panel* (b)*. Each square in this figure contains one sampling point z at its center. We also show in panel* (b) *the outline of the true scatterer as a smooth solid line, and as a white line the fit of the trigonometric series to the reconstruction. In this case C is chosen too small and the computed boundary lies outside the true scatterer.*

employ a more elaborate smoothing procedure.) Thus for an angle θ the radius of the reconstruction is given by

$$r(\theta) = \Re\left(\sum_{n=-M}^{M} r_n \exp(in\theta)\right),$$

where r is measured from the origin (since in all the examples here the origin is within the scatterer). The coefficients r_n are found using a least squares fit to the boundary identified in the previous step of the algorithm. Once we have a parameterization of the reconstructed boundary, we can compute the normal to the boundary and evaluate (4.17) for some choice of z_0 (in the examples always $z_0 = (0,0)^T$) using the trapezoidal rule with 100 points. This provides our reconstruction of η.

Figure 6.3 shows the main steps for evaluating our prediction of η for the ellipse. Here we choose $\eta = 1$ on the entire ellipse (fully coated). In (a) we see a plot of $1/\|\vec{g}\|$ (normalized so that the maximum value is 1) as a function of position. In this case the choice $\epsilon = 0.01$ for the additional error in the far field pattern gives an error of 1.3% in the spectral norm for $\mathcal{A}$.

We then make the arbitrary choice $C = 0.3$ (i.e., due to the normalization, the value is 0.3 times the maximum of $1/\|\vec{g}\|$). Figure 6.3(b) shows a plot of the pixels separating regions where $1/\|\vec{g}\| > C$ and $1/\|\vec{g}\| < C$. For clarity, we have graphed only the region $[-0.6, 0.6] \times [-0.6, 0.6]$. The black pixels in Figure 6.3(b) are then fitted using $M = 8$ in the trigonometric polynomial for $r(\theta)$, and the resulting curve is shown as a light curve on the figure. We also indicate, using a thick black line, the true ellipse. We have deliberately chosen a contour value C that does not give the best reconstruction of the ellipse so that the different geometric features can be easily seen. Using this reconstruction results in a predicted value of $\eta = 0.8372$ (compared to the true value $\eta = 1$).

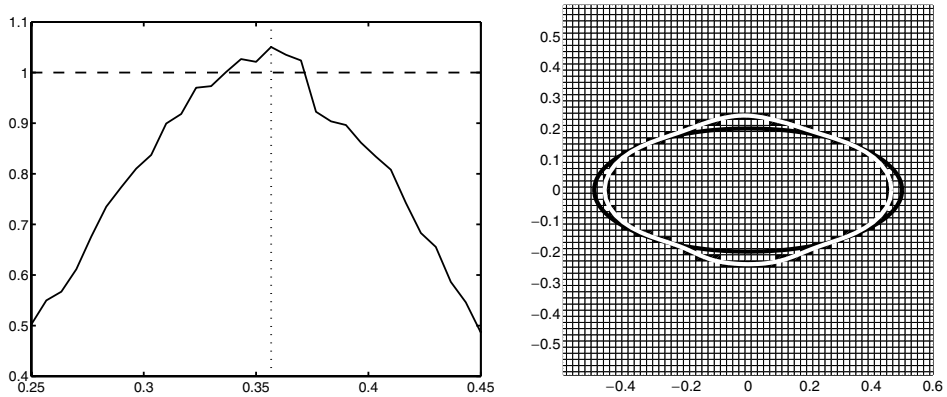


FIG. 6.4. *Panel* (a) *shows the computed value of* η *as a function of the cutoff* C*. The dashed line is the true value* $\eta = 1$*, and the dotted line marks the maximum predicted* η*. The corresponding reconstruction of the ellipse is shown in panel* (b) *using the same convention as in panel* (b) *of Figure* 6.3.

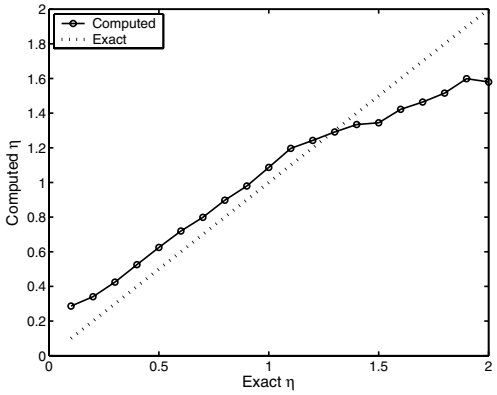


FIG. 6.5. *Reconstruction of a range of* η*. For each exact* η *we apply the reconstruction algorithm using a range of cutoffs and plot the corresponding reconstruction. An exact reconstruction would lie on the dotted line.*

With both scatterers in this study we have observed that a poor choice of the cutoff C tends to result in a predicted value of η that is too small. Therefore we now suggest sweeping through a range of values of C, and we find that the maximum value of η correlates with a good reconstruction of the scatterer and a better approximation of the true value of η. We show this in Figure 6.4 for the fully coated ellipse. The largest predicted value of η is $\eta = 1.05$ when $C = .3567$, and the reconstruction of the scatterer is better than choosing $C = 0.3$.

The reconstruction algorithm is next investigated for a range of values of η. For each exact η we apply the reconstruction algorithm using multiple cutoffs and plot the corresponding reconstruction of η. The results are shown in Figure 6.5 and should be compared to those in Figure 6.2(a). Given that the shape of the object and the parameter η are both being reconstructed, the results show reasonable agreement of the reconstruction up to approximately $\eta = 1.5$. For larger values of η the reconstruction deteriorates, perhaps because the field inside the scatterer diminishes as η

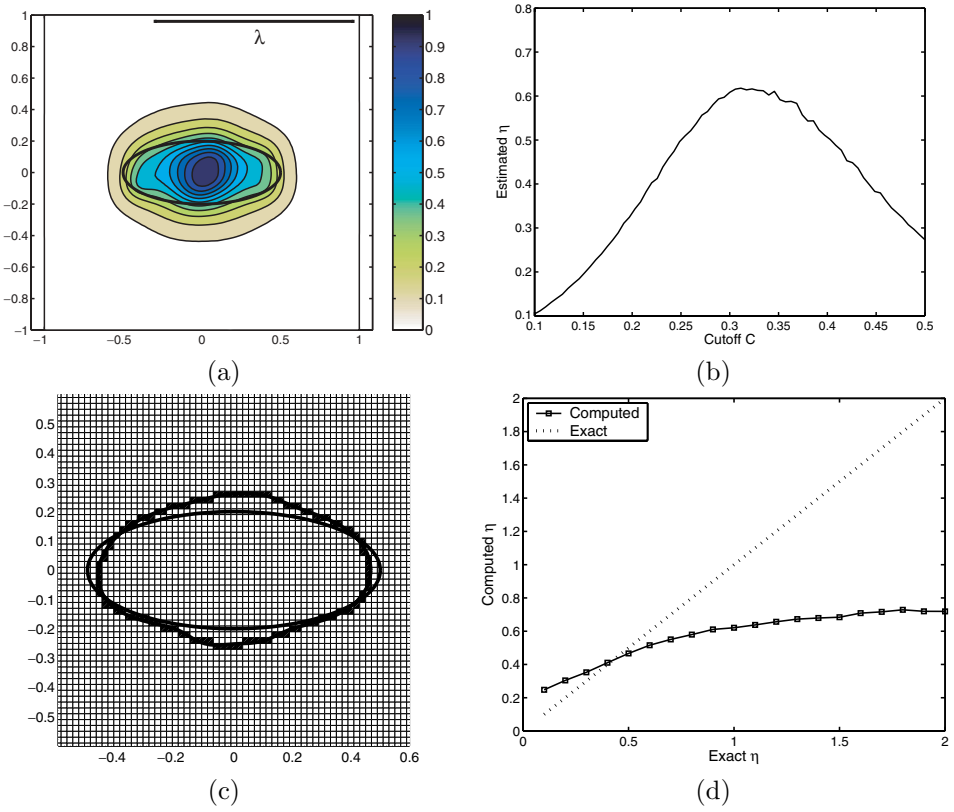


FIG. 6.6. *Reconstruction of the partially coated ellipse for* $\eta = 1$. (a) *The indicator function* $1/\|\vec{g}\|$ *resulting from the standard linear sampling method.* (b) *The computed value of* η *for a range of cutoffs* C. *The best reconstruction (maximum value of* η*) is* $\eta = 0.61793$ *when* $C = 0.3114$. (c) *The reconstruction of the ellipse using* $C = 0.3114$. (d) *The reconstruction of a range of* η*; this should be compared to Figure* 6.2(a).

increases. In this case the linear sampling method is able to provide a sufficiently accurate approximation of the ellipse so that the reconstruction of η in Figures 6.2 and 6.5 is of comparable accuracy.

Next we consider the partially coated ellipse (see Figure 6.1). The inversion algorithm is unchanged (both the boundary of the scatterer and η are reconstructed). The results are shown in Figure 6.6(a)–(c) when $\eta = 1$, and the results for a range of η are shown in Figure 6.6(d). The linear sampling method can still reconstruct the ellipse with reasonable fidelity despite the partial coating, and so the results in Figure 6.6(d) and Figure 6.2(a) are comparable. Recall that, for a partially coated obstacle, (4.17) provides only a lower bound for η.

6.3. Rectangular scatterer. Finally we show the reconstruction of the surface conductivity of the fully coated rectangular scatterer. Results for a range of η are shown in Figure 6.7. Comparing this to the reconstruction computed using the exact boundary (shown in Figure 6.2(b)), the results are much worse.

The deterioration of the results for the full inversion scheme can be explained by considering one choice of η in detail. In Figure 6.8 we show the full reconstruction procedure for $\eta = 1$. As in the case of the ellipse, we use the linear sampling method

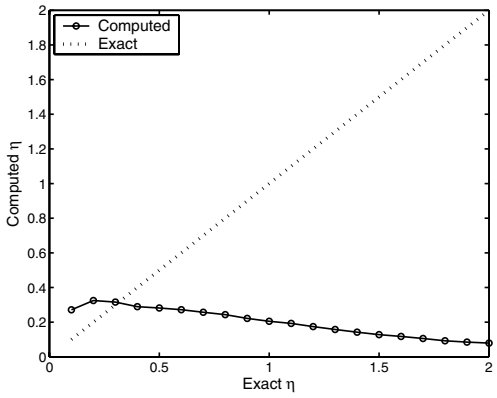


FIG. 6.7. *Results of reconstructing a range of conductivities for the fully coated rectangle. We plot the computed conductivity against its exact value. The results should be compared to Figure* 6.2(b)*, and are seen to be substantially worse than that case.*

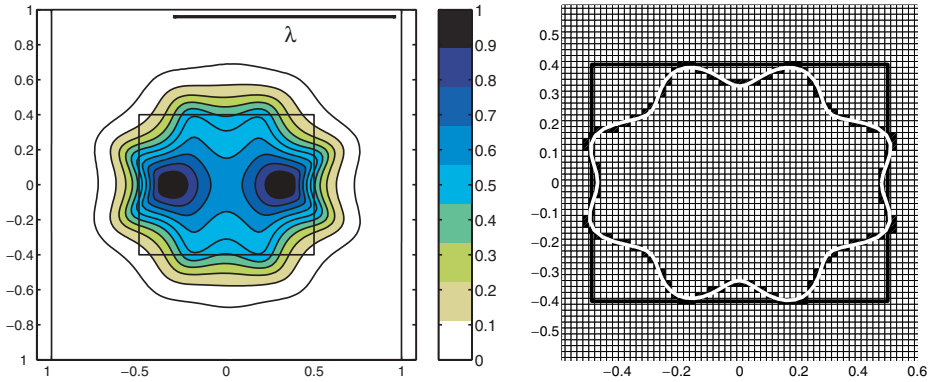


FIG. 6.8. *Details of the reconstruction of η when the exact value is $\eta = 0.5$.* (a) *The indicator function computed by the linear sampling method. Clearly, whatever the choice of C, the reconstruction of the scatterer will not provide an accurate normal.* (b) *The best reconstruction corresponding to $C = 0.52$, which yields the computed value of $\eta = 0.28$.*

to provide an indicator function for the boundary of the rectangle, but, compared to the ellipse, the reconstruction of the boundary shown in Figure 6.8(b) (at the best cutoff C) is now quite poor. From this reconstruction we need to compute the normal derivative of w_{z_0}. It is clear that this will be poorly approximated, and thus (4.17) will provide a poor approximation to η.

7. Conclusion. We have provided a method for estimating the surface conductivity of a scatterer from far field measurements. Numerical experiments show that this method can be combined with the linear sampling method to simultaneously identify the shape of the scatterer and the conductivity, provided that the shape of the scatterer can be computed with sufficient accuracy. Limitations include the fact that the method becomes inaccurate for large values of the surface conductivity, and the quality of the reconstruction of the conductivity can also be adversely influenced by the quality of the reconstruction of the scatterer. This may in part be due to the need to use the normal derivative of the Herglotz wave function in (4.17). We now

plan to investigate the use of the electric far field pattern, which should allow us to avoid the normal derivative. However considerable mathematical difficulties need to be overcome, and in particular the existence of a solution of the interior transmission problem is not known in this case.

REFERENCES

[1] T. S. Angell and A. Kirsch, *The conductive boundary condition for Maxwell's equations*, SIAM J. Appl. Math., 52 (1992), pp. 1597–1610.

[2] J. P. Bérenger, *A perfectly matched layer for the absorption of electromagnetics waves*, J. Comput. Phys., 114 (1994), pp. 185–200.

[3] F. Cakoni, D. Colton, and P. Monk, *The direct and inverse scattering problems for partially coated obstacles*, Inverse Problems, 17 (2001), pp. 1997–2015.

[4] F. Cakoni and D. Colton, *Combined far field operators in electromagnetic inverse scattering theory*, Math. Methods Appl. Sci., 26 (2003), pp. 413–429.

[5] F. Cakoni and D. Colton, *The determination of the surface impedance of a partially coated obstacle from far field data*, SIAM J. Appl. Math., 64 (2004), pp. 709–723.

[6] F. Cakoni and D. Colton, *A target signature for distinguishing perfect conductors from anisotropic media of finite conductivity*, Math. Comput. Simulation, 66 (2004), pp. 315–324.

[7] F. Cakoni, D. Colton, and H. Haddar, *The linear sampling method for anisotropic media*, J. Comput. Appl. Math., 146 (2002), pp. 285–299.

[8] F. Cakoni and D. Colton, *A uniqueness theorem for an inverse electromagnetic scattering problem in inhomogeneous anisotropic media*, Proc. Edinburgh Math. Soc., 46 (2003), pp. 293–314.

[9] F. Cakoni and H. Haddar, *The Linear Sampling Method for Anisotropic Media: Part* 2, Preprint 2001/26, Mathematical Sciences Research Institute, Berkeley, California, 2001.

[10] F. Collino and P. Monk, *The perfectly matched layer in curvilinear coordinates*, SIAM J. Sci. Comput., 19 (1998), pp. 2061–2090.

[11] D. Colton, H. Haddar, and M. Piana, *The linear sampling method in inverse electromagnetic scattering theory*, Inverse Problems, 19 (2003), pp. S105–S138.

[12] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer-Verlag, New York, 1998.

[13] D. Colton and R. Kress, *Eigenvalues of the far-field operator and inverse scattering theory*, SIAM J. Math. Anal., 26 (1995), pp. 601–615.

[14] D. Colton, M. Piana, and R. Potthast, *A simple method using Morozov's discrepancy principle for solving inverse scattering problems*, Inverse Problems, 13 (1997), pp. 1477–1493.

[15] P. Hähner, *On the uniqueness of the shape of a penetrable, anisotropic obstacle*, J. Comput. Appl. Math., 116 (2000) pp. 167–180.

[16] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, UK, 2000.

[17] P. Monk, *Finite Element Methods for Maxwell's Equations*, Oxford University Press, Oxford, UK, 2003.

[18] P. Monk and E. Süli, *The adaptive computation of far-field patterns by a posteriori error estimation of linear functionals*, SIAM J. Numer. Anal., 36 (1998), pp. 251–274.

SIAM J. APPL. MATH.
Vol. 65, No. 3, pp. 790–817

THE EXISTENCE AND STABILITY OF SPIKE PATTERNS IN A CHEMOTAXIS MODEL*

B. D. SLEEMAN†, MICHAEL J. WARD‡, AND J. C. WEI§

Abstract. In the limit of small chemoattractant diffusivity ϵ, the existence, stability, and dynamics of spiky patterns in a chemotaxis model are studied in a bounded multidimensional domain. In this model, the transition probability density function $\Phi(w)$ is assumed to have a power law form $\Phi(w) = w^p$, and the production of chemoattractant w is assumed to saturate according to a Michaelis–Menten kinetic function. In the limit $\epsilon \to 0$, it is proved that there is a steady-state single boundary spike solution located at the maximum of the mean curvature of the boundary. Moreover, a steady-state interior spike solution is proved to concentrate at a maximum of the distance function. The single interior spike solution is shown to be metastable for certain ranges of p and the dimension N. The stability of a single boundary spike solution is also analyzed in detail. Finally, a formal asymptotic analysis is used to characterize the metastable interior spike dynamics in both a one-dimensional and a multidimensional domain.

Key words. chemotaxis, spike patterns, stability, nonlocal eigenvalue problems, metastability

AMS subject classifications. Primary, 35B40, 35B45; Secondary, 35J40

DOI. 10.1137/S0036139902415117

1. Introduction.

In this paper, we investigate qualitative properties of a class of solutions to a special case of the following generalized chemotaxis system. Let $\Omega \subset \mathbb{R}^N$ be a bounded domain with boundary $\partial\Omega$. We seek solutions, $P \in \mathbb{R}$ and $W \in \mathbb{R}^{m+1}$, of the system

$$P_t = D\nabla \cdot \left(P\nabla \left(\log\left(\frac{P}{\Phi(W)} \right) \right) \right), \quad W_t = d\Delta W + F(P, W), \quad (x,t) \in \Omega \times (0,T), \tag{1.1}$$

subject to the "no-flux" boundary condition

$$P\nabla \left(\frac{\log P}{\Phi(W)} \right) \cdot \nu(x) = 0, \qquad \nabla W \cdot \nu(x) = 0. \tag{1.2}$$

Here $\nu(x)$ denotes the inward pointing normal to $\partial\Omega$. To close the system we prescribe the initial conditions

$$P(x,0) = P_0(x) > 0, \qquad W(x,0) = W_0(x) \geq 0, \qquad \text{for } x \in \overline{\Omega}.$$

In this system, D is a constant diffusion coefficient, d is a positive semidefinite diagonal matrix, P is a population density, $\log \Phi(W)$ is the chemotactical sensitivity function,

*Received by the editors September 24, 2002; accepted for publication (in revised form) June 25, 2004; published electronically February 25, 2005.
http://www.siam.org/journals/siap/65-3/41511.html

†School of Mathematics, University of Leeds, Leeds, UK, LS2 9JT (bds@amsta.leeds.ac.uk). The research of this author was supported by a grant from the Royal Society.

‡Department of Mathematics, University of British Columbia, Vancouver, BC, Canada V6T 1Z2 (ward@math.ubc.ca). The research of this author was supported by the NSERC (Canada) under grant 81541.

§Department of Mathematics, Chinese University of Hong Kong, Hong Kong, P.R.C. (wei@math.cuhk.edu.hkj). The research of this author was supported by an earmarked grant from RGC of Hong Kong and a direct grant from CUHK.

and W is a vector of nutrients or chemicals whose dynamics influences the movement of P. The function $\Phi(W)$ is a prescribed "transition probability function." This general system includes the so-called Keller–Segel model of biology [17]. To biologically motivate our study we begin by outlining two themes, one basic to developmental biology and the other from angiogenesis. Our contribution to the mathematical analysis of chemotactic systems is discussed at the end of the introduction.

A basic quality of all living systems is that they sense the environment in which they live and respond to it. The response often involves movement toward or away from an external stimulus. The mechanism for the response is called taxis. Any taxis involves two components: an external signal and the response of the organism to the signal. The response also involves two steps: the detection of the signal and the transduction of the external signal into an internal signal that triggers the response.

In many mathematical models analyzing taxis the signal is transported by diffusion, convection, or some other mechanism. There are, however, instances in which the organism seems to modify its environment in a strictly local manner and there is little or no transport of the modifying substance. A typical example of this is myxobacteria, which produce slime over which their cohorts can easily move. Myxobacteria are soil bacteria which glide on suitable surfaces or at air-water interfaces. Under starvation conditions they tend to move close together, forming complex patterns. Finally they aggregate to build fruiting bodies. Inside the fruiting bodies they survive as dormant myxospores. An account of this intriguing sequence of events can be found in [1].

A novel approach to the modeling of aggregation in myxobacteria is due to Othmer and Stevens [29]. Their inspiration and motivation comes from the work of Davis [6] concerning reinforced random walks. The models developed by Othmer and Stevens are of the form (1.1), (1.2) with $d = 0$ and have been studied in depth by Levine and Sleeman [19], who were able to provide some understanding of the numerical findings in [29], particularly with regard to aggregation, blow-up, and collapse of solutions.

More recently the idea of mathematical modeling based on the idea of reinforced random walks has been developed to gain some understanding of tumor angiogenesis. Angiogenesis is a morphogenetic process whereby new blood vessels are induced to grow out of a pre-existing vasculature. It is fundamental to the formation of any new blood vessels during embryonic development and contributes to the maintenance of tissue functionality in the adult (e.g., placental growth). Angiogenesis is also an important feature of various pathological processes such as wound healing and cancer progression. We are particularly interested in tumor angiogenesis.

Capillaries, which are the main microvessels involved in tumor angiogenesis, are composed of three components: (a) the basement membrane, which is a complex extra cellular matrix (ECM) encircling and supporting the cellular components; (b) the endothelial cells (EC), which form a monolayer of flattened and extended cells lining the lumen of the vessel and resting on the basement membrane; and (c) pericyte cells, which form a periendothelial cellular network embedded within the basement membrane.

The first event of tumor-induced angiogenesis is triggered by the secretion of a number of chemicals, collectively called tumor angiogenesis factors (TAFs) (cf. [4], [8], [9]) from a colony of cancerous cells of a solid tumor. These factors diffuse through the tissue, creating a chemotactic gradient which eventually reaches neighboring capillaries and other blood vessels. In response to TAFs the EC in nearby capillaries appear to thicken (i.e., aggregate) and produce a proteolytic enzyme, which in turn

degrades the basement membrane.

In response to the TAFs the normally smooth EC surface begins to develop pseudopodia which penetrate the weakened basement membrane. Capillary sprouts are formed by the accumulation of EC from the parent vessel. The sprouts grow in length, proliferate, form loops leading to microcirculation of blood (i.e., anastomoses); and also branch successively. The resulting capillary network continues to progress through the tissue ECM, forming a microvasculature, and eventually invades the tumor colony, leading to rapid growth and metastasis. The means of progress through the tissue ECM is via chemotaxis and haptotaxis.

Chemotaxis is the response of EC to chemical gradients set up by the TAFs. A major component of the tissue ECM is fibronectin. It has been verified experimentally that fibronectin stimulates directional migration of EC by establishing an adhesive gradient, i.e., via haptotaxis.

Mathematical modeling of the complex processes involved in tumor angiogenesis has been vigorously pursued in recent years. For recent overviews of some of the modeling ideas, see [36] and [33].

Driven by the need to understand the underlying biochemistry and also to attempt to bridge the gap between micro- and macrocellular events, Sleeman, together with Levine, Pamuk, Nilsen-Hamilton [20], [21], [22], Holmes [14], Plank [31], and Wallis [35], has modeled angiogenesis on the basis of reinforced random walks and Michaelis–Menten kinetics. In these modeling ideas, systems of equations of the form (1.1) play a crucial role.

Systems of the form (1.1) enjoy very rich dynamics. Consider, for example, the following two-component system in one space dimension

$$P_t = D\left(P_{xx} - a\left(P\frac{W_x}{W}\right)_x\right), \quad W_t = \lambda PW - \mu W \qquad \text{for } 0 < x < \ell,\ t > 0,$$

$$\frac{P_x}{P} - a\frac{W_x}{W} = 0 \qquad \text{for } x = 0, \ell,\ t > 0, \tag{1.3}$$

$$P(x,0) = P_0(x) > 0, \qquad W(x,0) = W_0(x) > 0 \qquad \text{for } 0 \le x \le \ell.$$

In [19] it is shown, among other things, that when $a = 1$, there are solution pairs (P, W) for which $P > 0$ but for which P blows up in finite time and that the power spectrum converges to that of the delta function in finite time. Indeed, it is possible to construct an explicit family of such solutions. When $a = -1$, there exist solution pairs (P, W) for which $P > 0$ and P collapses to a constant in infinite time but exponentially fast.

In this context we mention the related work of Rascle and Ziti [34]. In our notation they considered the following system in $\mathbb{R}^N$:

$$P_t = D_1\Delta P - \nabla \cdot [PW^{-\alpha}\nabla W], \quad W_t = D_2\Delta W - kW^m P. \tag{1.4}$$

Here all the constants are positive unless otherwise specified. They constructed similarity solutions of the form $(P, W) = ((T-t)^a P(\xi), (T-t)^b w(\xi))$, where $\xi = (T-t)^{-1}|x|^2$ for $x \in \mathbb{R}^N$ in one, two, or three space dimensions, when $0 < m < \alpha = 1$ and $D_2 = 0$. (Here $a, b \ge 0$.) When $D_1 = 0$ as well, they construct such solutions which blow up in infinite time in one and two dimensions. In the case $D_1 > 0$ they are able to construct only global self-similar solutions.

An important question which again is motivated by the need to understand how new capillaries sprout via angiogenesis from a preexisting vasculature is: Can we expect solutions of system (1.1) (with $d = 0$) to possess spatially nonconstant, piecewise constant aggregating solutions? Here we define aggregation as follows: $P(x,t)$ as a solution to (1.1), aggregates if it converges to a nonconstant steady state in finite or infinite time.

From numerical experiments, Othmer and Stevens [29] show that $P(x,t)$ can evolve to an aggregating solution through the formation of a "shock." In their experiments they consider system (1.1) with

$$\Phi(W) = \left(\frac{\beta + W}{\gamma + W}\right)^a, \qquad F(P,W) = \frac{PW}{1+\nu W} - \mu W + \gamma_r \frac{P}{1+P}. \tag{1.5}$$

In [19] it is argued that the seeds of such shock formation are already present in the case of the simpler system (1.3).

During the initiation of angiogenesis, as outlined above, the EC in capillaries near the tumor produce a proteolytic enzyme in response to the TAFs. While a detailed analysis of the process involved in angiogenesis initiation has been given in [20], a simple model involving only the EC and the proteolytic enzyme can be formulated. Such a model is of the form (1.1), in which P represents EC density and W is enzyme concentration. In this model d is small since in the capillary diffusion takes place on a much longer time scale than the kinetic reactions. This, of course, is not the case for the developing angiogenesis in the ECM.

It has been demonstrated recently by Holash et al. (cf. [13]) that once a tumor has become vascularized, the resulting capillary network may undergo periods of dramatic collapse and remodeling. Paradoxically, the coopted vasculature does not undergo angiogenesis to support the growing tumor, but instead regresses via a process that involves disruption of EC/smooth muscle cell interactions and EC apoptosis (i.e., programmed cell death). This vessel regression in turn results in necrosis within the central part of the tumor. However vigorous angiogenesis is initiated at the tumor boundary, rescuing the surviving tumor and supporting further growth. This behavior could be modeled by systems of the form (1.1) and suggests the existence of point-condensation solutions or spike-type patterns.

It is the purpose of this paper to investigate the existence, stability, and dynamics of the spike patterns in the following variant of (1.1) and (1.5). That is, we consider the system

$$\begin{cases} P_t = D_1 \nabla \cdot \left(P \nabla \left(\log \frac{P}{\Phi(W)} \right) \right), \quad W_t = D_2 \Delta W - \mu W + \frac{PW}{1+\gamma W} & \text{in } \Omega \times (0, +\infty), \\ \frac{\partial P}{\partial \nu} = \frac{\partial W}{\partial \nu} = 0 \quad \text{on } \partial\Omega \times (0, +\infty), \quad P(x,0) = P_0(x) \geq 0, \quad W(x,0) = W_0(x) \geq 0, \end{cases} \tag{1.6}$$

where $D_1 > 0$ and $D_2 > 0$ are two diffusion coefficients, $P(x,t)$ is the particle density of a particular species, $W(x,t)$ is the concentration of the "active agent," $\Omega \subset \mathbb{R}^N (N \leq 3)$ is a smooth and bounded domain, μ and γ are positive constants, and $\nu = \nu(x)$ is the unit normal derivative at $x \in \partial\Omega$. The term $\frac{W}{1+\gamma W}$ is a typical Michaelis–Menten saturating function. Throughout the paper, we take $\Phi(W) = W^p$, where $p > 1$, which corresponds to a logarithmic chemotactical sensitivity function.

We will show that the inclusion of a *small diffusion coefficient* D_2 can produce *stable* spiky patterns. By nondimensionalizing (1.6), we may assume, without loss of

generality, that

$$\mu = 1, \qquad D_1 = 1, \qquad D_2 = \epsilon^2 \ll 1. \tag{1.7}$$

In this new setting, (1.6) becomes

$$\begin{cases} P_t = \nabla \cdot \Big(P \nabla \Big(\log \frac{P}{W^p} \Big) \Big), \quad W_t = \epsilon^2 \Delta W - W + \frac{PW}{1+\gamma W} \quad \text{in } \ \Omega \times (0, +\infty), \\ \frac{\partial P}{\partial \nu} = \frac{\partial W}{\partial \nu} = 0 \quad \text{on } \ \partial\Omega \times (0, +\infty), \quad P(x,0) = P_0(x) \geq 0, \quad W(x,0) = W_0(x) \geq 0. \end{cases} \tag{1.8}$$

For $\gamma \gg 1$, (1.8) is the Keller–Segel model with a *logarithmic* chemotactical sensitivity function. A survey of the existence and regularity properties of solutions of the classic Keller–Segel model with a *linear* chemotactical sensitivity function, and for certain extensions of the basic model, is given in [15]. Results for global existence of solutions for the limiting model (1.8) where $\gamma \gg 1$ are surveyed in section 6.1.1 of [15]. Our results for the stability of spike-type patterns of (1.8) for $\gamma > 0$ are, to our knowledge, new. In particular, for certain ranges of p, we prove that an interior spike solution for (1.8) is metastable. The existence of metastable phenomena is known in certain reaction-diffusion systems, including the shadow Gierer–Meinhardt model (cf. [16] and [5]), but to our knowledge has never been shown previously in a chemotaxis system. After submission of this article, the occurrence of metastability has been shown asymptotically and numerically in [32] for the volume-filling chemotaxis model of [12] and [30].

Integrating the equation for P over Ω and using the Divergence theorem, we obtain

$$\int_\Omega P(x,t) = \int_\Omega P(x,0) = m. \tag{1.9}$$

This conservation of mass condition plays a central role in stabilizing nontrivial spatial patterns. To simplify the computations, we will assume that $m = 1$. The steady-state problem for (1.8) becomes

$$\begin{cases} \nabla \cdot \Big(P \nabla \Big(\log \frac{P}{W^p} \Big) \Big) = 0 & \text{in } \ \Omega, \\ \epsilon^2 \Delta W - W + \frac{PW}{1+\gamma W} = 0 & \text{in } \ \Omega, \\ \frac{\partial P}{\partial \nu} = \frac{\partial W}{\partial \nu} = 0 \text{ on } \ \partial\Omega, & \int_\Omega P(x) = 1. \end{cases} \tag{1.10}$$

Note that both P and W must be nonnegative. From the equation for P in (1.10) and the condition that $m = 1$, we get

$$P(x) = \frac{1}{\int_\Omega W^p} W^p(x). \tag{1.11}$$

Defining the new function $\hat{W}$ by

$$W(x) = \frac{1}{\gamma \int_\Omega \hat{W}^p} \hat{W} \tag{1.12}$$

and substituting (1.11) and (1.12) into (1.10), we obtain an equation for $\hat{W}$:

(1.13)
$$\epsilon^2\Delta\hat{W} - \hat{W} + \frac{\hat{W}^{p+1}}{\int_\Omega \hat{W}^p + \hat{W}} = 0 \quad \text{in } \Omega, \quad \hat{W} > 0 \quad \text{in } \Omega, \quad \frac{\partial \hat{W}}{\partial \nu} = 0 \quad \text{on } \partial\Omega.$$

Equation (1.13) without the nonlocal term $\int_\Omega \hat{W}^p$ has a variational structure and has been studied by numerous authors. For results on the existence of boundary spike solutions, see [2], [27], [28], [40], [43], and the references therein. Results for the existence of interior spike solutions are given in [11], [38], [39], and the references therein. A survey of some of these previous results is given in [26].

Throughout the paper, $C > 0$ is a generic constant, which is independent of ϵ, that may change from line to line. The notation $O(A), o(A)$ means that $|O(A)| \le C|A|$, $\lim_{\epsilon\to 0} \frac{o(A)}{|A|} = 0$.

The organization of the paper is as follows. In section 2, we state our main results. In section 3, we construct both single boundary and single interior spikes. In sections 4–6, we analyze the spectrum of the linearized problem. In the spectrum there are eigenvalues that are $O(1)$ as $\epsilon \to 0$, called the large eigenvalues, and eigenvalues that tend to zero as $\epsilon \to 0$, called the small eigenvalues. In section 4, we study the linearized eigenvalue problem and reduce the study of the large eigenvalues to a nonlocal eigenvalue problem. In section 5, we analyze this nonlocal eigenvalue problem. In section 6, we analyze the small eigenvalues. In section 7, we derive the dynamical law for the motion of an interior spike and present some numerical results. Finally, a brief discussion in given in section 8.

2. Statements of main results. We first state our main results on the existence of steady-state solutions.

THEOREM 2.1. *Assume that*

$$1 < p < +\infty \quad \text{if } N = 1, 2; \qquad 1 < p < 5, \quad \text{if } N = 3. \tag{2.1}$$

Then, for $\epsilon \ll 1$, there exists a steady-state solution for (1.10) *of the following form:*

$$(P_\epsilon, W_\epsilon) = \left(\frac{1}{\int_\Omega \hat{W}_\epsilon^p}\hat{W}_\epsilon^p, \frac{\hat{W}_\epsilon}{\gamma\int_\Omega \hat{W}_\epsilon^p}\right), \quad \text{where} \quad \hat{W}_\epsilon = w\left(\frac{x - Q_\epsilon}{\epsilon}\right) + O(\epsilon). \tag{2.2}$$

Here $w(y)$ is the unique solution of

$$\Delta w - w + w^p = 0, \quad w > 0 \quad \text{in } \mathbb{R}^N, \quad w(0) = \max_{y\in\mathbb{R}^N} w(y), \quad w(y) \to 0 \quad \text{as } |y| \to +\infty. \tag{2.3}$$

The point Q_ϵ is classified either by

(a) *(single boundary spike) $Q_\epsilon \in \partial\Omega$, $H(Q_\epsilon) \to \max_{Q\in\partial\Omega} H(Q)$, where $H(Q)$ is the mean curvature function at $Q \in \partial\Omega$, or*

(b) *(single interior spike) $Q_\epsilon \in \Omega$, $d(Q_\epsilon, \partial\Omega) \to \max_{Q\in\Omega} d(Q, \partial\Omega)$, where $d(Q, \partial\Omega)$ is the distance function at $Q \in \Omega$.*

Next, we study the linearized stability of the solutions constructed in Theorem 2.1. To this end, we linearize (1.8) around (P_ϵ, W_ϵ), as given in Theorem 2.1, to obtain

the following linearized eigenvalue problem:

$$\nabla \cdot \left(\psi_\epsilon \nabla \log \left(\frac{P_\epsilon}{W_\epsilon^p} \right) \right) + \nabla \cdot \left(P_\epsilon \nabla \left(\frac{\psi_\epsilon}{P_\epsilon} - p \frac{\phi_\epsilon}{W_\epsilon} \right) \right) = \lambda_\epsilon \psi_\epsilon \quad \text{in } \Omega, \tag{2.4}$$

$$\epsilon^2 \Delta \phi_\epsilon - \phi_\epsilon + \frac{P_\epsilon}{(1+\gamma W_\epsilon)^2} \phi_\epsilon + \frac{W_\epsilon}{1+\gamma W_\epsilon} \psi_\epsilon = \lambda_\epsilon \phi_\epsilon \quad \text{in } \Omega, \tag{2.5}$$

$$\frac{\partial \phi_\epsilon}{\partial \nu} = \frac{\partial \psi_\epsilon}{\partial \nu} = 0 \quad \text{on } \partial\Omega, \tag{2.6}$$

where $\lambda_\epsilon \in \mathcal{C}$—the set of complex numbers.

Note that (2.4)–(2.6) is *not self-adjoint*, and so complex eigenvalues are expected. We say that (P_ϵ, W_ϵ) is *linearly stable* if for all eigenvalues λ_ϵ of (2.4)–(2.6) we have $Re(\lambda_\epsilon) < 0$. We say that (P_ϵ, W_ϵ) is *linearly unstable* if there exists an eigenvalue λ_ϵ of (2.4)–(2.6) such that $Re(\lambda_\epsilon) > 0$. We say that (P_ϵ, W_ϵ) is *metastable* if for all eigenvalues λ_ϵ of (2.4)–(2.6) we have either $Re(\lambda_\epsilon) < 0$ or $|\lambda_\epsilon| = O(e^{-d/\epsilon})$ for some $d > 0$ independent of $\epsilon > 0$. With these definitions, we now give our main results classifying the stability of (P_ϵ, W_ϵ).

THEOREM 2.2. *Assume that*

$$1 < p < +\infty \quad \text{if } N = 1; \qquad 2 \le p \le 5 \quad \text{if } N = 2; \qquad 2 \le p \le 3 \quad \text{if } N = 3. \tag{2.7}$$

Let (P_ϵ, W_ϵ) be the solution given in Theorem 2.1.

(a) (metastability) *The single interior spike is metastable.*

(b) (stability) *If $N = 1$, then the single boundary spike is linearly stable.*

(c) (stability) *If $\Omega = B_R(0) = \{x \mid |x| < R\}$ and $(P(x,t), W(x,t)) = (P(|x|,t), W(|x|,t))$, then the single interior spike is linearly stable.*

(d) (stability) *If $N = 2, 3$ and Q_0 is a nondegenerate global maximum point of $H(P)$, where $Q_\epsilon \to Q_0$, then the single boundary spike is linearly stable.*

Remark 2.3. The condition on the exponent p given in (2.7) is needed for the analysis in section 5 of a nonlocal eigenvalue problem. Certainly it is not optimal. We conjecture that the same conclusion holds if p satisfies (2.1) only.

Remark 2.4. As $\epsilon \to 0$, we have $\int_\Omega \hat{W}_\epsilon^p \sim \epsilon^N$. Therefore, we see from (2.2) that $P_\epsilon(Q_\epsilon) \sim \epsilon^{-N}$, $W_\epsilon(Q_\epsilon) \sim \epsilon^{-N}$, and $P_\epsilon(x), W_\epsilon(x) \to 0$ for all $x \in \Omega$ with $|x - Q_\epsilon| \ge \delta > 0$.

Remark 2.5. Theorems 2.1 and 2.2 remain true for the following Keller–Segel model with logarithmic growth ([17], [23], [24]):

$$\begin{cases} P_t = \nabla \cdot \left(P \nabla \left(\log \frac{P}{W^p} \right) \right), \quad W_t = \epsilon^2 \Delta W - W + P \quad \text{in } \Omega \times (0, +\infty), \\ \frac{\partial P}{\partial \nu} = \frac{\partial W}{\partial \nu} = 0 \quad \text{on } \partial\Omega \times (0, +\infty), \quad P(x,0) = P_0(x) \ge 0, \quad W(x,0) = W_0(x) \ge 0. \end{cases} \tag{2.8}$$

Remark 2.6. In Theorems 2.1 and 2.2, we have assumed that $D_1 = 1$, where D_1 is the diffusion coefficient of P. Theorem 2.1 holds true for any $D_1 > 0$. It is not difficult to see that Theorem 2.2 also holds, provided that

$$\frac{\epsilon^2}{D_1} \ll 1. \tag{2.9}$$

Biologically, this means that if the active agent W diffuses more slowly than the species P, the species will move toward the boundary and form nontrivial stable spiky patterns. It is unclear what happens if $\frac{\epsilon^2}{D_1} \sim 1$.

Remark 2.7. There may be solutions with multiple spikes. We will not discuss this case here, as most likely multiple spike solutions are unstable.

The existence of spiky patterns for the steady states of Keller–Segel model (2.8) has been established in [23], [27], [28]. Theorem 2.1 establishes the existence of spiky patterns for the more general case (1.8). As far as the authors know, Theorem 2.2 is the first rigorous result on the stability of spiky patterns for a chemotaxis system.

3. Construction of the steady state: Proof of Theorem 2.1. In this section we construct steady-state solutions for (1.8) and prove Theorem 2.1. By the transformation leading to (1.13), we need to find a $\hat{W}$ satisfying (1.13). We do this in two steps.

In the first step, we fix δ small and solve the following problem:

$$\epsilon^2 \Delta \hat{W} - \hat{W} + \frac{\hat{W}^{p+1}}{\delta + \hat{W}} = 0, \quad \hat{W} > 0 \quad \text{in } \Omega, \qquad \frac{\partial \hat{W}}{\partial \nu} = 0 \quad \text{on } \partial\Omega. \tag{3.1}$$

This yields a solution $\hat{W}_{\epsilon,\delta}$. In the second step, we solve the algebraic equation

$$\delta = \int_\Omega \hat{W}^p_{\epsilon,\delta}. \tag{3.2}$$

The first step is more or less standard, but we have to treat the dependence of $\hat{W}$ on δ. In the second step we have to make sure that the function on the right-hand side of (3.2) is continuous in δ. We begin with the following simple but important observation.

LEMMA 3.1. *There exists a unique solution to*

$$\begin{aligned} &\Delta w - w + \frac{w^{p+1}}{\delta + w} = 0, \quad w > 0 \quad \text{in } \mathbb{R}^N, \quad w(0) = \max_{y \in \mathbb{R}^N} w(y), \\ &w(y) \to 0 \quad \text{as } |y| \to +\infty. \end{aligned} \tag{3.3}$$

We call such a solution $w_\delta(y)$. As $\delta \to 0$, we have

$$|w_\delta(y) - w(y)| \le C\delta e^{-\min(1,p-1)|y|}, \tag{3.4}$$

where C is independent of $\delta > 0$ and w is the unique solution of (2.3).

Proof. By the well-known Gidas–Ni–Nirenberg theorem, all solutions to (3.3) are radially symmetric. Let $f_\delta(u) = \frac{u^{p+1}}{\delta+u}$. Then we have $\left(\frac{f_\delta(u)}{u}\right)' \ge 0$. By [18], there exists a unique solution, called $w_\delta(|y|)$, to (3.3). Since p is subcritical and $f_\delta(u) \le u^p$, we see that w_δ is uniformly bounded in δ. By compactness and the uniqueness of w_δ, it follows that as $\delta \to 0$, $w_\delta \to w(y)$, where $w(y)$ is the unique solution of (2.3). This implies that

$$w_\delta(y) < Ce^{-|y|}, \tag{3.5}$$

where C is independent of $\delta > 0$. Next we consider $w_\delta = w + \delta\hat{w}_\delta$. It is easy to see that $\hat{w}_\delta$ satisfies

$$\Delta_y \hat{w}_\delta - \hat{w}_\delta + pw^{p-1}\hat{w}_\delta + \delta^{-1}\left(\frac{(w+\delta\hat{w}_\delta)^{p+1}}{\delta + w_\delta} - w^p - p\delta w^{p-1}\hat{w}_\delta\right) = 0, \tag{3.6}$$

where, by (3.5),

$$\left|\delta^{-1}\left(\frac{(w+\delta\hat{w}_\delta)^{p+1}}{\delta+w_\delta}-w^p-p\delta w^{p-1}\hat{w}_\delta\right)\right|\le Cw_\delta^{p-1}\le Ce^{-(p-1)|y|}. \tag{3.7}$$

Since the operator $L_0:=\Delta-1+pw^{p-1}$ is invertible from $H_r^2(\mathbb{R}^N)=H^2(\mathbb{R}^N)\cap\{u(y)=u(|y|)\}$ to $L_r^2(\mathbb{R}^N)=L^2(\mathbb{R}^N)\cap\{u(y)=u(|y|)\}$ (see Lemma 4 of [45]), we see from (3.6) and (3.7) that

$$\|\hat{w}_\delta\|_{H^2(\mathbb{R}^N)}\le C \quad\text{and}\quad |\hat{w}_\delta|\le Ce^{-\min(1,p-1)|y|}. \qquad \square$$

To analyze single boundary spikes we proceed as follows. For each $\delta>0$ small, we define

$$J_{\epsilon,\delta}[u]=\frac{\epsilon^2}{2}\int_\Omega|\nabla u|^2+\frac{1}{2}\int_\Omega u^2-\int_\Omega F_\delta(u)\qquad\text{for } u\in H^1(\Omega), \tag{3.8}$$

where $F_\delta(u)=\int_0^u\frac{s^{p+1}}{s+\delta}ds$. By taking a function $e(x)\equiv k$ for some constant k in Ω and by choosing k large enough, we have $J_{\epsilon,\delta}(e)<0$ for all $\epsilon\in(0,1)$ and $\delta\in(0,1)$. Then, for each $\epsilon\in(0,1)$ and $\delta\in(0,1)$, we can define the so-called mountain-pass value

$$c_{\epsilon,\delta}=\inf_{h\in\Gamma}\max_{0\le t\le1}J_{\epsilon,\delta}[h(t)], \tag{3.9}$$

where $\Gamma=\{h:[0,1]\to H^1(\Omega)|h(t)\text{ is continuous, }h(0)=0,\,h(1)=u_0\}$.

Since p satisfies (2.1), p is subcritical. Furthermore, $\left(\frac{f_\delta(u)}{u}\right)'\ge0$. Thus f_δ satisfies all the assumptions in [27]. Similar to the analysis in section 2 of [27], $c_{\epsilon,\delta}$ is attained by some function $\hat{W}_{\epsilon,\delta}$, which satisfies (3.1), and $c_{\epsilon,\delta}$ is the least among all nonzero critical values of $J_{\epsilon,\delta}$. Furthermore, the analysis in section 3 of [27] shows that for ϵ sufficiently small, and uniformly in δ, $\hat{W}_{\epsilon,\delta}$ has a unique maximum point $Q_{\epsilon,\delta}$, with $Q_{\epsilon,\delta}\in\partial\Omega$ and $H(Q_{\epsilon,\delta})\to\max_{P\in\partial\Omega}H(P)$ as $\epsilon\to0$. This completes the first step in the proof.

Next we define the following set:

$$\mathcal{S}_{\epsilon,\delta}=\{W\mid W\text{ satisfies (3.1) and }\ J_{\epsilon,\delta}[W]=c_{\epsilon,\delta}\}. \tag{3.10}$$

In other words, $\mathcal{S}_{\epsilon,\delta}$ contains the set of all least energy solutions of (3.1). By our first step above, $\mathcal{S}_{\epsilon,\delta}$ is not empty. Moreover, $\mathcal{S}$ is a compact set (uniformly in δ and ϵ small) since

$$\sup_{W\in\mathcal{S}_{\epsilon,\delta}}\|W\|_\epsilon\le C, \tag{3.11}$$

where C is independent of ϵ and δ and

$$\|W\|_\epsilon^2:=\epsilon^{-N}\left(\epsilon^2\int_\Omega|\nabla W|^2+\int_\Omega W^2\right). \tag{3.12}$$

(In fact, by Lemma 3.1 and by using a test function, we have $c_{\epsilon,\delta}\le C\epsilon^N$, where C is independent of ϵ and δ. Integrating the equation (3.1) gives (3.11).)

We consider the algebraic equation

$$\beta(\delta)=\delta-\rho(\delta)=0,\qquad\text{where}\quad\rho(\delta):=\inf_{W\in\mathcal{S}_{\epsilon,\delta}}\int_\Omega W^p. \tag{3.13}$$

Note that for any $W \in \mathcal{S}_{\epsilon,\delta}$, the same asymptotic analysis holds for mountain-pass solutions (since they have the energy level). Thus we have

$$\int_\Omega W^p = \epsilon^N \left(\frac{1}{2} \int_{\mathbb{R}^N} w_\delta^p + o(1) \right), \tag{3.14}$$

where $o(1) \to 0$ as $\epsilon \to 0$ uniformly in δ. We also remark that, by the compactness of the set $\mathcal{S}_{\epsilon,\delta}$, $\rho(\delta)$ is attained and is a continuous function in δ.

Note that $\beta(0) < 0$ and $\beta(\epsilon^N \int_{\mathbb{R}^N} w^p) > 0$. Therefore, by the mean-value theorem, there exists a $\delta_\epsilon \in (0, \epsilon^N \int_{\mathbb{R}^N} w^p)$ such that $\beta(\delta_\epsilon) = 0$. That is, there exists a $\hat{W}_{\epsilon,\delta_\epsilon} \in \mathcal{S}_{\epsilon,\delta_\epsilon}$ such that $\delta_\epsilon = \int_\Omega \hat{W}^p_{\epsilon,\delta_\epsilon}$.

Let $\hat{W}_\epsilon = \hat{W}_{\epsilon,\delta_\epsilon}$ and $Q_\epsilon = Q_{\epsilon,\delta_\epsilon}$. Then, $\hat{W}_\epsilon$ is a single boundary spike and satisfies the properties stated in Theorem 2.1. This completes our second step for single boundary spikes.

Finally, we use a similar method to discuss the case of single interior spikes. We follow the analysis in [11]. By [11], there exists solution $\hat{W}_{\epsilon,\delta}$ to (3.1), with a single interior spike for ϵ small (uniformly for δ small). Moreover, $\hat{W}_{\epsilon,\delta}$ has a unique local maximum point $Q_{\epsilon,\delta}$ such that $d(Q_{\epsilon,\delta}, \partial\Omega) \to \max_{Q\in\Omega} d(Q, \partial\Omega)$. Now we fix one such solution $\bar{W}_{\epsilon,\delta}$, and consider the following set:

$$\mathcal{S}'_{\epsilon,\delta} = \{u | u \text{ satisfies (3.1) and } \|u - \bar{W}_{\epsilon,\delta}\|_\epsilon \le \epsilon\}. \tag{3.15}$$

Similarly $\mathcal{S}'_{\epsilon,\delta}$ is a nonempty compact set, and the following problem has a solution δ_ϵ:

$$\beta'(\delta) = \delta - \rho'(\delta) = 0, \qquad \text{where} \qquad \rho'(\delta) := \inf_{\hat{W}\in\mathcal{S}'_{\epsilon,\delta}} \int_\Omega \hat{W}^p. \tag{3.16}$$

Let $\hat{W}_\epsilon = \hat{W}_{\epsilon,\delta_\epsilon}$ and $Q_\epsilon = Q_{\epsilon,\delta_\epsilon}$. Then, $\hat{W}_\epsilon$ is a single interior spike solution and satisfies the properties stated in Theorem 2.1. □

Remark 3.2. If $N = 1$, both the single boundary spike solution and the single interior spike solution are unique.

Remark 3.3. If $\Omega = B_R(0)$, (P_ϵ, W_ϵ) can be chosen to be radially symmetric if we restrict our analysis to the class of radially symmetric functions.

We list several properties of $\hat{W}_\epsilon$ for later use. Their proofs can be found in [27], [28], and [11].

LEMMA 3.4. *Let $\hat{W}_\epsilon$ be given in Theorem* 2.1. *Then, we have*

(1)

$$\delta_\epsilon = \int_\Omega \hat{W}^p_\epsilon = \begin{cases} \epsilon^N \left(\displaystyle\int_{\mathbb{R}^N} w^p + o(1) \right) & \text{for } Q_\epsilon \in \Omega, \\ \epsilon^N \left(\dfrac{1}{2} \displaystyle\int_{\mathbb{R}^N} w^p + o(1) \right) & \text{for } Q_\epsilon \in \partial\Omega. \end{cases}$$

(2) $\hat{W}_\epsilon(x) \le Ce^{-c|x-Q_\epsilon|/\epsilon}$ *for some constants* $C, c > 0$.

(3) $\epsilon \frac{|\nabla_x \hat{W}_\epsilon(x)|}{\hat{W}_\epsilon(x)} \ge \sqrt{1-\eta}$ *for* $|x - Q_\epsilon| > \epsilon R$, *where* $0 < \eta < 1$ *is a fixed constant,* R *is large, and* ϵ *is sufficiently small.*

(4)

$$\hat{W}_\epsilon = w_{\delta_\epsilon}\left(\frac{x - Q_\epsilon}{\epsilon} \right) + \begin{cases} O(e^{-d/\epsilon}) & \text{for } Q_\epsilon \in \Omega, \\ O(\epsilon) & \text{for } Q_\epsilon \in \partial\Omega. \end{cases}$$

Here w_{δ_ϵ} is the unique solution of (3.3)*, and $d > 0$ is some constant (independent of ϵ).*

4. Study of the linearized eigenvalue problem. In this section, we begin to study the stability of (P_ϵ, W_ϵ). We consider the interior spike case first. The boundary spike case will be treated later.

We introduce a perturbation around (P_ϵ, W_ϵ) of the following:

$$P(x,t) = P_\epsilon(x) + \eta e^{\lambda_\epsilon t}\psi_\epsilon, \qquad W(x,t) = W_\epsilon(x) + \eta e^{\lambda_\epsilon}\phi_\epsilon. \tag{4.1}$$

Substituting (4.1) into (1.8) and discarding higher order terms, we obtain the following linearized eigenvalue problem:

$$\nabla \cdot \left(P_\epsilon \nabla \left(\frac{\psi_\epsilon}{P_\epsilon} - p\frac{\phi_\epsilon}{W_\epsilon} \right) \right) = \lambda_\epsilon \psi_\epsilon \qquad \text{in } \ \Omega, \tag{4.2}$$

$$\epsilon^2 \Delta\phi_\epsilon - \phi_\epsilon + \frac{P_\epsilon}{(1+\gamma W_\epsilon)^2}\phi_\epsilon + \frac{W_\epsilon}{1+\gamma W_\epsilon}\psi_\epsilon = \lambda_\epsilon \phi_\epsilon \qquad \text{in } \ \Omega, \tag{4.3}$$

$$\frac{\partial \phi_\epsilon}{\partial \nu} = \frac{\partial \psi_\epsilon}{\partial \nu} = 0 \qquad \text{on } \ \partial\Omega, \tag{4.4}$$

where $\lambda_\epsilon \in \mathcal{C}$. Note that the conservation of mass (1.9) requires that $\int_\Omega \psi_\epsilon = 0$.

Recall from (2.2) that

$$P_\epsilon = \frac{1}{\int_\Omega W_\epsilon^p} W_\epsilon^p = \frac{1}{\int_\Omega \hat{W}_\epsilon^p}\hat{W}_\epsilon^p, \qquad W_\epsilon = \frac{1}{\gamma \int_\Omega \hat{W}_\epsilon^p}\hat{W}_\epsilon, \tag{4.5}$$

where $\hat{W}_\epsilon$ satisfies (1.13) and has all the properties listed in Lemma 3.4.

We begin by simplifying (4.2) and (4.3). We introduce $\tilde{\psi}_\epsilon$ by

$$\psi_\epsilon = p\frac{P_\epsilon}{W_\epsilon}\phi_\epsilon - \eta_\epsilon P_\epsilon + \tilde{\psi}_\epsilon, \tag{4.6}$$

where η_ϵ is a constant and $\tilde{\psi}_\epsilon$ satisfies

$$\int_\Omega \tilde{\psi}_\epsilon = 0. \tag{4.7}$$

Since $\int_\Omega \psi_\epsilon = 0$, from (4.6), we obtain that $\eta_\epsilon = (\int_\Omega P_\epsilon)^{-1} p \int_\Omega \frac{P_\epsilon}{W_\epsilon}\phi_\epsilon$. Therefore, using (4.5), we get

$$\psi_\epsilon = p\gamma \hat{W}_\epsilon^{p-1}\phi_\epsilon - p\gamma \frac{\int_\Omega \hat{W}_\epsilon^{p-1}\phi_\epsilon}{\int_\Omega \hat{W}_\epsilon^p}\hat{W}_\epsilon^p + \tilde{\psi}_\epsilon. \tag{4.8}$$

Substituting (4.5) and (4.8) into (4.2), we obtain that

$$\nabla \cdot \left(P_\epsilon \nabla \left(\frac{\tilde{\psi}_\epsilon}{P_\epsilon} \right) \right) = \lambda_\epsilon \left(p\gamma \hat{W}_\epsilon^{p-1}\phi_\epsilon - p\gamma \frac{\int_\Omega \hat{W}_\epsilon^{p-1}\phi_\epsilon}{\int_\Omega \hat{W}_\epsilon^p}\hat{W}_\epsilon^p + \tilde{\psi}_\epsilon \right). \tag{4.9}$$

We introduce local coordinates y and $\hat{\psi}_\epsilon$ by

$$x = Q_\epsilon + \epsilon y, \qquad y \in \Omega_\epsilon := \{y | \epsilon y + Q_\epsilon \in \Omega_\epsilon\}, \qquad \tilde{\psi}_\epsilon(x) = \epsilon^2 \lambda_\epsilon \gamma \hat{W}_\epsilon^{p/2} \hat{\psi}_\epsilon(y), \tag{4.10}$$

where Q_ϵ is the unique maximum point of $\hat{W}_\epsilon$. We still use $\hat{W}_\epsilon, \phi_\epsilon$, etc. to denote the functions $\hat{W}_\epsilon, \phi_\epsilon$, etc. under the new coordinate y.

A simple computation shows that (4.7) and (4.9) become

$$\int_{\Omega_\epsilon} \hat{W}_\epsilon^{\frac{p}{2}} \hat{\psi}_\epsilon(y) = 0, \tag{4.11}$$

$$\Delta_y \hat{\psi}_\epsilon - h(\hat{W}_\epsilon)\hat{\psi}_\epsilon - \epsilon^2 \lambda_\epsilon \hat{\psi}_\epsilon = p \hat{W}_\epsilon^{\frac{p}{2}-1} \phi_\epsilon - p \frac{\int_\Omega \hat{W}_\epsilon^{p-1} \phi_\epsilon}{\int_\Omega \hat{W}_\epsilon^p} \hat{W}_\epsilon^{\frac{p}{2}}, \qquad y \in \Omega_\epsilon, \tag{4.12}$$

where $h(\hat{W}_\epsilon) = \frac{\Delta_y \hat{W}_\epsilon^{p/2}}{\hat{W}_\epsilon^{p/2}}$. Substituting (4.5) and (4.8) into (4.3), we obtain that ϕ_ϵ satisfies

$$\Delta_y \phi_\epsilon - \phi_\epsilon + f'_{\delta_\epsilon}(\hat{W}_\epsilon)\phi_\epsilon - p \frac{\int_{\Omega_\epsilon} \hat{W}_\epsilon^{p-1} \phi_\epsilon}{\int_{\Omega_\epsilon} \hat{W}_\epsilon^p} \frac{\hat{W}_\epsilon^{p+1}}{\delta_\epsilon + \hat{W}_\epsilon} = \lambda_\epsilon \left(\phi_\epsilon - \frac{\epsilon^2 \hat{W}_\epsilon^{1+p/2} \hat{\psi}_\epsilon}{\delta_\epsilon + \hat{W}_\epsilon} \right), \quad y \in \Omega_\epsilon, \tag{4.13}$$

where $f'_{\delta_\epsilon}(\hat{W}_\epsilon)$ is defined by

$$f'_{\delta_\epsilon}(\hat{W}_\epsilon) = \frac{p \hat{W}_\epsilon^p}{\delta_\epsilon + \hat{W}_\epsilon} + \frac{\delta_\epsilon \hat{W}_\epsilon^p}{(\delta_\epsilon + \hat{W}_\epsilon)^2}. \tag{4.14}$$

The boundary condition (4.4) becomes

$$\frac{\partial \phi_\epsilon}{\partial \nu_\epsilon} = \frac{\partial \hat{\psi}_\epsilon}{\partial \nu_\epsilon} = 0 \qquad \text{on } \partial\Omega_\epsilon, \tag{4.15}$$

where ν_ϵ is the unit normal derivative of $\partial\Omega_\epsilon$ at y.

We now need to solve the reformulated eigenvalue problem (4.12) and (4.13), subject to (4.11) and (4.15). We begin with the following simple observation.

LEMMA 4.1. *There exists a constant $C > 0$ such that for ϵ sufficiently small we have*

$$\int_{\Omega_\epsilon} (|\nabla \psi|^2 + h(\hat{W}_\epsilon)\psi^2) \geq C \int_{\Omega_\epsilon} \psi^2 \quad \forall \psi \in H^1(\Omega_\epsilon) \quad \textit{such that} \quad \int_{\Omega_\epsilon} \hat{W}_\epsilon^{p/2} \psi = 0. \tag{4.16}$$

Proof. We just need to show that the principal eigenvalue ν_1^ϵ of

$$\Delta_y \psi^\epsilon - h(\hat{W}_\epsilon)\psi^\epsilon = \nu_1^\epsilon \psi^\epsilon \quad \text{in } \Omega_\epsilon, \quad \int_{\Omega_\epsilon} \hat{W}_\epsilon^{p/2} \psi^\epsilon = 0, \quad \frac{\partial \psi^\epsilon}{\partial \nu_\epsilon} = 0 \quad \text{on } \partial\Omega_\epsilon \tag{4.17}$$

satisfies $\nu_1^\epsilon < -C < 0$. Suppose not. Multiplying (4.17) by ψ_ϵ and integrating over Ω_ϵ, we obtain

$$\nu_1^\epsilon \int_{\Omega_\epsilon} (\psi^\epsilon)^2 = \int_{\Omega_\epsilon} \left(\nabla \cdot \left(P_\epsilon \nabla \left(\frac{\psi^\epsilon}{\sqrt{P_\epsilon}} \right) \right) \right) \frac{\psi^\epsilon}{\sqrt{P_\epsilon}} = - \int_{\Omega_\epsilon} P_\epsilon \left| \nabla \cdot \left(\frac{\psi^\epsilon}{\sqrt{P_\epsilon}} \right) \right|^2. \tag{4.18}$$

Hence $\nu_1^\epsilon \le 0$ and $\nu_1^\epsilon = 0$ if and only if $\frac{\psi^\epsilon}{\sqrt{P_\epsilon}}$ is identically a constant. Since $\int_{\Omega_\epsilon} \psi^\epsilon \sqrt{P_\epsilon} = 0$, we see that $\nu_1^\epsilon < 0$. Suppose now that as $\epsilon \to 0$ we have $\nu_1^\epsilon < 0$ with $\nu_1^\epsilon \to 0$. We proceed to derive a contradiction.

We calculate

$$h(\hat{W}_\epsilon) = \frac{p}{2}\left(\frac{p}{2} - 1\right)\frac{|\nabla \hat{W}_\epsilon|^2}{\hat{W}_\epsilon^2} + \frac{p}{2} - \frac{p}{2}\frac{\hat{W}_\epsilon^p}{\delta_\epsilon + \hat{W}_\epsilon}. \tag{4.19}$$

By Lemma 3.4 we see that, for ϵ sufficiently small, the inequalities

$$\frac{|\nabla \hat{W}_\epsilon|}{\hat{W}_\epsilon} \ge \sqrt{1-\eta}, \qquad \frac{\hat{W}_\epsilon^p}{\delta_\epsilon + \hat{W}_\epsilon} < \eta \tag{4.20}$$

hold for any η small and $|y|$ large. Hence, for $|y|$ large, we have

$$h(\hat{W}_\epsilon) \ge \frac{p}{2}\left(\frac{p}{2} - 1\right)(1-\eta) + \frac{p}{2} - \frac{p}{2}\eta = \frac{p^2}{4}(1-\eta).$$

Therefore, by the maximum principle, we get

$$|\psi^\epsilon(y)| \le C\|\psi^\epsilon\|_{H^1(\Omega_\epsilon)} e^{-\delta|y|}, \tag{4.21}$$

where $\delta = \frac{p^2}{8}(1-\eta)$. By (4.21) and a compactness argument we can now take a subsequence $\epsilon \to 0$ such that $\psi^\epsilon \to \psi^0$ in $H^1(\Omega_\epsilon)$, and ψ^0 satisfies

$$\Delta_y \psi^0 - h(w^{p/2})\psi^0 = 0, \qquad \psi^0 \in H^1(\mathbb{R}^N), \qquad \int_{\mathbb{R}^N} w^{p/2}\psi^0 = 0.$$

This is impossible by the same reasoning leading to (4.18). (Note that $\psi^0 \in H^1(\mathbb{R}^N)$ and w decays exponentially fast.) □

From Lemma 4.1, we have the following result.

LEMMA 4.2. *Let λ_ϵ be such that $\mathrm{Re}(\lambda_\epsilon) \ge 0$. Then, there exists a constant $C > 0$ such that $|\lambda_\epsilon| \le C$, uniformly for ϵ small.*

Proof. Multiplying (4.12) by the conjugate function of $\hat{\psi}_\epsilon$, labeled by $\overline{\hat{\psi}_\epsilon}$, and integrating over Ω_ϵ, we obtain

$$\begin{aligned}&\left|\int_{\Omega_\epsilon}(|\nabla\hat{\psi}_\epsilon|^2 + h(\hat{W}_\epsilon)|\hat{\psi}_\epsilon|^2 + \epsilon^2\lambda_\epsilon|\hat{\psi}_\epsilon|^2)\right| \\ &= \left|p\int_{\Omega_\epsilon}\hat{W}_\epsilon^{p/2-1}\phi_\epsilon\overline{\hat{\psi}_\epsilon} - p\frac{\int_{\Omega_\epsilon}\hat{W}_\epsilon^{p-1}\phi_\epsilon}{\int_{\Omega_\epsilon}\hat{W}_\epsilon^p}\int_{\Omega_\epsilon}\hat{W}_\epsilon^{p/2}\overline{\hat{\psi}_\epsilon}\right| \le C\|\hat{\psi}_\epsilon\|_{L^2(\Omega_\epsilon)}\|\phi_\epsilon\|_{L^2(\Omega_\epsilon)}.\end{aligned} \tag{4.22}$$

Now applying Lemma 4.1 and the fact that $\mathrm{Re}(\lambda_\epsilon) \ge 0$, we arrive at

$$\|\hat{\psi}_\epsilon\|_{H^1(\Omega_\epsilon)} \le C\|\phi_\epsilon\|_{L^2(\Omega_\epsilon)}. \tag{4.23}$$

Multiplying (4.13) by $\bar{\phi}_\epsilon$, the conjugate function of ϕ_ϵ, and integrating over Ω_ϵ, we get

$$\left|\int_{\Omega_\epsilon}(|\nabla\phi_\epsilon|^2 + |\phi_\epsilon|^2 + \lambda_\epsilon|\phi_\epsilon|^2)\right| \le C\left(\int_{\Omega_\epsilon}|\phi_\epsilon|^2 + \epsilon^2|\lambda_\epsilon|\int_{\Omega_\epsilon}|\phi_\epsilon||\hat{\psi}_\epsilon|\right). \tag{4.24}$$

This yields that $|\lambda_\epsilon| \le C$, using (4.23). ∎

A corollary of Lemma 4.2 is the following.

COROLLARY 4.3. *Let λ_ϵ be such that* $\mathrm{Re}(\lambda_\epsilon) \ge 0$. *Assume that for a subsequence $\epsilon \to 0$, $\lambda_\epsilon \to \lambda_0$. Then λ_0 is an eigenvalue of the nonlocal eigenvalue problem*

$$\Delta_y \phi_0 - \phi_0 + p w^{p-1}\phi_0 - p\frac{\int_{\mathbb{R}^N} w^{p-1}\phi_0}{\int_{\mathbb{R}^N} w^p} w^p = \lambda_0 \phi_0, \quad \phi_0 \in H^1(\mathbb{R}^N), \tag{4.25}$$
$$\lambda_0 \in \mathcal{C}, \qquad \mathrm{Re}(\lambda_0) \ge 0.$$

Proof. By Lemma 4.2, we have $|\lambda_\epsilon| \le C$. We may assume that for a subsequence $\epsilon \to 0$, $\lambda_\epsilon \to \lambda_0$. Assume that $\|\phi_\epsilon\|_{H^1(\Omega_\epsilon)} = 1$. Then, from Lemma 4.1, we have

$$\|\hat{\psi}_\epsilon\|_{H^1(\Omega_\epsilon)} \le C, \qquad \|\hat{\psi}_\epsilon\|_{L^\infty(\Omega_\epsilon)} \le C.$$

By taking a limit in (4.13), we see that $\phi_\epsilon \to \phi_0$ in $H^1(\Omega_\epsilon)$ and $\lambda_\epsilon \to \lambda_0$, where (λ_0, ϕ_0) satisfies (4.25). ∎

Finally we discuss the boundary spike case. Let $Q_\epsilon \in \partial\Omega$ be the global maximum point $\hat{W}_\epsilon$. Without loss of generality we may assume from now on that $Q_\epsilon = 0$ and that the normal derivative at Q_ϵ is $\nu(Q_\epsilon) = (0, \dots, -1)$. Similarly as in the interior spike case, we have the following conclusion.

COROLLARY 4.4. *Let (P_ϵ, W_ϵ) be a single boundary spike at Q_ϵ. Let λ_ϵ be such that $Re(\lambda_\epsilon) \ge 0$. Assume that for a subsequence $\epsilon \to 0$, $\lambda_\epsilon \to \lambda_0$. Then, λ_0 is an eigenvalue of the nonlocal eigenvalue problem*

$$\Delta_y \phi_0 - \phi_0 + p w^{p-1}\phi_0 - p\frac{\int_{\mathbb{R}^N_+} w^{p-1}\phi_0}{\int_{\mathbb{R}^N_+} w^p} w^p = \lambda_0 \phi_0 \quad \text{in } \mathbb{R}^N_+, \tag{4.26}$$
$$\frac{\partial \phi_0}{\partial y_N} = 0 \quad \text{on } \partial\mathbb{R}^N_+, \quad \phi_0 \in H^1(\mathbb{R}^N_+),$$

with $Re(\lambda_0) \ge 0$, where $\mathbb{R}^N_+ = \{(y', y_N) \in \mathbb{R}^N | y_N > 0\}$.

Let ϕ_0 be a solution of (4.26) on $\mathbb{R}^N_+$. By an even extension of ϕ_0 to $\mathbb{R}^N$, it is easy to see that the new function, denoted by $\tilde{\phi}_0$, satisfies (4.25).

5. Study of a nonlocal eigenvalue problem. In this section, we study the following nonlocal eigenvalue problem derived in Corollary 4.3:

(5.1)
$$L\phi := \Delta_y \phi - \phi + p w^{p-1}\phi - p\frac{\int_{\mathbb{R}^N} w^{p-1}\phi}{\int_{\mathbb{R}^N} w^p} w^p = \lambda_0 \phi, \qquad \phi \in H^1(\mathbb{R}^N),\ Re(\lambda_0) \ge 0,$$

where w is the unique solution of (2.3) and $\lambda_0 \in \mathcal{C}$ is the set of complex numbers. Nonlocal eigenvalues of this type have been studied in several papers. For the case of $N = 1$, we refer to [7] and [41]. For the case of $p = 2$, we refer to [5] and [41]. In the general (p, N) case, we follow an approach in [45].

We first characterize the kernel of L as follows.

LEMMA 5.1. *We have*

$$\{\phi \in H^1(\mathbb{R}^N) | L\phi = 0\} = K_0 := span\ \left\{ \frac{\partial w}{\partial y_j}, j = 1, \dots, N \right\},$$

$$\left\{\phi \in H^1(\mathbb{R}^N_+) | L\phi = 0,\ \frac{\partial \phi}{\partial y_N} = 0 \text{ on } \partial\mathbb{R}^N_+\right\} = \text{ span } \left\{\frac{\partial w}{\partial y_j}, j = 1, \ldots, N-1\right\}.$$

Proof. The proof is similar to that of Lemma 5.1 of [41]. We omit the details here. ∎

From Lemma 5.1, we may assume that $\lambda_0 \neq 0$. The following result was proved in [44].

LEMMA 5.2. *Assume that $N = 1$ and $1 < p < +\infty$. Then, for any nonzero eigenvalue λ_0 of (5.1), we have $Re(\lambda_0) \leq -C < 0$ for some constant $C > 0$.*

We are left with the case of $N = 2, 3$. We now use a continuation argument, similar to [45], where p is a continuation parameter. For $p = 2$, (5.1) was studied in [41]. Applying Theorem 1 of [45], we have the next result.

THEOREM 5.3. *Suppose that p satisfies (2.7) in Theorem 2.2. Then, for any nonzero eigenvalue λ_0 of (5.1), we have $\mathrm{Re}(\lambda_0) \leq -C < 0$ for some constant $C > 0$.*

Proof. Let $r = p$ and $\gamma = \frac{p}{p-1}$. Using (2.7), it is easy to see that $F(p) = 1 - \frac{p-1}{2p}N \geq 0$. Applying Theorem 1 of [45], we just need to check that

$$F(p) \geq \frac{\gamma - 2}{\gamma} F(p+1) + \frac{|\gamma - 2|}{\gamma}\sqrt{F(p+1)(F(p+1) - F(2))}, \tag{5.2}$$

where $F(r) = 1 - \frac{p-1}{2r}N$. Note that for $2 \leq p$, we have $\frac{\gamma-2}{\gamma} = \frac{2-p}{p} \leq 0$.

If $N = 2$, we have $F(p) = \frac{1}{p}, F(p+1) = \frac{2}{p+1}, F(2) = \frac{3-p}{2}$. By simple computations, (5.2) is equivalent to $p^2 - 6p + 5 \leq 0$, which holds if $2 \leq p \leq 5$.

If $N = 3$, we have $F(p) = \frac{3-p}{2p}, F(p+1) = \frac{5-p}{2(p+1)}, F(2) = \frac{7-3p}{4}$. By simple computations, $\frac{\gamma-2}{\gamma}F(p+1) + \frac{|\gamma-2|}{\gamma}\sqrt{F(p+1)(F(p+1) - F(2))} \leq 0 \leq F(p)$ if $2 \leq p \leq 3$. ∎

6. Study of the small eigenvalues: Proof of Theorem 2.2. In this section, we study the asymptotic behavior of the small eigenvalues that tend to zero as $\epsilon \to 0$. We also prove Theorem 2.2.

Suppose that p satisfies (2.7) and that λ_ϵ is an eigenvalue with $\mathrm{Re}(\lambda_\epsilon) \geq 0$. Then, by Corollaries 4.3 and 4.4, Lemma 5.1, and Theorem 5.3, we must have that $\lim_{\epsilon\to 0}\lambda_\epsilon = \lambda_0 = 0$. Namely, if $\mathrm{Re}(\lambda_\epsilon) \geq 0$, then necessarily $\lambda_\epsilon \to 0$. By Lemma 5.1, we have $\phi_0 \in \text{span}\,\{\frac{\partial w}{\partial y_j}, j = 1, \ldots, N\}$ for $Q_\epsilon \in \Omega$, and $\phi_0 \in \text{span}\,\{\frac{\partial w}{\partial y_j}, j = 1, \ldots, N-1\}$ for $Q_\epsilon \in \partial\Omega$. This result is summarized as follows.

LEMMA 6.1. *Suppose that p satisfies (2.7). Let λ_ϵ be an eigenvalue of (4.2) and (4.3) with $\mathrm{Re}(\lambda_\epsilon) \geq 0$. Then, for $\epsilon \to 0$ and $y \in \Omega_\epsilon$, we must have*

$$\phi_\epsilon = \begin{cases} \sum_{j=1}^N a_j^\epsilon \frac{\partial w}{\partial y_j}(y) + o(1) & \text{if} \quad Q_\epsilon \in \Omega, \\ \sum_{j=1}^{N-1} a_j^\epsilon \frac{\partial w}{\partial y_j}(y) + o(1) & \text{if} \quad Q_\epsilon \in \partial\Omega. \end{cases} \tag{6.1}$$

Lemma 6.1 immediately implies that for $N = 1$ the single boundary spike is linearly stable. Next, consider the radially symmetric case where $(P(x,t), W(x,t)) = (P(|x|,t), W(|x|,t))$. Then $\phi_0(y) = \phi(|y|)$, and by Lemma 6.1 we conclude that $\phi_0 = 0$. Hence, there are no small eigenvalues. Moreover, if p satisfies (2.7), then $\mathrm{Re}(\lambda_0) < 0$ for $\lim_{\epsilon\to 0}\lambda_\epsilon = \lambda_0 \neq 0$. Therefore, we conclude that in the radially symmetric case (P_ϵ, W_ϵ) is linearly stable. This proves (b) and (c) of Theorem 2.2.

Now we prove (a) of Theorem 2.2. Let (P_ϵ, W_ϵ) be a single interior spike solution. We now show that (P_ϵ, W_ϵ) is metastable. Namely, we need to show that for $\mathrm{Re}(\lambda_\epsilon) \geq 0$ we must have $|\lambda_\epsilon| = O(e^{-d/\epsilon})$.

Suppose now that $\mathrm{Re}(\lambda_\epsilon) \geq 0$. From (4.13), the equation for ϕ_ϵ becomes

(6.2)
$$\Delta_y \phi_\epsilon - \phi_\epsilon + f'_{\delta_\epsilon}(\hat{W}_\epsilon)\phi_\epsilon - p\frac{\int_{\Omega_\epsilon} \hat{W}_\epsilon^{p-1}\phi_\epsilon}{\int_{\Omega_\epsilon} \hat{W}_\epsilon^{p}} \frac{\hat{W}_\epsilon^{p+1}}{\delta_\epsilon + \hat{W}_\epsilon} = \lambda_\epsilon \left(\phi_\epsilon - \epsilon^2 \frac{\hat{W}_\epsilon^{1+p/2}}{\delta_\epsilon + \hat{W}_\epsilon}\hat{\psi}_\epsilon \right), \quad y \in \Omega_\epsilon.$$

We introduce the new function $\hat{\phi}_\epsilon$ by

$$\phi_\epsilon = \hat{\phi}_\epsilon + c_\epsilon \hat{W}_\epsilon, \qquad \text{where} \qquad c_\epsilon = -p\frac{\int_{\Omega_\epsilon} \hat{W}_\epsilon^{p-1}\hat{\phi}_\epsilon}{\int_{\Omega_\epsilon} \hat{W}_\epsilon^{p}}. \tag{6.3}$$

Then, it is easy to see that $\hat{\phi}_\epsilon$ satisfies

$$\Delta_y \hat{\phi}_\epsilon - \hat{\phi}_\epsilon + f'_{\delta_\epsilon}(\hat{W}_\epsilon)\hat{\phi}_\epsilon + c_\epsilon \frac{\delta_\epsilon \hat{W}_\epsilon^{p+1}}{(\delta_\epsilon + \hat{W}_\epsilon)^2} = \lambda_\epsilon \left(\hat{\phi}_\epsilon + c_\epsilon \hat{W}_\epsilon - \epsilon^2 \frac{\hat{W}_\epsilon^{1+p/2}}{\delta_\epsilon + \hat{W}_\epsilon}\hat{\psi}_\epsilon \right). \tag{6.4}$$

Let $\eta(x)$ be a smooth cut-off function such that $\eta(x) = 1$ for $|x| \leq 1$, and $\eta(x) = 0$ for $|x| > 2$. Set $r = \frac{1}{4}d(Q_\epsilon, \partial\Omega)$. Consider the following functions:

$$\phi_{\epsilon,j}(y) = \frac{\partial w_{\delta_\epsilon}}{\partial y_j}(y)\eta\left(\frac{\epsilon y}{r}\right), \qquad j = 1, \ldots, N, \qquad y \in \Omega_\epsilon. \tag{6.5}$$

Then, by Lemma 3.4, we have

(6.6)
$$\Delta_y \phi_{\epsilon,j} - \phi_{\epsilon,j} + f'_{\delta_\epsilon}(\hat{W}_\epsilon)\phi_{\epsilon,j} = O(e^{-d/\epsilon}), \quad \int_{\Omega_\epsilon} \phi_{\epsilon,j}\phi_{\epsilon,k} = O(e^{-d/\epsilon}) \qquad \text{for } \ j \neq k.$$

Next, we decompose $\hat{\phi}_\epsilon$ as follows:

$$\hat{\phi}_\epsilon = \sum_{j=1}^{N} c_j^\epsilon \phi_{\epsilon,j} + \hat{\phi}_\epsilon^\perp,$$

where $\hat{\phi}_\epsilon^\perp \perp \phi_{\epsilon,j}$ for $j = 1, \ldots, N$ and $\sum_{j=1}^N |c_j^\epsilon|^2 = 1$. Lemma 6.1 implies $\|\hat{\phi}_\epsilon^\perp\|_\epsilon \to 0$ as $\epsilon \to 0$.

Note that from Lemma 3.4 we have

$$\left| \int_{\Omega_\epsilon} \hat{W}_\epsilon^{p-1}\phi_{\epsilon,j} \right| + \left| \int_{\Omega_\epsilon} \hat{W}_\epsilon^{p}\phi_{\epsilon,j} \right| = O(e^{-d/\epsilon}), \quad j = 1, \ldots, N. \tag{6.7}$$

The proof of part (a) of Theorem 2.2 proceeds in two steps. First, we obtain the estimates for $\hat{\phi}_\epsilon^\perp$. Then, we deduce the equation for λ_ϵ. Similar arguments have been used in section 4 of [41].

It is easy to see that $\hat{\phi}_\epsilon^\perp$ satisfies

$$\Delta_y \hat{\phi}_\epsilon^\perp - \hat{\phi}_\epsilon^\perp + f'_{\delta_\epsilon}(\hat{W}_\epsilon)\hat{\phi}_\epsilon^\perp = -c_\epsilon \frac{\delta_\epsilon \hat{W}_\epsilon^{p+1}}{(\delta_\epsilon + \hat{W}_\epsilon)^2} \tag{6.8}$$

$$+\sum_{j=1}^{N} c_j^\epsilon O(e^{-d/\epsilon}) + \lambda_\epsilon \left(\hat{\phi}_\epsilon + \sum_{j=1}^{N} c_j^\epsilon \phi_{\epsilon,j} + c_\epsilon \hat{W}_\epsilon - \epsilon^2 \frac{\hat{W}_\epsilon^{1+p/2}}{\delta_\epsilon + \hat{W}_\epsilon} \hat{\psi}_\epsilon \right).$$

Note that from the definition of c_ϵ, we have that

$$c_\epsilon = -p \frac{\int_{\Omega_\epsilon} \hat{W}_\epsilon^{p-1} (\sum_{j=1}^{N} c_j^\epsilon \phi_{\epsilon,j} + \hat{\phi}_\epsilon^\perp)}{\int_{\Omega_\epsilon} \hat{W}_\epsilon^p} = O\left(e^{-d/\epsilon} + \left| \int_{\Omega_\epsilon} \hat{W}_\epsilon^{p-1} \hat{\phi}_\epsilon^\perp \right| \right). \tag{6.9}$$

Let us define

(6.10)
$$\mathcal{K}_\epsilon = \text{span}\,\{\phi_{\epsilon,j}, j = 1, \ldots, N\} \subset H^2_{\nu_\epsilon}(\Omega_\epsilon), \quad \mathcal{C}_\epsilon = \text{span}\,\{\phi_{\epsilon,j}, j = 1, \ldots, N\} \subset L^2(\Omega_\epsilon),$$

where $H^2_{\nu_\epsilon}(\Omega_\epsilon) = H^2(\Omega_\epsilon) \cap \{\frac{\partial u}{\partial \nu_\epsilon} = 0 \text{ on } \partial\Omega_\epsilon\}$. Let $\mathcal{K}_\epsilon^\perp$ and $\mathcal{C}_\epsilon^\perp$ be the orthogonal space of $\mathcal{K}_\epsilon$ and $\mathcal{C}_\epsilon$, respectively, under the $L^2(\Omega_\epsilon)$ inner product. Set

$$L_\epsilon := \Delta_y \phi - \phi + f'_{\delta_\epsilon}(\hat{W}_\epsilon)\phi : \qquad H^2_{\nu_\epsilon}(\Omega_\epsilon) \to L^2(\Omega_\epsilon). \tag{6.11}$$

Then, as in Lemma 2.3 of [41], we have that the map $\mathcal{L}_\epsilon = \pi_\epsilon \circ L_\epsilon : \mathcal{K}_\epsilon^\perp \to \mathcal{C}_\epsilon^\perp$ is invertible and the inverse is bounded uniformly in ϵ. Here π_ϵ is the projection from $L^2(\Omega_\epsilon)$ into $\mathcal{C}_\epsilon^\perp$.

From (6.8) and (4.23) we obtain

$$\|\hat{\phi}_\epsilon^\perp\|_{H^2(\Omega_\epsilon)} \le C \left(|c_\epsilon|\delta_\epsilon + |c_\epsilon||\lambda_\epsilon| + e^{-d/\epsilon} + \epsilon^2 |\lambda_\epsilon| \right).$$

From (6.9), this implies that

$$\|\hat{\phi}_\epsilon^\perp\|_{H^2(\Omega_\epsilon)} \le C(e^{-d/\epsilon} + \epsilon^2 |\lambda_\epsilon|). \tag{6.12}$$

Multiplying (6.8) by $\phi_{\epsilon,k}$ and integrating over Ω_ϵ, we obtain

(6.13)
$$\begin{aligned} \lambda_\epsilon \left(\sum_{j=1}^{N} c_j^\epsilon \int_{\Omega_\epsilon} \phi_{\epsilon,j}\phi_{\epsilon,k} + O(\epsilon^2) \right) &= O(e^{-d/\epsilon}) + \int_{\Omega_\epsilon} \phi_{\epsilon,k} \left(\Delta \hat{\phi}_\epsilon^\perp - \hat{\phi}_\epsilon^\perp + f'_{\delta_\epsilon}(\hat{W}_\epsilon)\hat{\phi}_\epsilon^\perp \right) \\ &= O(e^{-d/\epsilon}). \end{aligned}$$

This shows that $|\lambda_\epsilon| = O(e^{-d/\epsilon})$, which proves part (a) of Theorem 2.2.

Finally, we prove (d) of Theorem 2.2. Let us assume that $Q_\epsilon \to Q_0$, where Q_0 is a nondegenerate global maximum point of $H(P)$. For single boundary spikes, we have from Theorem 1.3 of [42] or Theorem 2.2 of [3] that the eigenvalue problem

$$\Delta_y \phi - \phi + f'_{\delta_\epsilon}(\hat{W}_\epsilon)\phi = \tau_\epsilon \phi \quad \text{in } \Omega_\epsilon, \qquad \frac{\partial \phi}{\partial \nu_\epsilon} = 0 \quad \text{on } \partial\Omega_\epsilon, \tag{6.14}$$

has $N-1$ normalized eigenfunctions $\{\phi_{\epsilon,j}, j = 1, \ldots, N-1\}$ with N eigenvalues (multiplicity is allowed)

$$\tau_1^\epsilon \le \cdots \le \tau_{N-1}^\epsilon, \qquad \tau_j^\epsilon = c_0 \epsilon^2 \lambda_j + o(\epsilon^2). \tag{6.15}$$

Here λ_j is the eigenvalue of the matrix $(\nabla^2 H(Q_0))$, and $c_0 > 0$ is a generic constant. Moreover, we have

$$\int_{\Omega_\epsilon} \phi_{\epsilon,j}^2 = 1, \qquad \int_{\Omega_\epsilon} \phi_{\epsilon,j}\phi_{\epsilon,k} = 0, \quad \text{for } j \neq k, \tag{6.16}$$
$$\phi_{\epsilon,j}(y) = \sum_{k=1}^{N-1} a_{jk} \frac{\partial w}{\partial y_k} + O(\epsilon), \qquad j = 1, \dots, N-1,$$

for some constants a_{jk}.

Similar to the analysis above, we now have from Lemma 3.4 that

$$\hat{W}_\epsilon(y) = w_{\delta_\epsilon}(y) + O(\epsilon), \quad \left| \int_{\Omega_\epsilon} \hat{W}_\epsilon^{p-1} \phi_{\epsilon,j} \right| + \left| \int_{\Omega_\epsilon} \hat{W}_\epsilon^{p} \phi_{\epsilon,j} \right| = O(\epsilon), \quad j = 1, \dots, N-1. \tag{6.17}$$

We decompose

$$\hat{\phi}_\epsilon = \sum_{j=1}^{N-1} c_j^\epsilon \phi_{\epsilon,j} + \hat{\phi}_\epsilon^\perp, \qquad \hat{\phi}_\epsilon^\perp \perp \phi_{\epsilon,j}, \quad j = 1, \dots, N-1.$$

Hence, we have

$$c_\epsilon = -p \frac{\int_{\Omega_\epsilon} \hat{W}_\epsilon^{p-1} (\sum_{j=1}^N c_j^\epsilon \phi_{\epsilon,j} + \hat{\phi}_\epsilon^\perp)}{\int_{\Omega_\epsilon} \hat{W}_\epsilon^p} = O\left(\epsilon + \left| \int_{\Omega_\epsilon} \hat{W}_\epsilon^{p-1} \hat{\phi}_\epsilon^\perp \right| \right). \tag{6.18}$$

From (6.8) we obtain that

$$\begin{aligned} \Delta_y \hat{\phi}_\epsilon^\perp - \hat{\phi}_\epsilon^\perp + f_\delta'(\hat{W}_\epsilon)\hat{\phi}_\epsilon^\perp &= O\left(\epsilon\delta_\epsilon + \delta_\epsilon \int_{\Omega_\epsilon} \hat{W}_\epsilon^{p-1} |\hat{\phi}_\epsilon^\perp| \right) \\ &\quad - \sum_{j=1}^{N-1} c_j^\epsilon \tau_j^\epsilon \phi_{\epsilon,j} + \lambda_\epsilon \left(\sum_{j=1}^{N-1} c_j^\epsilon \phi_{\epsilon,j} + \hat{\phi}_\epsilon^\perp + c_\epsilon \hat{W}_\epsilon - \epsilon^2 \frac{\hat{W}_\epsilon^{1+p/2}}{\delta_\epsilon + \hat{W}_\epsilon} \hat{\psi}_\epsilon \right). \end{aligned} \tag{6.19}$$

Similar arguments leading to (6.12) imply that

$$\|\hat{\phi}_\epsilon^\perp\|_{H^2(\Omega_\epsilon)} \leq C \left(\epsilon\delta_\epsilon + \epsilon^2 |\lambda_\epsilon| + |c_\epsilon||\lambda_\epsilon| \right). \tag{6.20}$$

Multiplying (6.19) by $\phi_{\epsilon,k}$ and integrating over Ω_ϵ, we obtain

$$\lambda_\epsilon \left(\sum_{j=1}^{N-1} c_j^\epsilon \int_{\Omega_\epsilon} \phi_{\epsilon,j}\phi_{\epsilon,k} + O(\epsilon^2) \right) = O(\epsilon\delta_\epsilon) + \sum_{j=1}^{N-1} c_j^\epsilon \tau_j^\epsilon \int_{\Omega_\epsilon} \phi_{\epsilon,j}\phi_{\epsilon,k}. \tag{6.21}$$

Therefore,

$$\lambda_\epsilon (c_k^\epsilon + O(\epsilon^2)) = O(\epsilon^{N+1}) + c_k^\epsilon \tau_k^\epsilon. \tag{6.22}$$

Since $\sum_{k=1}^N |c_k^\epsilon|^2 = 1$, we see that there exists some k such that $\lambda_\epsilon = \tau_k^\epsilon + o(\epsilon^2)$. Since $\tau_k^\epsilon < 0$, we see that $\mathrm{Re}(\lambda_\epsilon) \leq -c_1\epsilon^2 < 0$, where $c_1 > 0$ is a generic constant. This finishes the proof of part (d) of Theorem 2.2. ∎

7. Metastable dynamics and numerical results. In this section, we assume that p satisfies the conditions (2.7) under Theorem 5.3. We then use a formal asymptotic analysis to characterize the metastable dynamics of an interior spike solution for (1.8). The metastability analysis is similar to that given in [16] for an interior spike solution to the Gierer–Meinhardt model of [10]. We look for a solution to (1.8) in the form

(7.1)
$$W(x,t) = W_\epsilon\left(\epsilon^{-1}|x-x_0|\right) + R(x,t), \qquad P(x,t) = P_\epsilon\left(\epsilon^{-1}|x-x_0|\right) + H(x,t).$$

Here (W_ϵ, P_ϵ) satisfy the PDEs of (1.10), and the error terms R and W are such that $R \ll W_\epsilon$ and $H \ll P_\epsilon$. In (7.1), $x_0 = x_0(t) \in \Omega$ is the unknown location of the center of the spike. Our goal is to derive an equation of motion for $x_0(t)$. We will only consider the long-time evolution of the spike, and do not discuss the transient process by which a spike forms from initial data. Therefore, we assume that at $t = 0$ we have $x_0(0) = x_0^0 \in \Omega$ and $R(x,0) = H(x,0) = 0$. Since the linearized problem has an exponentially small principal eigenvalue, we expect that the speed $x_0^{'}$ is exponentially small as $\epsilon \to 0$.

Substituting (7.1) into (1.8) and linearizing the resulting system, we obtain

(7.2a)
$$\nabla\cdot\left(P_\epsilon\nabla\left(\frac{H}{P_\epsilon} - \frac{pR}{W_\epsilon}\right)\right) = \partial_t P_\epsilon + \partial_t H \quad \text{in } \Omega; \qquad \frac{\partial R}{\partial \nu} = -\frac{\partial W_\epsilon}{\partial \nu} \quad \text{on } \Omega,$$

(7.2b)
$$\epsilon^2\Delta R - R + \frac{P_\epsilon}{(1+\gamma W_\epsilon)^2}R + \frac{W_\epsilon}{1+\gamma W_\epsilon}H = \partial_t W_\epsilon + \partial_t R; \qquad \frac{\partial H}{\partial \nu} = -\frac{\partial P_\epsilon}{\partial \nu} \quad \text{on } \partial\Omega.$$

Since both $\partial H/\partial\nu$ on $\partial\Omega$ and the right-hand side of the PDE in (7.2a) are small, we can asymptotically solve (7.2a) for H to get

(7.3)
$$H \sim \left(\frac{pP_\epsilon}{W_\epsilon}\right)R - pP_\epsilon\int_\Omega \frac{P_\epsilon R}{W_\epsilon}.$$

Substituting (7.3) into (7.2b), we obtain

(7.4)
$$\epsilon^2\Delta R - R + f^{'}(W_\epsilon)R - \frac{pW_\epsilon P_\epsilon}{1+\gamma W_\epsilon}\int_\Omega \frac{P_\epsilon R}{W_\epsilon} = \partial_t W_\epsilon + \partial_t R \quad \text{in } \Omega,$$
$$\frac{\partial R}{\partial \nu} = -\frac{\partial W_\epsilon}{\partial \nu} \quad \text{on } \partial\Omega.$$

Here $f^{'}(W_\epsilon)$ is defined by

(7.5)
$$f^{'}(W_\epsilon) = \frac{P_\epsilon}{(1+\gamma W_\epsilon)^2} + \frac{pP_\epsilon}{1+\gamma W_\epsilon}.$$

Next, we use (2.2) to write (7.4) as

(7.6a)
$$\mathcal{L}_\epsilon R \equiv \epsilon^2\Delta R - R + f^{'}_{\delta_\epsilon}(\hat{W}_\epsilon)R - \frac{p\hat{W}_\epsilon^{p+1}}{\delta_\epsilon + \hat{W}_\epsilon}\frac{\int_\Omega \hat{W}_\epsilon^{p-1}R}{\int_\Omega \hat{W}_\epsilon^{p}} = \partial_t W_\epsilon + \partial_t R \quad \text{in } \Omega,$$

(7.6b)
$$\frac{\partial R}{\partial \nu} = -\frac{\partial W_\epsilon}{\partial \nu} \quad \text{on } \partial\Omega.$$

Here δ_ϵ and $f'_{\delta_\epsilon}(\hat{W}_\epsilon)$ are defined in Lemmas 3.4 and 4.14, respectively, and $\hat{W}_\epsilon$ satisfies (1.13).

We define the local operator in (7.6) as

$$L_\epsilon R \equiv \epsilon^2 \Delta R - R + f'_{\delta_\epsilon}(\hat{W}_\epsilon) R. \tag{7.7}$$

By translation invariance, we find upon differentiating (1.13) that

$$L_\epsilon \left(\partial_{x_j} \hat{W}_\epsilon \right) = 0, \qquad j = 1, \ldots, N. \tag{7.8}$$

In addition, since $x_0 \in \Omega$ and $\hat{W}_\epsilon$ is locally radially symmetric near x_0, then $\mathcal{L}_\epsilon(\partial_{x_j} \hat{W}_\epsilon)$ is exponentially small as $\epsilon \to 0$. Moreover, $\partial_{x_j} \hat{W}_\epsilon$ is exponentially small on $\partial\Omega$ for $j = 1, \ldots, N$.

From part 4 of Lemma 3.4, we recall that

$$\hat{W}_\epsilon = w_{\delta_\epsilon}\left[\epsilon^{-1}|x - x_0|\right] + O(e^{-d/\epsilon}), \tag{7.9}$$

where $d > 0$ is some constant independent of ϵ. Here w_{δ_ϵ} satisfies (3.3), with $\delta_\epsilon = \int_\Omega \hat{W}_\epsilon^p = O(\epsilon^N)$. The far-field behavior of the solution, valid for $|x - x_0| \gg O(\epsilon)$, is

$$\hat{W}_\epsilon \sim a \left(\frac{|x - x_0|}{\epsilon} \right)^{(1-N)/2} e^{-|x-x_0|/\epsilon}. \tag{7.10}$$

Here a is a positive constant that depends on N, p, and ϵ. However, as $\epsilon \to 0$ we have $a \to a_0 > 0$, where a_0 is determined from the far-field behavior of (2.3), which corresponds to setting $\delta = 0$ in (3.3). In section 7.1 we derive an ODE for $x_0(t)$ for the multidimensional case where $N \geq 2$. The one-dimensional case is studied in section 7.2.

7.1. The multidimensional case. To derive an equation of motion for $x_0(t)$, we first must determine the eigenfunctions of $\mathcal{L}_\epsilon$ in (7.6) corresponding to the exponentially small eigenvalues. Let $(\lambda_{0j}, \phi_{0j})$, for $j = 1, \ldots, N$, be the eigenpairs of $\mathcal{L}_\epsilon \phi_0 = \lambda_0 \phi_0$, where λ_{0j} is exponentially small as $\epsilon \to 0$. From (7.8), we note that these eigenfunctions are given asymptotically, in the interior of the domain, by $\phi_{0j} \sim \partial_{x_j} \hat{W}_\epsilon$ for $j = 1, \ldots, N$. However, as in [16], we must insert a boundary layer correction term near Ω to ensure that ϕ_{0j} satisfies the homogeneous boundary condition $\partial\phi_{0j}/\partial\nu = 0$ on $\partial\Omega$. In order to resolve the boundary layer, we define a local coordinate system. Let $\hat{\eta}$ represent the distance from a point in Ω to $\partial\Omega$, where $\hat{\eta} < 0$ corresponds to the interior of Ω. Let s correspond to the other $N - 1$ orthogonal coordinates. To localize the region near $\partial\Omega$, we let $\eta = \epsilon^{-1}\hat{\eta}$. The eigenfunction is then approximated by

$$\phi_{0j} \sim \partial_{x_j} \hat{W}_\epsilon + \hat{\phi}_j, \qquad j = 1, \ldots, N. \tag{7.11}$$

Substituting (7.11) into (7.6a), we obtain that $\hat{\phi}_j$ satisfies

$$\partial_{\eta\eta}\hat{\phi}_j - \phi_j = 0, \quad \eta < 0, \qquad \partial_\eta \hat{\phi}_j = -\epsilon \partial_{\hat{\eta}}(\partial_{x_j} \hat{W}_\epsilon)|_{\eta=0} \quad \text{on} \quad \eta = 0, \tag{7.12}$$

with $\hat{\phi}_j \to 0$ as $\eta \to -\infty$. Below, we require a formula for ϕ_{0j} on $\partial\Omega$. Solving (7.12) and using the far-field form (7.10), we substitute the resulting expression into (7.11) to get for $j = 1, \ldots, N$ that

$$\phi_{0j} \sim -a\epsilon^{(N-3)/2} r^{-(1+N)/2} (x_j - x_{0j})\, e^{-r/\epsilon} (1 + \hat{r} \cdot \hat{n}) \qquad \text{on} \quad \Omega. \tag{7.13}$$

Here x_j denotes the jth coordinate of x, $r \equiv |x - x_0|$, $\hat{r} = (x - x_0)/r$, and $\hat{n}$ is the unit outward normal to $\partial\Omega$. A similar calculation was done in [16] with regards to metastable behavior in the Gierer–Meinhardt model [10]. For further details of the calculation leading to (7.13), see [16].

Next, we multiply (7.6a) by ϕ_{0j} and integrate over Ω. Integrating the resulting equation by parts over Ω and assuming that $\partial_t R$ is asymptotically small, we obtain

$$(\partial_t W_\epsilon, \phi_{0j}) = -\epsilon^2 \int_{\partial\Omega} \phi_{0j}\, \partial_\nu \hat{W}_\epsilon \, dS + (R, \mathcal{L}_\epsilon{}^* \phi_{0j})\,. \tag{7.14}$$

We now evaluate the terms in (7.14) using $W_\epsilon = C\hat{W}_\epsilon$, for some constant C. The dominant contribution to the integral on the left-hand side of (7.14) arises from the region near $x = x_0$. Using $\phi_{0j} \sim \partial_{x_j}\hat{W}_\epsilon$, $W_\epsilon = C\hat{W}_\epsilon$, and (7.9) for $\hat{W}_\epsilon$, we calculate for $j = 1, \ldots, N$ that

$$(\partial_t W_\epsilon, \phi_{0j}) \sim -C\epsilon^{N-2} x_{0j}^{'} \left(\frac{\omega_N \beta_N}{N}\right), \qquad \beta_N \equiv \int_0^\infty \rho^{N-1} \left[w_{\delta_\epsilon}^{'}(\rho)\right]^2 \, d\rho. \tag{7.15}$$

Here $(u, v) \equiv \int_\Omega uv$, $x_{0j}^{'} \equiv dx_{0j}/dt$, $w_{\delta_\epsilon}(\rho)$ is the radially symmetric solution to (3.3), and ω_N is the surface area of the unit N-sphere. Next, we use (7.10) and (7.13) to estimate the boundary integral term in (7.14) as

$$-\epsilon^2 \int_{\partial\Omega} \phi_{0j}\, \partial_\nu \hat{W}_\epsilon \, dS \sim -Ca^2\epsilon^{N-1} \int_{\partial\Omega} r^{1-N} e^{-2r/\epsilon} \frac{(x_j - x_{0j})}{r} \hat{r}\cdot\hat{n}\,(1 + \hat{r}\cdot\hat{n})\; dS. \tag{7.16}$$

This expression shows that the boundary integral term in (7.14) is $O(\epsilon^q e^{-2r_0/\epsilon})$, where q is some constant, and $r_0 = \operatorname{dist}(x_0, \partial\Omega)$.

We now show that the inner product term on the right-hand side of (7.14) is asymptotically negligible in comparison with (7.16). We calculate, using $\delta_\epsilon = O(\epsilon^N)$ and symmetry, that

$$\mathcal{L}_\epsilon{}^* \phi_{0j} \sim -\frac{p\hat{W}_\epsilon^{p-1}}{\int_\Omega \hat{W}_\epsilon^p} \int_\Omega \frac{\hat{W}_\epsilon^{p+1}}{\delta_\epsilon + \hat{W}_\epsilon} \partial_{x_j}\hat{W}_\epsilon \sim -\frac{p\hat{W}_\epsilon^{p-1}}{N\,(p+1)\int_\Omega \hat{W}_\epsilon^p} \int_\Omega \nabla\cdot\hat{W}_\epsilon^{p+1}. \tag{7.17}$$

Then, using the divergence theorem and the far-field behavior (7.10), we obtain

$$\mathcal{L}_\epsilon{}^* \phi_{0j} \sim -\frac{p\hat{W}_\epsilon^{p-1}}{N\int_\Omega \hat{W}_\epsilon^p} \int_{\partial\Omega} \frac{\hat{W}_\epsilon^{p+1}}{p+1} = \hat{W}_\epsilon^{p-1}\, O\left(\epsilon^q e^{-(p+1)r_0/\epsilon}\right). \tag{7.18}$$

Here q is a constant, and $r_0 = \operatorname{dist}(x_0, \partial\Omega)$. Since $R \ll 1$ and $p > 1$, it follows that the exponentially small term $(R, \mathcal{L}_\epsilon{}^* \phi_{0j})$ is asymptotically negligible in comparison with the boundary integral term in (7.16).

Therefore, substituting (7.15) and (7.16) into (7.14) and neglecting $(R, \mathcal{L}_\epsilon{}^* \phi_{0j})$ in (7.14), we obtain

$$\frac{dx_0}{dt} \sim \frac{a^2 N\epsilon}{\beta_N \omega_N} \int_{\partial\Omega} r^{1-N} e^{-2r/\epsilon}\, \hat{r}\; (1 + \hat{r}\cdot\hat{n})\; \hat{r}\cdot\hat{n}\, dS. \tag{7.19}$$

Assuming that there is a unique point $x_m \in \partial\Omega$ closest to the initial center $x_0(0)$ of the spike, we can evaluate the surface integral in (7.19) using Laplace's method. This leads to the following explicit result.

PROPOSITION 7.1. *For $\epsilon \ll 1$, a metastable spike solution for* (1.8) *is given by*

$$W_\epsilon \sim \frac{w_{\delta_\epsilon}\left[\epsilon^{-1}(x-x_0)\right]}{\gamma\epsilon^N \int_{\mathbb{R}^N}[w_{\delta_\epsilon}(y)]^p\, dy}, \qquad P_\epsilon \sim \frac{\left(w_{\delta_\epsilon}\left[\epsilon^{-1}(x-x_0)\right]\right)^p}{\epsilon^N \int_{\mathbb{R}^N}[w_{\delta_\epsilon}(y)]^p\, dy}. \tag{7.20a}$$

Here w_{δ_ϵ} satisfies (3.3). *Let x_m be the point on $\partial\Omega$ closest to $x_0(0)$. Then, for $t > 0$, the spike moves in the direction of x_m, and the distance $r_m(t) = |x_m - x_0(t)|$ satisfies the first order nonlinear differential equation*

$$\frac{dr_m}{dt} \sim -\xi r_m \left(\frac{\epsilon}{r_m}\right)^{(N+1)/2} K(r_m) e^{-2r_m/\epsilon}, \tag{7.20b}$$

where $\xi > 0$ and the functions $K(r_m)$ are defined by

$$\xi \equiv \frac{2Na^2}{\omega_N \beta_N}\pi^{(N-1)/2}, \qquad K(r_m) \equiv \left(1-\frac{r_m}{R_1}\right)^{-1/2}\left(1-\frac{r_m}{R_2}\right)^{-1/2}\cdots\left(1-\frac{r_m}{R_{N-1}}\right)^{-1/2}. \tag{7.20c}$$

Here $R_j > 0$, for $j = 1, \ldots, N-1$, are the principal radii of curvature of $\partial\Omega$ at x_m, ω_N is the surface area of the unit N-sphere, and a and β_N were defined in (7.10) *and* (7.15), *respectively.*

This result is valid up until the time when the spike approaches to within an $O(\epsilon)$ distance of x_m. If the initial condition for (7.20b) is $r_m(0) = r_0$, then the time T needed for $r_m(T) = 0$, is readily found to be

$$T \sim \frac{\epsilon^{(1-N)/2} r_0^{(N-1)/2}}{2K(r_0)\xi} e^{2r_0/\epsilon}. \tag{7.21}$$

7.2. The one-dimensional case. Let (λ_0, ϕ_0) be the principal eigenpair of $\mathcal{L}_\epsilon\phi_0 = \lambda_0\phi_0$ with $\phi_0'(\pm 1) = 0$. Here $\mathcal{L}_\epsilon$ is defined in (7.6), and λ_0 is exponentially small. Similar to the analysis for the multidimensional case, ϕ_0 has the boundary layer form

$$\phi_0 \sim \partial_x \hat{W}_\epsilon + \phi_l\left[\epsilon^{-1}(1+x)\right] + \phi_r\left[\epsilon^{-1}(1-x)\right]. \tag{7.22}$$

Substituting (7.22) into (7.6), we obtain that the boundary layer correction terms $\phi_l(\eta)$ and $\phi_r(\eta)$ satisfy $v'' - v = 0$. Imposing that $\phi_0'(\pm 1) = 0$, and using the far-field behavior (7.10) with $N = 1$, we get

$$\phi_l(\eta) = a\epsilon^{-1}e^{-(1+x_0)/\epsilon}e^{-\eta}, \qquad \phi_r(\eta) = -a\epsilon^{-1}e^{-(1-x_0)/\epsilon}e^{-\eta}. \tag{7.23}$$

Combining (7.10), (7.22), and (7.23), and where a is defined in (7.10), we calculate

$$\phi_0(-1) \sim 2\epsilon^{-1}ae^{-(1+x_0)/\epsilon}, \qquad \phi_0(+1) \sim -2\epsilon^{-1}ae^{-(1-x_0)/\epsilon}. \tag{7.24}$$

To derive a differential equation for x_0, we proceed as in the multidimensional case. We multiply both sides of (7.6) by ϕ_0 and integrate over the domain. Assuming that $\partial_t R$ on the right-hand side of (7.6) is asymptotically small, we integrate the resulting expression by parts, and then use the boundary condition (7.6b), to obtain

$$(\phi_0, \partial_t W_\epsilon) = -\epsilon^2\phi_0\partial_x W_e|_{-1}^1 + (R, \mathcal{L}_\epsilon{}^*\phi_0). \tag{7.25}$$

Here $\mathcal{L}_\epsilon{}^*$ denotes the adjoint operator of $\mathcal{L}_\epsilon$. As in the multidimensional case, the second term on the right-hand side of (7.25) is asymptotically negligible in comparison to the other two terms in (7.25). Then, since $W_\epsilon = C\hat{W}_\epsilon$ for some constant C, (7.25) reduces to

$$\left(\phi_0, \partial_t \hat{W}_\epsilon\right) \sim -\epsilon^2 \phi_0 \hat{W}_{\epsilon x}|_{-1}^{1}. \tag{7.26}$$

Using (7.9), (7.10), (7.22), and (7.24), we calculate

$$\left(\phi_0, \partial_t \hat{W}_\epsilon\right) \sim -\epsilon^{-1} x_0^{'} \int_{-\infty}^{\infty} \left[\hat{W}_\epsilon^{'}(y)\right]^2 dy, \tag{7.27}$$

$$\epsilon^2 \phi_0 \hat{W}_{\epsilon x}|_{-1}^{1} \sim 2a^2 \left(e^{-2(1-x_0)/\epsilon} - e^{-2(1+x_0)/\epsilon}\right).$$

Substituting (7.28) into (7.26), we obtain the ODE for the spike location. The corresponding asymptotic solution is obtained by combining (2.2) and part 4 of Lemma 3.4. We summarize the result as follows.

PROPOSITION 7.2. *For $\epsilon \ll 1$, a metastable spike solution for* (1.8) *is given by*

$$W_\epsilon \sim \frac{w_{\delta_\epsilon}\left[\epsilon^{-1}(x-x_0)\right]}{\gamma\epsilon \int_{-\infty}^{\infty} \left[w_{\delta_\epsilon}(y)\right]^p \, dy}, \qquad P_\epsilon \sim \frac{\left(w_{\delta_\epsilon}\left[\epsilon^{-1}(x-x_0)\right]\right)^p}{\epsilon \int_{-\infty}^{\infty} \left[w_{\delta_\epsilon}(y)\right]^p \, dy}. \tag{7.28a}$$

Here w_{δ_ϵ} satisfies (3.3), *and the spike location $x_0(t)$ satisfies the differential equation*

$$\frac{dx_0}{dt} \sim \frac{2a^2\epsilon}{\beta}\left(e^{-2(1-x_0)/\epsilon} - e^{-2(1+x_0)/\epsilon}\right), \quad x_0(0) = x_0^0, \qquad \beta \equiv \int_{-\infty}^{\infty} \left[w_{\delta_\epsilon}^{'}(y)\right]^2 dy. \tag{7.28b}$$

The constant $a > 0$ is defined in (7.10).

Following ideas in [37], the homoclinic orbit constants a, β, and δ_ϵ can be computed numerically in terms of the nonlinearity of (3.3). In this way, we obtain the explicit formulae for a and β:

$$\log a \equiv \log w_m + \int_0^{w_m} \left(\frac{1}{\left[\eta^2 - 2Q(\eta)\right]^{1/2}} - \frac{1}{\eta}\right) d\eta, \quad \beta = 2\int_0^{w_m} \left[\eta^2 - 2Q(\eta)\right]^{1/2} d\eta. \tag{7.29a}$$

Here $w_m = w_{\delta_\epsilon}(0)$, which denotes the maximum of $w_{\delta_\epsilon}(y)$ on $y \geq 0$, satisfies the transcendental equation

$$w_m^2 = 2Q(w_m), \qquad \text{where} \qquad Q(w) = \int_0^w \frac{y^{p+1}}{\delta_\epsilon + y}\, dy. \tag{7.29b}$$

The nonlocal term δ_ϵ, defined in Lemma 3.4, can be rewritten by determining a formula for $w_{\delta_\epsilon}^{'}$ in terms of w_{δ_ϵ} for $y \geq 0$. In this way, we get

$$\delta_\epsilon = 2\epsilon \int_0^{w_m} \frac{w^p}{\sqrt{w^2 - 2Q(w)}}\, dw. \tag{7.29c}$$

We use Newton's method to solve the coupled system (7.29b) and (7.29c) for w_m and δ_ϵ for various values of ϵ. In terms of these values, we then calculate a and β from

TABLE 1
Numerical values for the homoclinic orbit constants at different values of ϵ when $p = 2$.

ϵ	w_m	δ_ϵ	a	β
0.01	1.587	0.0625	5.34	1.38
0.02	1.679	0.137	5.27	1.57
0.03	1.776	0.225	5.28	1.79
0.04	1.880	0.331	5.34	2.03
0.05	1.993	0.458	5.45	2.31
0.06	2.115	0.610	5.59	2.64
0.07	2.248	0.793	5.77	3.01
0.08	2.396	1.017	5.98	3.45
0.09	2.559	1.286	6.22	3.97
0.10	2.741	1.627	6.51	4.60

TABLE 2
Numerical values for the homoclinic orbit constants for different values of p when $\epsilon = 0.06$.

p	w_m	w_{m0}	δ_ϵ	δ_0	a	β
2	2.115	1.500	0.610	0.360	5.59	2.64
3	1.604	1.414	0.351	0.267	2.85	1.77
4	1.456	1.357	0.276	0.228	2.17	1.59
5	1.379	1.316	0.240	0.207	1.89	1.51

(7.29a). An important feature of these formulae is that they do not require an explicit pointwise expression for the solution $w_{\delta_\epsilon}(y)$ to (3.3).

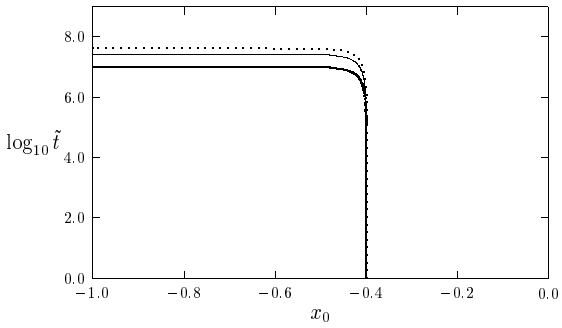
When $\epsilon \to 0$, then $\delta_\epsilon \to 0$. Consequently, in this limit, the solution to (3.3) is given to leading order by the solution to (2.3), which is given explicitly by

$$w(y) = \left(\frac{p+1}{2}\right)^{1/(p-1)} \left(\cosh\left[\frac{(p-1)y}{2}\right]\right)^{-2/(p-1)}. \tag{7.30}$$

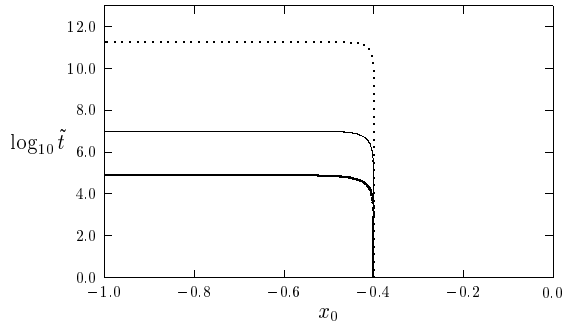
Therefore, as $\epsilon \to 0$, we have for $p = 2$ that $w_m \to w_{m0} = 1.5$, $\delta_\epsilon \to \delta_0 = \epsilon \int_{-\infty}^{\infty} w^2\, dy = 6\epsilon$, $a \to a_0 = 6$, and $\beta \to \beta_0 = 6/5$. In Table 1 we give numerical values for the homoclinic orbit constants for different values of ϵ when $p = 2$. In Table 2 numerical values are given for these constants for different values of p when $\epsilon = 0.06$. In the latter table we compare the numerical values for w_m and δ_ϵ with the numerical values for the corresponding leading order approximations w_{m0} and δ_0 at different values of p. These values show that the nonlocal term δ_ϵ is significant at this value of ϵ.

The ODE (7.28b) shows that the spike moves exponentially slowly towards the right or the left boundary when $x_0(0) > 0$ or $x_0(0) < 0$, respectively. In Figure 1(a) we plot the numerical solution to (7.28b) in the form $\log_{10}(1 + t)$ versus x_0 when $\epsilon = 0.06$ for $p = 2$, $p = 3$, and $p = 4$. The homoclinic orbit constants needed in (7.28b) are given in Table 2. The initial value for (7.28b) is $x_0(0) = -0.4$. The spike motion is found to be slower for larger values of p. For $p = 2$ and the initial value $x_0(0) = -0.4$, in Figure 1(b) we plot $\log_{10}(1 + t)$ versus x_0 for $\epsilon = 0.04$, $\epsilon = 0.06$, and $\epsilon = 0.08$. Notice the dramatic change in the time-scale of the metastability for a small change in ϵ. For $p = 2$ and $\epsilon = 0.06$, in Figure 2(a) and (b) we plot the leading order solutions for W_ϵ and P_ϵ, respectively, at different times. These leading order solutions are obtained by replacing w_{δ_ϵ} in (7.28a) by w as given in (7.30).

Finally, we compare full numerical results for the evolution of an interior spike for (1.8) with the asymptotic dynamical behavior given in (7.28). The routine D03PCF of the NAG library [25] is used to compute the numerical solution to (1.8). The initial



(a) $\log_{10}\tilde{t}$ versus x_0: $\tilde{t} = 1 + t$.



(b) $\log_{10}\tilde{t}$ versus x_0: $\tilde{t} = 1 + t$.

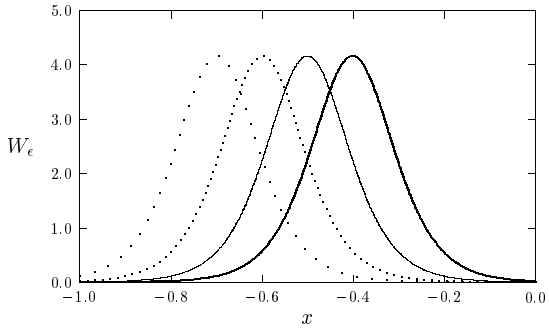
FIG. 1. (a): *Plot of asymptotic spike location computed from* (7.28*b*) *when* $\epsilon = 0.06$*, for* $p = 2$ *(heavy solid curve),* $p = 3$ *(solid curve), and* $p = 4$ *(dashed curve).* (b): *Similar plot when* $p = 2$*, for* $\epsilon = 0.08$ *(heavy solid curve),* $\epsilon = 0.06$ *(solid curve), and* $\epsilon = 0.04$ *(dashed curve).*

condition for the numerical solution to (1.8) is

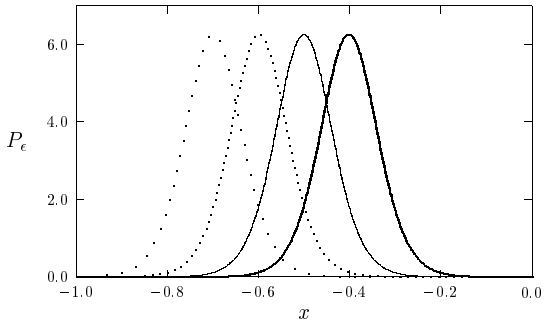
$$W(x,0) = \frac{w\left[\epsilon^{-1}(x - x_0^0)\right]}{\gamma\epsilon\int_{-\infty}^{\infty}[w(y)]^p\,dy}, \qquad P(x,0) = \frac{\left(w\left[\epsilon^{-1}(x - x_0^0)\right]\right)^p}{\epsilon\int_{-\infty}^{\infty}[w(y)]^p\,dy}, \tag{7.31}$$

where w is given in (7.30). In the computations, we choose $p = 2$, $\gamma = 1$, $\epsilon = 0.06$, and $x_0^0 = -0.4$. The spike location is determined numerically from the maximum of the numerical solution for W. In Figure 3, we show a favorable comparison between the asymptotic and numerical results for x_0. For this value of ϵ, the relative error between the asymptotic and numerical predictions for the time at which the spike hits the boundary at $x = -1$ is about 3%.

8. Concluding remarks. Motivated by the general system (1.1), (1.2) in which we have a single equation governing the population density P coupled with a system governing several substrates or nutrients, we have focussed on the simpler system (1.5) in which the transition probability function is represented by a power law.



(a) W_ϵ versus x.



(b) P_ϵ versus x.

FIG. 2. *Plot of the leading order solutions for W_ϵ and P_ϵ at four different times when $p = 2$, $\gamma = 1$, and $\epsilon = 0.06$. The heavy solid curve is for $t = 0$, the solid curve is for $t = 4.01 \times 10^5$, the dashed curve is for $t = 1.0214 \times 10^7$, and the widely spaced dots are for $t = 1.0227 \times 10^7$.*

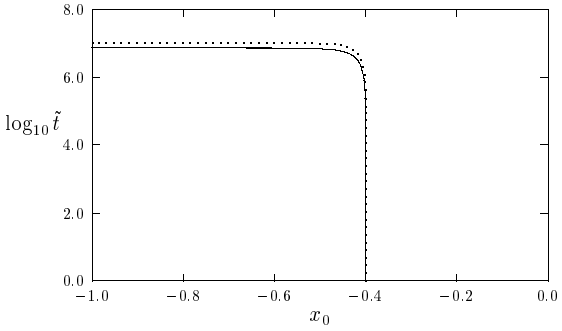


FIG. 3. *Comparison of numerical and asymptotic results for $\log_{10} \tilde{t}$, where $\tilde{t} = 1 + t$, versus x_0 when $p = 2$, $\gamma = 1$, $x_0^0 = -0.4$, and $\epsilon = 0.06$. The solid curve is the full numerical result computed from* (1.8). *The dashed curve is the asymptotic result computed from the ODE* (7.28b).

Under weak diffusion of the substrate, we have established the existence of spike solutions and investigated their stability. Of particular interest is the existence of metastable spikes. With regards to the modeling of the movement of myxobacteria our results suggest that bacteria which aggregate when diffusion of the substrate is neglected are, in fact, metastable spikes under weak diffusion. A similar observation can be made regarding the initiation of capillary sprouts in tumor angiogenesis under the simple model indicated in section 1. Furthermore, it is an important problem to investigate whether spike behavior can account for the observed vigorous angiogenesis in a vascularised tumor.

The investigations of this paper suggest a number of open problems. For example, it is of interest to investigate whether spikes arise for more a general transition probability function Φ. In addition, we may ask whether spike patterns exist for systems of the form (1.1) in which there are several substrates or nutrients.

Acknowledgments. We would like to thank the referees for their careful reading of the manuscript.

REFERENCES

[1] B. Alberts, D. Bray, J. Lewis, M. Raffs, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell*, 3rd ed., Garland Publishing, New York, 1994.

[2] P. Bates, E. N. Dancer, and J. Shi, *Multi-spike stationary solutions of the Cahn–Hilliard equation in higher-dimension and instability*, Adv. Differential Equations, 4 (1999), pp. 1–69.

[3] P. Bates and J. Shi, *Existence and instability of spike layer solutions to singular perturbation problems*, J. Funct. Anal., 196 (2002), pp. 211–264.

[4] M. A. Chaplain and A. R. A. Anderson, *Modeling the growth and form of capillary networks*, in On Growth and Form, M. A. Chaplain, G. D. Smith, and J. C. McLachlan, eds., Wiley, New York, 1999, pp. 225–249.

[5] X. Chen and M. Kowalczyk, *Slow dynamics of interior spikes in the shadow Gierer–Meinhardt system*, Adv. Differential Equations, 6 (2001), pp. 847–872.

[6] B. Davis, *Reinforced random walks*, Probability and Related Fields, 84 (1990), pp. 203–229.

[7] A. Doelman, R. A. Gardner, and T. J. Kaper, *Stability of singular patterns in the 1-D Gray–Scott model: A matched asymptotic approach*, Phys. D, 122 (1998), pp. 1–36.

[8] J. Folkman, *Tumor angiogenesis*, Adv. Cancer Res., 43 (1985), pp. 175–203.

[9] J. Folkman, *Angiogenesis in cancer, vascular, rheumatical and other diseases*, Nature Medicine, 1 (1995), pp. 21–31.

[10] A. Gierer and H. Meinhardt, *A theory of biological pattern formation*, Kybernetik, 12 (1972), pp. 30–39.

[11] C. Gui and J. Wei, *Multiple interior peak solutions for some singular perturbation problems*, J. Differential Equations, 158 (1999), pp. 1–27.

[12] T. Hillen and K. J. Painter, *A parabolic model with bounded chemotaxis—Prevention of overcrowding*, Adv. Appl. Math., 26 (2001), pp. 280–301.

[13] J. Holash, P. C. Maisonpierre, D. Compton, P. Boland, C. R. Alexander, D. Zagzag, G. D. Yancopoulos, and S. J. Wiegand, *Vessel cooption, regression and growth in tumors mediated by angioproteins and VEGF*, Science, 284 (1999), pp. 1994–1998.

[14] J. Holmes and B. D. Sleeman, *A mathematical model of tumor angiogenesis incorporating cellular traction and viscoelastic effects*, J. Theoret. Biol., 202 (2000), pp. 95–112.

[15] D. Horstmann, *From* 1970 *until present: The Keller-Segel model in chemotaxis and its consequences* I, Jahresber. Deutch. Math.-Verein, 105 (2003), pp. 103–165.

[16] D. Iron and M. J. Ward, *A metastable spike solution for a nonlocal reaction-diffusion model*, SIAM J. Appl. Math., 60 (2000), pp. 778–802.

[17] E. F. Keller and L. A. Segel, *Initiation of slime mold aggregation viewed as an instability*, J. Theoret. Biol., 26 (1970), pp. 399–415.

[18] M. K. Kwong and L. Zhang, *Uniqueness of positive solutions of* $\Delta u + f(u) = 0$ *in an annulus*, Differential Integral Equations, 4 (1991), pp. 583–599.

[19] H. A. Levine and B. D. Sleeman, *A system of reaction diffusion equations arising in the theory of reinforced random walks*, SIAM J. Appl. Math., 57 (1997), pp. 683–730.

[20] H. A. Levine, B. D. Sleeman, and M. Nilsen-Hamilton, *Mathematical modeling of the onset of capillary formation initiating angiogenesis*, J. Math. Biol., 42 (2001), pp. 195–238.
[21] H. A. Levine, B. D. Sleeman, and M. Nilsen-Hamilton, *A mathematical model for the roles of pericytes and macrophages in the onset of angiogenesis* I: *The role of protease inhibitors in preventing angiogenesis*, Math. Biosci., 168 (2000), pp. 77–115.
[22] H. A. Levine, S. Pamuk, B. D. Sleeman, and M. Nilsen-Hamilton, *Mathematical modeling of capillary formation and development in tumor angiogenesis: Penetration into the stroma*, Bull. Math. Biol., 63 (2001), pp. 801–863.
[23] C. S. Lin, W. M. Ni, and I. Takagi, *Large amplitude stationary solutions to a chemotaxis system*, J. Differential Equations, 72 (1988), pp. 1–27.
[24] J. D. Murray, *Mathematical Biology*, Springer, Berlin, Heidelberg, 1989.
[25] *NAG Fortran Library Mark* 17, *Routine* D03PCF, Numerical Algorithms Group, Oxford, U.K., 1995.
[26] W. Ni, *Diffusion, cross-diffusion, and their spike-layer steady-states*, Notices Amer. Math. Soc., 45 (1998), pp. 9–18.
[27] W. M. Ni and I. Takagi, *On the shape of least-energy solutions to a semilinear Neumann problem*, Comm. Pure Appl. Math., 44 (1991), pp. 819–851.
[28] W. M. Ni and I. Takagi, *Locating the peaks of least-energy solutions to a semilinear Neumann problem*, Duke Math. J., 70 (1993), pp. 247–281.
[29] H. G. Othmer and A. Stevens, *Aggregation, blowup, and collapse: The ABC's of taxis in reinforced random walks*, SIAM J. Appl. Math., 57 (1997), pp. 1044–1081.
[30] K. J. Painter and T. Hillen, *Volume-filling and quorum-sensing in models for chemosensitive movement*, Canadian Appl. Math. Quart., 10 (2003), pp. 501–544.
[31] M. J. Plank and B. D. Sleeman, *A reinforced random walk model of tumour angiogenesis and anti-angiogenic strategies*, Math. Med. Biol., 20 (2003), pp. 135–181.
[32] A. B. Potapov and T. Hillen, *Metastability in chemotaxis models*, J. Dynam. Differential Equations, (2005), to appear.
[33] L. Preziosi, *Cancer Modelling and Simulation*, Chapman and Hall/CRC, Bacon Raton, London, 2003.
[34] M. Rascle and C. Ziti, *Finite time blow-up in some models of chemotaxis growth*, J. Math. Biol., 33 (1995), pp. 388–414.
[35] B. D. Sleeman and I. P. Wallis, *Tumor induced angiogenesis as a reinforced random walk: Modeling capillary network formation without endothelial cell proliferation*, Math. Comput. Modeling, 36 (2002), pp. 339–358.
[36] C. L. Stokes and D. A. Lauffenburger, *Analysis of the roles of microvessel endothelial cell random motility and chemotaxis in angiogenesis*, J. Theoret. Biol., 152 (1991), pp. 377–403.
[37] M. J. Ward, *Eliminating indeterminacy in singularly perturbed boundary value problems with translation invariant potentials*, Stud. Appl. Math., 87 (1992), pp. 95–135.
[38] M. J. Ward, *An asymptotic analysis of localized solutions for some reaction-diffusion models in multi-dimensional domains*, Stud. Appl. Math., 97 (1996), pp. 103–126.
[39] J. Wei, *On the construction of single-peaked solutions to a singularly perturbed Neumann problem*, J. Differential Equations, 129 (1996), pp. 315–333.
[40] J. Wei, *On the boundary spike layer solutions of singularly perturbed semilinear Neumann problem*, J. Differential Equations, 134 (1997), pp. 104–133.
[41] J. Wei, *On single interior spike solutions for the Gierer–Meinhardt system: Uniqueness and stability estimates*, European J. Appl. Math., 10 (1999), pp. 353–378.
[42] J. Wei, *Uniqueness and critical spectrum of boundary spike solutions*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 1457–1480.
[43] J. Wei and M. Winter, *Multi-peak solutions for a wide class of singular perturbation problems*, J. London Math. Soc., 59 (1999), pp. 585–606.
[44] J. Wei and M. Winter, *A nonlocal eigenvalue problem and the stability of spikes for reaction-diffusion systems with fractional reaction rates*, Int. J. Bifur. Chaos Appl. Sci. Engrg., 13 (2003), pp. 1529–1543.
[45] J. Wei and L. Zhang, *On a nonlocal eigenvalue problem*, Ann. Scuola Norm. Sup. Pisa, 30 (2001), pp. 41–61.

SIAM J. APPL. MATH.
Vol. 65, No. 3, pp. 818–837

FAST INVERSION OF THE RADON TRANSFORM USING LOG-POLAR COORDINATES AND PARTIAL BACK-PROJECTIONS*

FREDRIK ANDERSSON†

Abstract. In this paper a novel filtered back-projection algorithm for inversion of a discretized Radon transform is presented. It makes use of invariance properties possessed by both the Radon transform and its dual. By switching to log-polar coordinates, both operators can be expressed in a displacement invariant manner. Explicit expressions for the corresponding transfer functions are calculated. Furthermore, by dividing the back-projection into several partial back-projections, inversion can be performed by means of finite convolutions and hence implemented by an FFT-algorithm. In this way, a fast and accurate reconstruction method is obtained.

Key words. Radon transform, filtered back-projection

AMS subject classifications. 44A12, 64R10, 92C55

DOI. 10.1137/S0036139903436005

1. Introduction.

Computerized tomography, CT, is a well established image acquisition technique with a growing range and diversity of applications. The most common algorithm used in X-ray CT, SPECT, and many other image modalities is called filtered back-projection. It consists of two steps: a one-dimensional filtering of the data followed by a back-projection step, where the filtered data values are distributed along the lines of measurement. In standard implementations of filtered back-projection algorithms, the latter part is by far the most time-consuming. Doubling the image resolution requires four times the amount of measurement data, while the computational cost in the back-projection step increases by a factor of eight. For the reconstruction of an image with resolution $q \times q$, the number of computations needed in the back-projection part is $\mathcal{O}(q^3)$.

The principle of CT is mathematically described by the *Radon transform*. The two-dimensional Radon transform is the mapping of a function in $\mathbb{R}^2$ to its line integral values, and the reconstruction problem of CT lies in inverting this mapping. The back-projection mentioned above is mathematically described by the dual Radon transform which integrates over all lines passing through a point.

Several suggestions on how to invert the Radon transform in $\mathcal{O}(q^2 \log q)$ time have appeared in the literature during the years. Most common are Fourier-based methods. Using such an approach results in data of the two-dimensional Fourier transform of the sought function on a nonuniform grid. Direct interpolation to a rectangular grid followed by FFT inversion results in $\mathcal{O}(q^2 \log q)$ complexity but gives rise to unacceptable artifacts. To cope with this one can, e.g., use over-sampling combined with more sophisticated interpolation, as suggested in [14], or use fast Fourier algorithms for unequally spaced data; cf. [2], [16], [8].

For issues of quality, the algorithms of filtered back-projection type have traditionally been preferred among the manufacturers of CT machines. Fast techniques for filtered back-projection algorithms are presented in [15], [6], [3], [4], where the back-projection is calculated recursively in $\mathcal{O}(q^2 \log q)$ time.

*Received by the editors October 6, 2003; accepted for publication (in revised form) June 19, 2004; published electronically February 25, 2005.
http://www.siam.org/journals/siap/65-3/43600.html
†Center for Mathematical Sciences, Lund University/LTH, P.O. Box 118, S-22100 Lund, Sweden (fa@maths.lth.se).

In this paper we exploit the fact that both the Radon transform and its dual possess similar invariance properties. By a change of coordinates to log-polar coordinates, the operators can be expressed as convolutions. In particular, by introducing an appropriate Fourier transform, it is possible to compute parts of the back-projection in a fast manner by means of the FFT. In this way a filtered back-projection algorithm is constructed which works in $\mathcal{O}(q^2 \log q)$ time.

A log-polar grid is hence introduced for treating the back-projection part in a fast manner. Log-polar grids have previously been used in an ART fashion; cf. [7]. However, according to [7], this approach was not successful in competition with the standard filtered back-projection algorithm, in either speed or quality.

The methods developed in this paper may also be used in fast calculation of the (forward) Radon transform, although the focus of the paper lies on the inverse problem. Fast computation of the forward problem has also appeared in, e.g., [9], [3]. Besides simulation possibilities, this is, e.g., useful in fast computation of the Hough transform, a tool used in image analysis and pattern recognition for finding lines.

The paper begins with a brief review of the Radon transform and its inversion. In section 3.1 follows a discussion on the invariance properties possessed by the Radon transform and its dual. We show that when expressed by log-polar coordinates, both can be written in terms of convolutions. Explicit expressions for the corresponding kernels are given in section 3.2. However, the geometry of log-polar grids makes it impractical to use the convolutional structure directly. In section 3.3 the concept of partial back-projections is introduced to deal with this problem. The discrete setting is dealt with in section 4, where implementation techniques are discussed and numerical experiments displayed.

2. Preliminaries. The two-dimensional Radon transform is the mapping from (sufficiently regular) functions on $\mathbb{R}^2$ to line integrals in $\mathbb{R}^2$,

$$\mathcal{R}f(\theta, s) = \int_{x\cdot\theta=s} f(x)\, dx, \tag{2.1}$$

i.e., the integral of f over the line with normal direction θ and (signed) distance s to the origin. When defined on the unit cylinder $\mathbb{S} = S^1 \times \mathbb{R}$, the two-dimensional Radon transform (defined by (2.1)) is even, i.e., $\mathcal{R}f(-\theta, -s) = \mathcal{R}f(\theta, s)$, since each line can be parameterized in two ways. Here S^1 denotes the unit circle, parameterized by $[-\pi, \pi)$. When describing an element of S^1, it is sometimes advantageous to describe it via a unit vector and sometimes with the corresponding angle in a polar representation. To avoid cumbersome notation, it is customary to use the same symbol in both cases, and we will follow this tradition. Sometimes it is convenient to work on $\mathbb{S}_+ = S^1 \times \mathbb{R}_+$ or $\mathbb{S}_{1/2} = [-\pi/2, \pi/2) \times \mathbb{R}$ instead of $\mathbb{S}$, both able to describe all lines in the plane.

It is well known that the Radon transform as a mapping $\mathcal{S}(\mathbb{R}^2) \to \mathcal{S}(\mathbb{S})$ is injective, where $\mathcal{S}$ denotes the Schwartz space. In this paper we are especially interested in the Radon transform of compactly supported functions, and inversion techniques for such. The Radon transform can then (possibly combined with a suitable preceding translation) be viewed as an injective mapping $\mathcal{C}_0^\infty(\mathbb{S}_+) \to \mathcal{C}_0^\infty(\mathbb{S}_+)$.

The dual Radon transform $\mathcal{R}^\#$ integrates functions defined on $\mathbb{S}$ over subsets of $\mathbb{S}$ corresponding to lines passing through a point $x \in \mathbb{R}^2$,

$$\mathcal{R}^\# g(x) = \int_{S^1} g(\theta, x\cdot\theta)\, d\theta. \tag{2.2}$$

It is dual in the sense that

$$\int_{\mathbb{S}} (\mathcal{R}f)(\theta,s)g(\theta,s)\; ds d\theta = \int_{\mathbb{R}^2} f(x)(\mathcal{R}^{\#}g)(x)\; dx,$$

and in the literature is commonly referred to as the *back-projection operator*.

In order to write the inversion formula for the Radon transform in a simple form, we introduce the operator $\mathcal{J}$ acting on (sufficiently regular) functions on $\mathbb{S}$, defined by

$$\mathcal{J} = \frac{\partial}{\partial s}\mathcal{H},$$

where $\mathcal{H}$ is the Hilbert transform operator, applied to the second variable in $\mathbb{S}$. Here, the Hilbert transform is defined by

$$\mathcal{H}f(x) = \frac{1}{\pi}\int_{\mathbb{R}} \frac{1}{x-y} f(y)\, dy,$$

where the integral is interpreted as a principal value.

An inversion formula then reads as

$$f = \frac{1}{4\pi}\mathcal{R}^{\#}\mathcal{J}\mathcal{R}f. \tag{2.3}$$

For more details, see [14]. A factor of $1/2$ on the right-hand side is due to the fact that, in the interpretation of $\mathcal{R}^{\#}$, each line is taken into account twice. Using instead $\mathbb{S}_+$ or $\mathbb{S}_{1/2}$, the factor $1/4\pi$ is replaced by $1/2\pi$. The operator $\mathcal{J}$ cannot be defined directly on $\mathbb{S}_+$, and hence instead it is defined as the restriction of the action on $\mathbb{S}$ to $\mathbb{S}_+$.

In practice, where Radon data are given on a discrete sampling grid, the operator $\mathcal{J}$ is usually implemented by a discrete convolution in the s-variable with a band-limited filter. More specifically, one makes use of the formula

$$W * f = \mathcal{R}^{\#}(w \overset{s}{*} \mathcal{R}f), \tag{2.4}$$

where $\overset{s}{*}$ denotes one-dimensional convolution with respect to the second variable, s, in $\mathbb{S}$, and where W and w are related by $W = \mathcal{R}^{\#}w$; cf. [14]. By choosing W to approximate a δ-distribution, an approximate reconstruction is obtained. More specifically, one usually chooses $W = W_b$ to be radially symmetric and band-limited with some cut-off frequency b. When chosen in this way, $w = w_b$ depends only on the s-variable, in which it is band-limited by b; cf. [14].

Let us consider the problem of making reconstructions on a rectangular grid of size $q \times q$. To do that, the number of needed parallel lines and directions in data measurements are both of order q; cf. [14]. By approximating the continuous convolution in (2.4) by a discrete one, the filtering step $w \overset{s}{*} \mathcal{R}f$ can be accomplished, by use of FFT, with a time complexity $\mathcal{O}(q^2 \log q)$. A more time-consuming step is the computation of the back-projection. The straightforward numerical implementation of $\mathcal{R}^{\#}$ uses, for each discrete direction sample point θ_j, some kind of interpolation in the s-variable to approximate $g(\theta_j, x\cdot\theta_j)$ in (2.2), in combination with some quadrature rule on S^1:

$$\mathcal{R}_d^{\#}g(x) = \sum_j \alpha_j g(\theta_j, x\cdot\theta_j).$$

Hence, roughly q values are summed up at each reconstruction point, giving a time complexity of order $\mathcal{O}(q^3)$ for q^2 reconstruction points. Reconstruction methods which are based on (2.3) are referred to as *filtered back-projection algorithms*.

Next, we present a discussion about some properties of the continuous Radon transformation and its dual, which are both of interest on their own and useful for the discrete approximative inversion presented below. A somewhat reminiscent discussion is given in [1], where some closed-form formulas, involving Chebyshev polynomials, are given starting from a polar representation. A numerical implementation is presented in [1], but the number of computations needed is of order $\mathcal{O}(q^3)$.

3. Properties of the continuous Radon transform and its dual.

3.1. Convolution operators. Consider the cylinder $\mathbb{S}_+ = S^1 \times \mathbb{R}_+$ as a group, provided with the algebraic structure inherited from its components, i.e., the additive group $S^1 = \mathbb{R}/2\pi\mathbb{Z}$ (addition modulo 2π) and the multiplicative group $\mathbb{R}_+$ (positive real numbers).

Let $z = (\theta, s) \in \mathbb{S}_+$. The group operation on $\mathbb{S}_+$, written multiplicatively, is

$$(\theta_1, s_1)(\theta_2, s_2) = ((\theta_1 + \theta_2) \bmod 2\pi, s_1 s_2).$$

The Haar measure on $\mathbb{S}_+ = S^1 \times \mathbb{R}_+$ is inherited from the components and can be written $dh(z) = d\theta\, ds/s$. Hence

$$\int_{\mathbb{S}_+} f(z)\, dh(z) = \int_0^{2\pi} \int_0^{\infty} f(\theta, s)\, d\theta\, \frac{ds}{s} \quad \text{for} \quad f \in \dot{\mathcal{C}}_0^{\infty}(\mathbb{S}_+),$$

where $\dot{\mathcal{C}}_0^{\infty}(\mathbb{S}_+)$ is the $\mathcal{C}_0^{\infty}$ class on $\mathbb{S}_+$, with support outside the origin $S^1 \times \{0\}$. The Haar property means that

$$\int_{\mathbb{S}_+} f(wz)\, dh(z) = \int_{\mathbb{S}_+} f(z)\, dh(z) \quad \text{for} \quad w \in \mathbb{S}_+,\ f \in \dot{\mathcal{C}}_0^{\infty}(\mathbb{S}_+).$$

There exists a natural isomorphism between $\mathbb{S}_+$ and the punctured complex plane $\dot{\mathbb{C}} = \mathbb{C} \setminus \{0\}$ considered multiplicatively, parameterized by $(\theta, s) \longleftrightarrow s e^{i\theta}$. Using the Cartesian representation $z = x + iy$ for $\dot{\mathbb{C}}$, the Haar measure on $\mathbb{S}_+$ can be written

$$dh(z) = d\theta\, \frac{ds}{s} = \frac{dx\, dy}{x^2 + y^2}.$$

If $\dot{\mathbb{C}}$ is represented instead by coordinates

$$s e^{i\theta} = e^{\rho} e^{i\theta}, \quad \rho \in \mathbb{R}, \tag{3.1}$$

which we will refer to as *log-polar coordinates*, then the Haar measure on $\mathbb{S}_+$ becomes

$$dh(z) = d\theta\, d\rho.$$

Let λ be the distribution that represents integration over the line $x = 1$ in $\dot{\mathbb{C}}$,

$$\lambda : f \longmapsto \int_{-\infty}^{\infty} f(1, y)\, dy = \int_{\dot{\mathbb{C}}} f(x, y) \delta(x - 1)\, dx\, dy, \quad f \in \mathcal{S}(\mathbb{S}_+).$$

In the $(\theta, s) = z$ representation, it can be written

$$\lambda : f \longmapsto \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} f\left(\theta, \frac{1}{\cos\theta}\right) \frac{1}{\cos^2\theta}\, d\theta = \int_{\mathbb{S}_+} f(\theta, s)\delta(s\cos\theta - 1) s\, ds\, d\theta$$
$$= \int_{\mathbb{S}_+} f(z)\lambda(z)\, dh(z), \quad \text{with} \quad \lambda(z) = s^2\delta(s\cos\theta - 1).$$

Let lines in $\dot{\mathbb{C}}$ be denoted L_ξ, where ξ is the footprint of the normal of the line through the origin. The Radon transform (2.1) can be expressed as

$$g(\xi) = Rf(\xi) = |\xi| \int_{\mathbb{S}_+} f(z\xi)\lambda(z)\, dh(z). \tag{3.2}$$

This follows from the fact that, in the last integral, the distribution λ is applied to a function that is obtained from f by a similarity transformation of the coordinate plane, such that the line L_ξ is transferred to L_1, which is the support of λ. Using $|\xi|$ to compensate for the change in scale, we obtain the Radon transform.

The duality between points on a line in $\dot{\mathbb{C}}$ and lines through a point in $\dot{\mathbb{C}}$ motivates interest in the distribution

$$\lambda^\# : f \longmapsto \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} f(\theta, \cos\theta)\, d\theta = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_{\mathbb{R}_+} f(\theta, s)\delta(\cos\theta - s)\, ds\, d\theta$$
$$= \int_{\mathbb{S}_+} f(z)\lambda^\#(z)\, dh(z), \quad \text{with} \quad \lambda^\#(z) = s\delta(\cos\theta - s).$$

If f is interpreted as a function on lines in $\dot{\mathbb{C}}$, then this represents the integral of f over all lines through the point $1 \in \dot{\mathbb{C}}$. Also note that

$$\int_\alpha^\beta \int_{\mathbb{R}_+} f(s, \theta)\lambda^\#(s, \theta)\, ds\, d\theta \tag{3.3}$$

is the integral over all lines through the point $1 \in \dot{\mathbb{C}}$ with normal direction in the interval $[\alpha, \beta]$.

The distributions $\zeta(z) = \lambda(1/z)$ and $\zeta^\#(z) = \lambda^\#(1/z)$, formally defined by

$$\zeta : f \longmapsto \int_{\mathbb{S}_+} f\left(\frac{1}{z}\right) \lambda(z)\, dh(z),$$
$$\zeta^\# : f \longmapsto \int_{\mathbb{S}_+} f\left(\frac{1}{z}\right) \lambda^\#(z)\, dh(z), \quad f \in \dot{\mathcal{C}}_0^\infty(\mathbb{S}_+),$$

will be crucial in what follows. In the $(\theta, s) = z$ representation, these can explicitly be written

$$\zeta(z) = s^{-2}\delta(s^{-1}\cos\theta - 1),$$
$$\zeta^\#(z) = \delta(s\cos\theta - 1).$$

Using the Haar property on the integral, the formula (3.2) can also be written

$$\mathcal{Z}f(\xi) = \frac{\mathcal{R}f(\xi)}{|\xi|} = \int f(z)\zeta\left(\frac{\xi}{z}\right)\, dh(z). \tag{3.4}$$

Apart from the scaling factor $|\xi|$, the Radon transform can thus be expressed as a convolution on $\mathbb{S}_+$.

For the back-projecting operator (2.2), note that if $\xi \in \dot{\mathbb{C}}$ and if $g \in \dot{\mathcal{C}}_0^\infty(\mathbb{S}_+)$ is interpreted as a function on lines in $\dot{\mathbb{C}}$, then

$$\mathcal{R}^\# g(\xi) = \int_{\mathbb{S}_+} g(\xi z)\lambda^\#(z)\, dh(z)$$

is the integral of g over all lines that pass through ξ. Hence, again by using the Haar property,

$$\mathcal{R}^\# g(\xi) = \int_{\mathbb{S}_+} g(z)\zeta^\#\left(\frac{\xi}{z}\right) dh(z), \tag{3.5}$$

which in the log-polar representation can be expressed by the following convolution:

$$\mathcal{R}^\# g(\rho,\theta) = \int_{\mathbb{S}_+} g(\rho',\theta')\zeta^\#(\rho-\rho',\theta-\theta')\, d\rho'\, d\theta'. \tag{3.6}$$

This fact opens the possibility of performing the inversion in (2.3) by means of an appropriate Fourier transform.

Note also that by using (3.3) and the same arguments as above, it follows that the back-projection restricted to lines with normal directions in the interval $[\alpha,\beta]$ can be written

$$\mathcal{R}^\#_{[\alpha,\beta]} g(\rho,\theta) = \int_\alpha^\beta \int_{\mathbb{R}_+} g(\rho',\theta')\zeta^\#(\rho-\rho',\theta-\theta')\, d\rho'\, d\theta'. \tag{3.7}$$

3.2. Fourier analysis. The Fourier transform $\mathcal{F}_{\mathbb{S}_+} = \mathcal{F}$ on $\mathbb{S}_+$ is a compound of the Fourier series transform on S^1 and the Mellin transform on $\mathbb{R}_+$:

$$\mathcal{F} : f(\theta,s) \longmapsto g(\mu,\sigma) = \int_0^{2\pi}\int_0^\infty e^{-i\mu\theta} s^{-\sigma} f(\theta,r)\, \frac{ds}{s}\, d\theta, \quad \mu \in \mathbb{Z},\ \sigma \in \mathbb{C}.$$

In a log-polar representation, (3.1), the Fourier transform instead becomes a compound of the Fourier series transform and the Laplace transform on $\mathbb{R}$,

$$\mathcal{F} : f(\theta,\rho) \longmapsto g(\mu,\sigma) = \int_0^{2\pi}\int_{-\infty}^\infty e^{-i\mu\theta} e^{-\sigma\rho} f(\theta,r)\, d\rho\, d\theta, \quad \mu \in \mathbb{Z},\ \sigma \in \mathbb{C}.$$

As the operators $\mathcal{Z}$ in (3.4) and $\mathcal{R}^\#$ in (3.5) can be expressed as convolutions on $\mathbb{S}_+$, it suffices to calculate the corresponding transfer functions for determination of $\mathcal{F}\zeta$ and $\mathcal{F}\zeta^\#$.

It is readily verified using (3.2) that, for $\xi = re^{i\psi}$,

$$\mathcal{Z}f(\xi) = \int_{\mathbb{S}_+} f(\theta+\psi, rs)\delta(s\cos\theta - 1)s\, ds\, d\theta = \int_{-\pi/2}^{\pi/2} f\left(\theta+\psi, \frac{r}{\cos\theta}\right)\frac{1}{\cos^2\theta}\, d\theta.$$

Let $f(\theta,s) = s^\sigma e^{i\mu\theta}$. Then the transfer function for $\mathcal{Z}$ is obtained by

$$\begin{aligned}\mathcal{Z}f(\xi) &= \int_{-\pi/2}^{\pi/2} e^{i\mu(\theta+\psi)}\left(\frac{r}{\cos\theta}\right)^\sigma \frac{1}{\cos^2\theta}\, d\theta = r^\sigma e^{i\mu\psi}\int_{-\pi/2}^{\pi/2} e^{i\mu\theta}\big(\cos\theta\big)^{-\sigma-2}\, d\theta \\ &= f(\xi)\mathcal{F}\zeta(\mu,\sigma).\end{aligned}$$

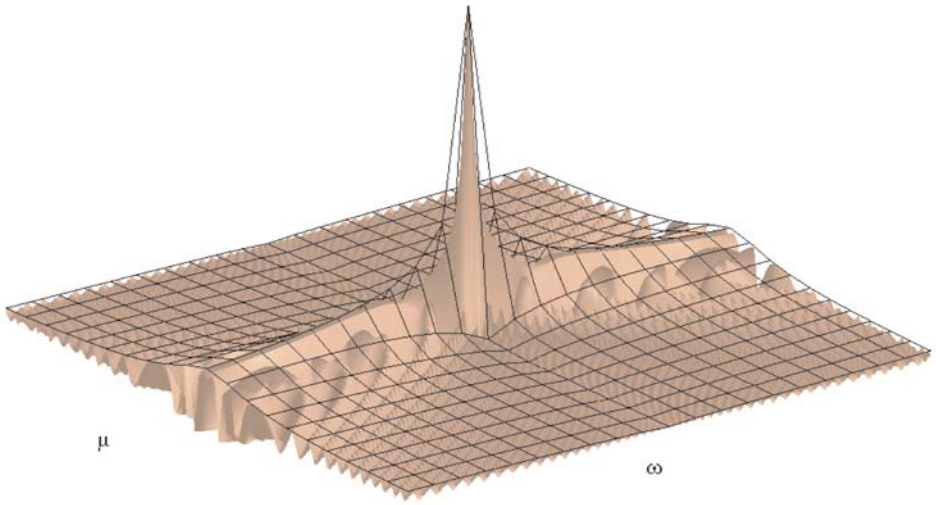


FIG. 3.1. *The absolute value and real part of* $\mathcal{F}\zeta^{\#}(\mu, i\omega)$.

Here, $\mathcal{F}\zeta(\mu,\sigma)$ converges in the classical sense for $\operatorname{Re}\sigma < -1$, where it defines an analytic function.

Similarly, the transfer function for the back-projecting operator $\mathcal{R}^{\#}$ is obtained by

$$\begin{aligned}\mathcal{R}^{\#}f(\xi) &= \int_{-\pi/2}^{\pi/2} e^{i\mu\theta}(r\cos(\theta-\psi))^{\sigma}\,d\theta = \int_{-\pi/2}^{\pi/2} e^{i\mu(\theta+\psi)}(r\cos\theta)^{\sigma}\,d\theta \\ &= r^{\sigma}e^{i\mu\psi}\int_{-\pi/2}^{\pi/2} e^{i\mu\theta}(\cos\theta)^{\sigma}\,d\theta = f(\xi)\mathcal{F}\zeta^{\#}(\mu,\sigma),\end{aligned}$$

where $\mathcal{F}\zeta^{\#}(\mu,\sigma)$ converges in the classical sense for $\operatorname{Re}\sigma > -1$.

Note that $\mathcal{F}\zeta(\mu,\sigma)$ and $\mathcal{F}\zeta^{\#}(\mu,\sigma)$ are not simultaneously (classically) well defined, and note further the resemblance between the two Fourier transforms:

$$\mathcal{F}\zeta(\mu,\sigma) = \mathcal{F}\zeta^{\#}(\mu,-\sigma-2).$$

An analogue to the following formula can be found in [10, p. 372].

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} e^{i\mu\theta}\cos^{\sigma}(\theta)\,d\theta = \frac{\pi\Gamma(\sigma+1)}{2^{\sigma}\Gamma(\frac{\sigma+\mu}{2}+1)\Gamma(\frac{\sigma-\mu}{2}+1)}, \qquad \operatorname{Re}\sigma > -1.$$

Hence we conclude that

$$\begin{aligned}\mathcal{F}\zeta(\mu,\sigma) &= \frac{\pi\Gamma(-\sigma-1)}{2^{-\sigma-2}\Gamma(\frac{-\sigma+\mu}{2})\Gamma(\frac{-\sigma-\mu}{2})}, \quad \operatorname{Re}\sigma < -1, \\ \mathcal{F}\zeta^{\#}(\mu,\sigma) &= \frac{\pi\Gamma(\sigma+1)}{2^{\sigma}\Gamma(\frac{\sigma+\mu}{2}+1)\Gamma(\frac{\sigma-\mu}{2}+1)}, \quad \operatorname{Re}\sigma > -1. \end{aligned} \tag{3.8}$$

Figure 3.1 illustrates the function $\mathcal{F}\zeta^{\#}(\mu, iw)$. The real part is displayed in the form of a surface plot, and the absolute value is shown on a coarser grid in a mesh

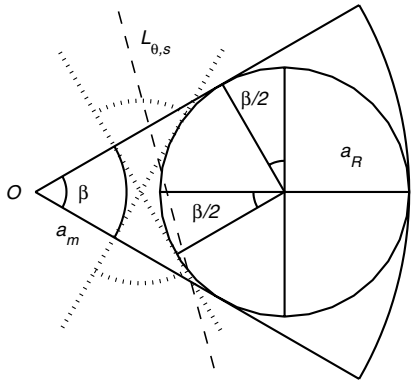


FIG. 3.2. *Lines with directions* $\theta \in [-\frac{\beta}{2}, \frac{\beta}{2}]$, *passing through a circle inscribed in a segment.*

plot. Its energy content is mainly contained in a neighborhood of the origin and along the line $\mu = 0$. Note that, although its absolute value is rather smooth, the real and imaginary parts are strongly oscillating.

3.3. Partial back-projection. Our aim is to present a fast procedure for computing the back-projection as a discrete convolution on a uniformly sampled grid in log-polar coordinates. However, the counterpart in polar coordinates will be nonuniform, corresponding to a dense assembling of grid points close to the origin in $\mathbb{C}$. This assembling causes large variation in the density of both reconstruction grid points and data sample points. By moving the origin it is possible to obtain a more uniform structure within the grid points of the reconstruction region, but this does not simplify the treatment of the data sample points. The difficulty here is the need to deal with lines with various distances to the origin, and this problem exists no matter where the origin is situated, if lines with all directions must be considered.

To cope with this, we divide the measurement data into $m \geq 3$ (disjoint) sets, each containing lines with directions in an interval of length $\frac{\pi}{m}$. Corresponding to each such set, we choose an origin outside the region of interest, in such a way that the back-projection from lines with directions within each interval (cf. (3.3)) can be calculated by means of finite integrals. We refer to this procedure as *partial back-projection*. By putting these partial back-projections together, it is possible to obtain a total back-projection. To begin with, we consider one special case.

Let $\beta = \frac{\pi}{m}$ for some positive integer $m \geq 3$, and let a_R denote the radius of the largest circle inscribed in a sector with unit radius and central angle β; cf. Figure 3.2. It is straightforward to show that

$$a_R = \frac{\sin(\frac{\beta}{2})}{1 + \sin(\frac{\beta}{2})}. \tag{3.9}$$

Consider the set of all lines $L_{\theta,s}$, with normal directions $|\theta| \leq \frac{\beta}{2}$ with respect to the symmetry axis of the sector, passing through the inscribed circle; cf. Figure 3.2. It is clear that the normal distances to the origin, s, of these lines will be in the interval $(a_m, 1)$, where

$$a_m + a_R = (1 - a_R) \cos\left(\frac{\beta}{2}\right), \tag{3.10}$$

or, by using (3.9),

$$a_m = \frac{\cos(\frac{\beta}{2}) - \sin(\frac{\beta}{2})}{1 + \sin(\frac{\beta}{2})}. \tag{3.11}$$

Suppose that $f \in \mathcal{C}_0^\infty$ has its support inside the inscribed circle, and that $\mathcal{R}f(\theta, s)$ is known. Let

$$h^0(\theta, s) = \chi_\beta \mathcal{J}\mathcal{R}f(\theta, s),$$

where χ_β is the characteristic function of the set

$$\left\{\theta : -\frac{\beta}{2} \le \theta < \frac{\beta}{2}\right\}.$$

The contribution

$$f_0(x) = \int_{-\frac{\beta}{2}}^{\frac{\beta}{2}} \mathcal{J}\mathcal{R}f(\theta, \theta \cdot x)\, d\theta$$

from the back-projection of lines with directions in the interval $(-\frac{\beta}{2}, \frac{\beta}{2})$ in log-polar coordinates, for x inside the inscribed circle, may be written as (by using (3.7))

$$f_0(\theta, \rho) = \int_{-\frac{\beta}{2}}^{\frac{\beta}{2}} \int_{\ln(a_m)}^{0} h^0(\theta', \rho')\zeta^{\#}(\theta - \theta', \rho - \rho')\, d\rho' d\theta'.$$

Furthermore, if $\zeta_p^{\#}(\theta, \rho)$ is defined as the periodic extension of $\zeta^{\#}(\theta, \rho)$ for $\theta \in [\beta, \beta]$ and $\rho \in [\ln(a_m), 0]$, then

$$f_0(\theta, \rho) = \int_{-\beta}^{\beta} \int_{\ln(a_m)}^{0} h^0(\theta', \rho')\zeta_p^{\#}(\theta - \theta', \rho - \rho')\, d\rho' d\theta', \tag{3.12}$$

when $\theta \in [-\frac{\beta}{2}, \frac{\beta}{2}]$ and $\rho \in [1 - 2a_R, 1]$. This is due to the fact that the interval length in the θ'-direction is twice the θ'-support of $h^0(\theta', \rho')$, and to the fact that the values of $h^0(\theta', \rho')$ for ρ' outside $[\ln(a_m), 0]$ are of no importance in (partial) back-projecting for $x = (\theta, \rho)$ inside the inscribed circle. Thus, replacing $h^0(\theta', \rho')$ with a periodical extension of $h^0(\theta', \rho')$ as above (which is equivalent to extending $\zeta^{\#}$) does not influence the result. Hence, the partial back-projection at points of interest can be calculated as a periodic convolution.

Now, let $f \in \mathcal{C}_0^\infty(\Omega^2)$, where Ω^2 denotes the unit disc in $\mathbb{R}^2$, and suppose that $\mathcal{R}f(\theta, s)$ is known. By dividing the data into m different parts, each spanning an angle interval of length β, and making a suitable change of coordinates to each such set, one transforms the full back-projection problem into m subproblems of the form above. An illustration of the procedure, for $m = 3$, is given in Figure 3.3. The coordinate transformation $x \to x_\nu$ consists of, after a rescaling with a_R, a rotation by the angle β followed by a translation of the origin O to O_ν. Each part is then of the form discussed above. Adding the respective partial back-projections gives the total back-projection, since each partial back-projection integrates over disjunct intervals $(-\beta/2 + \nu\beta, \beta/2 + \nu\beta)$, which together covers an interval length of $m\beta = \pi$. For the sake of completeness, we include the details.

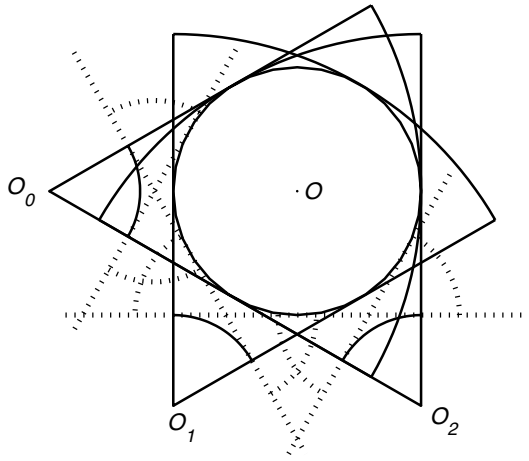


FIG. 3.3. *Reconstruction from three views.*

Let $\nu = 0, \ldots, m-1$; introduce for each ν new coordinates

$$x_\nu = a_R \begin{pmatrix} \cos(\nu\beta) & \sin(\nu\beta) \\ -\sin(\nu\beta) & \cos(\nu\beta) \end{pmatrix} x + \begin{pmatrix} 1 - a_R \\ 0 \end{pmatrix};$$

and let $f^\nu(x_\nu) = f(x)$. Each f^ν then has its support inside a circle inscribed in a sector with unit radius and central angle β, as in Figure 3.2. Let $h^\nu = \chi_\beta \mathcal{J}\mathcal{R} f^\nu$. The following relation, easily verified,

$$x_\nu \cdot \theta = x \cdot a_R(\theta + \nu\beta) + (1 - a_R)\cos(\theta), \tag{3.13}$$

will be useful. Any line in the x-coordinate system may be written

$$L_{\theta+\nu\beta,s} = \{x | x \cdot (\theta + \nu\beta) = s\},$$

for some $\nu \in \{0, \ldots, m-1\}$ and $-\frac{\beta}{2} \le \theta < \frac{\beta}{2}$. In the x_ν-coordinate system the corresponding line may then, by using (3.13), be written

$$L^\nu_{\theta, a_R s + (1-a_R)\cos(\theta)} = \{x_\nu | x_\nu \cdot \theta = a_R s + (1 - a_R)\cos(\theta)\},$$

and hence,

$$h^\nu(\theta, s) = \chi_\beta \mathcal{J}\mathcal{R} f \left(\theta + \nu\beta, \frac{s - (1 - a_R)\cos(\theta)}{a_R} \right). \tag{3.14}$$

The total back-projection for some point x inside the support of f can now, again by using (3.13), be written as

$$\begin{aligned} \mathcal{R}^\# \mathcal{J}\mathcal{R} f(x) &= \int_{S^1} \mathcal{J}\mathcal{R} f(\theta, x \cdot \theta)\, d\theta \\ &= 2 \sum_{\nu=0}^{m-1} \int_{-\frac{\beta}{2}}^{\frac{\beta}{2}} \mathcal{J}\mathcal{R} f(\theta + \nu\beta, (\theta + \nu\beta) \cdot x)\, d\theta \\ &= 2 \sum_{\nu=0}^{m-1} \int_{-\frac{\beta}{2}}^{\frac{\beta}{2}} h^\nu(\theta, a_R(\theta + \nu\beta) \cdot x + (1 - a_R)\cos(\theta))\, d\theta \\ &= 2 \sum_{\nu=0}^{m-1} \int_{-\frac{\beta}{2}}^{\frac{\beta}{2}} h^\nu(\theta, x_\nu \cdot \theta)\, d\theta, \end{aligned}$$

where the factor 2 in the last expression is due to the fact that the total integration is performed on only half of S^1.

We now have reduced the full back-projection problem to m partial reconstructions of the form discussed above. For future reference, let us introduce

$$f_r^\nu(x_\nu) = \frac{1}{2\pi}\int_{-\frac{\beta}{2}}^{\frac{\beta}{2}} h^\nu(\theta, x_\nu\cdot\theta)\,d\theta. \tag{3.15}$$

The inversion formula (2.3) then allows us to write

$$f(x) = \sum_{\nu=0}^{m-1} f_r^\nu(x_\nu). \tag{3.16}$$

We have deduced that it is possible to compute the total back-projection as a sum of partial back-projection. The sharp cutoff caused by χ_β will, however, in practice give rise to artifacts in the form of sharp lines with normal directions corresponding to the cutoff angles. Therefore it is desirable to make a smoother cutoff. This requires that the supports in the θ-direction for the respective h^ν overlap. The angle β is then not equal to $\frac{\pi}{m}$ (where m is the number of partial back-projections to be used) but larger. Suppose $\frac{\pi}{m} \le \beta \le \frac{2\pi}{m}$. By letting

$$\eta(t) = \begin{cases} e^{\frac{-\beta^2}{\beta^2-4t^2}} & \text{if } |t| < \frac{\beta}{2}, \\ 0 & \text{otherwise,} \end{cases}$$

a smooth cutoff function

$$\chi_{\beta,m}(\theta) = \frac{\eta(\theta)}{\eta(\theta)+\eta(\frac{\pi}{m}-\arccos(\cos\theta))}, \tag{3.17}$$

with support in $[-\frac{\beta}{2}, \frac{\beta}{2}]$, can be formed. Since

$$\sum_{k=-\infty}^{\infty} \chi_{\beta,m}\left(\theta + k\frac{\pi}{m}\right) \equiv 1,$$

it follows that the reconstruction (3.16) is still valid if χ_β in (3.14) is replaced by $\chi_{\beta,m}$.

4. Discrete log-polar reconstruction.

4.1. Principles. In this section we describe how to use the method of section 3.3 to make reconstructions from discrete measurements. For simplicity, in this paper we deal with the case of parallel beam data. Let $f \in \mathcal{C}_0^\infty(\Omega^2)$; let $g = \mathcal{R}f$ be sampled at (θ_j, s_l), $j = 0,\dots,p-1$, $l = 0,\dots,q-1$ (parallel beam geometry), where $\theta_j \in S^1$ and $s_l = \frac{2l}{q} - 1$; and let $h = w_b \overset{s}{*} g$, in accordance with (2.4). The latter quantity is approximated by the discrete convolution

$$w_b \overset{d}{*} g(\theta_j, s) = \frac{1}{q}\sum_{l=-q}^{q} w_b(s-s_l)g(\theta_j, s_l). \tag{4.1}$$

What now remains for reconstruction is the back-projection. To this end, let the integer m be the number of partial reconstructions to be used, let $\beta = \pi/m_\beta$, suppose that m and m_β are divisors of p, and let $\theta_j = -\frac{\beta}{2} + \frac{j\pi}{p}$.

For the construction of the discrete back-projecting operator we will make use of the interpolation results discussed in section A. To begin with, suppose $h(\theta, s)$ is known, and make a uniform sampling of h^ν defined in (3.14), with χ_β replaced by $\chi_{\beta,m}$ from (3.17), in a log-polar representation (θ_j, ρ_i) covering $(-\frac{\beta}{2}, \frac{\beta}{2}) \times (-\ln(a_m), 0)$. Let $I_{\Delta_2} = I_{(\Delta\theta,\Delta\rho)}$ be a quasi-interpolator of order o_2, with kernel $\varphi(\theta, \rho) = \varphi_\theta(\theta)\varphi_\rho(\rho)$, where $\Delta\theta$ and $\Delta\rho$ is the grid spacing of (θ_j, ρ_i).

Construct

$$h_*^\nu(\theta, \rho) = \sum_j \sum_i h^\nu[j, i]\varphi\left(\frac{\theta - \theta_j}{\Delta\theta}, \frac{\rho - \rho_i}{\Delta\rho}\right),$$

as an approximation of h^ν. Now the back-projection of h_*^ν, for x_ν represented by $\theta \in [-\frac{\beta}{2}, \frac{\beta}{2}]$ and $\rho \in [1 - 2a_R, 1]$, can be written

$$\begin{aligned}
\mathcal{R}^\# h_*^\nu(\theta, \rho) &= \int_{-\beta}^{\beta} \int_{\ln(a_m)}^{0} h_*^\nu(\theta - \theta', \rho - \rho')\zeta^\#(\theta', \rho')\, d\rho'\, d\theta' \\
&= \int_{-\beta}^{\beta} \int_{\ln(a_m)}^{0} \left(\sum_j \sum_i h^\nu[j, i]\varphi\left(\frac{\theta - \theta' - \theta_j}{\Delta\theta}, \frac{\rho - \rho' - \rho_i}{\Delta\rho}\right)\right)\zeta^\#(\theta', \rho')\, d\rho'\, d\theta' \\
&= \sum_j \sum_i h^\nu[j, i]Z(\theta - \theta_j, \rho - \rho_i), \qquad (4.2)
\end{aligned}$$

where

$$Z(\theta, \rho) = \int_{-\beta}^{\beta} \int_{\ln(a_m)}^{0} \varphi\left(\frac{\theta - \theta'}{\Delta\theta}, \frac{\rho - \rho'}{\Delta\rho}\right) \zeta^\#(\theta', \rho')\, d\theta' d\rho'. \qquad (4.3)$$

This follows by exchanging the order between sums and integrals. Note that Z is independent of the Radon data.

Due to this structure, it is particularly convenient to make reconstructions on some uniformly sampled (θ, ρ)-grids, e.g., the same upon which h^ν was resampled, as this enables computation of the discrete convolutions by means of two-dimensional FFT (after appropriate zero padding). Once the partial back-projections f_r^ν are computed on the (θ_j, ρ_i)-grid, it remains only to interpolate them onto a Cartesian representation and sum up the results to obtain a reconstruction f. A survey of relevant interpolation methods is given in the appendix.

The procedure discussed above involves three quasi interpolators. The first one (not discussed above) is needed in the resampling onto the uniform (θ_j, ρ_i)-grid; let it be denoted by $I_{\Delta_1} = I_{\frac{2}{q}}$, and its interpolating order by o_1. Hence, $h[j, i]$ above should be replaced by $(I_{\Delta_1} h)[j, i]$. The second quasi interpolator, represented by I_{Δ_2} above, is naturally incorporated into $Z(\theta, \rho)$ as described by (4.3), and the third, I_{Δ_3} of interpolating order o_3, is needed when adding up the parts. In short, a pseudocode for the reconstruction is presented below.

ALGORITHM 4.1.

```
function f=iradonlp(g,m,mbeta);
    [p,q]=size(g);
    h=wfilter(g);
    FZ=fft2(get_Z(p,q,mbeta));
    chi=get_chi(m,mbeta,q);
    f=0;
```

```
    for nu=0:m-1,
        hlp=interp_pol2lp(chi.*h(nu*p/m+(0:p/mbeta-1),:));
        hlp=[hlp;zeros(size(hlp))];
        flp=ifft2(FZ*fft2(hlp));
        f=f+interp_lp2cart(flp);
    end;
end;
```

Let us discuss the first interpolation step in slightly more detail. Since, for each fixed θ, $h(\theta, s)$ is band-limited by b, the first interpolation step can actually be computed exactly from $h(\theta_j, s_l)$ by (A.1), as long as the Nyquist condition $b < \frac{\pi q}{2}$ is satisfied. However, this is quite time-consuming and ruins the time gain achieved by using log-polar coordinates in the back-projection.

The nonuniformity between data in polar and log-polar representation requires use of more sample points in the log-polar representation in order to avoid too much loss of information. Suppose $\Delta\rho = \frac{-\ln(a_m)}{q'-1}$ and $\rho_i = 1 - i\Delta\rho$, $i = 0, \ldots, q'-1$, where $q' = \kappa q$ for some oversample factor κ. If the filter bandwidth of (4.1) is given by $b = \frac{\pi q}{2}$, then by choosing

$$\kappa > -\frac{\ln(a_m)}{2a_R},$$

the knowledge of $h(\theta, \rho_i)$ at all grid points suffices to reconstruct $h(\theta, \rho_i)$; cf. Theorem 3.1 in [22]. A typical choice here is $\kappa = 2$.

Remark. The combination of uniformly spaced FFT and the interpolation scheme described above is in principle the same as in the procedures used in fast Fourier transforms for unequally spaced data; cf. [2]. Usage of the fast implementations available for such routines allows simple and fast implementation of the algorithm described above. It should be stressed that, although the tools used are the same as in, e.g., [16], the underlying method is quite different; the approach of this paper is based on the filtered back-projection technique, whereas the others have been based on the Fourier slice theorem.

4.2. Error analysis. In the algorithm described above, errors are introduced at several places. The first one is in the filtering step, caused by the discrete convolution (4.1). This is common for algorithms of filtered back-projection type, and estimated in [14] by

$$|w_b \overset{d}{*} g - w_b \overset{s}{*} g|(\theta, s) \leq \frac{1}{2} \int_{\sigma \geq b} |\sigma|^{n-1} \hat{f}(\sigma\theta)\, d\sigma = e_1. \tag{4.4}$$

To describe the errors caused by interpolation we need to introduce modified Sobolev norms for polar and log-polar coordinates. The main reason for not using the natural definitions is that each filtering h of Radon data of some compactly supported function is not compactly supported. In particular, since the filtering does not vanish in a neighborhood of $s = 0$, the natural Sobolev norm in log-polar coordinates of h is in general, if it exists, infinite.

As in the definition (A.4), let

$$||f||_{H^\gamma_{\mathrm{pol}}} = \sum_{|\alpha| \leq \gamma} \int\int |D^\alpha_{\theta,s} f(\theta, s)|^2 \chi_{\mathrm{pol}}(\theta, \rho)\, d\theta\, ds$$

and

$$||f||_{H^\gamma_{\rm lp}} = \sum_{|\alpha|\le\gamma} \int\int |D^\alpha_{\theta,\rho} f(\theta,\rho)|^2 \chi_{\rm lp}(\theta,\rho)\, d\theta\, d\rho$$

be polar and log-polar Sobolev norms, respectively. Here $f(\theta,s)$ and $f(\theta,\rho)$ are connected through the change of variables

$$\rho = \ln\left(\frac{s-(1-a_R)\cos(\theta)}{a_R}\right), \tag{4.5}$$

$\chi_{\rm pol}$ is a smooth cutoff function equal to one on $(-\frac{\beta}{2},\frac{\beta}{2})\times(-1,1)$, and similarly $\chi_{\rm lp}$ equal to one on $(-\frac{\beta}{2},\frac{\beta}{2})\times(-\ln(a_m),0)$. Since the function defined by (4.5) is $\mathcal{C}^\infty$ for $s\in[-1,1]$, and correspondingly $\rho\in[-\ln(a_m),0]$, it is possible to choose $\chi_{\rm pol}$ and $\chi_{\rm lp}$ such that the two norms above are equivalent. It is clear that the results in the appendix are also valid with the norms above. By abuse of notation, we will denote all constants in the remainder of this section by C.

The error introduced in the first interpolation step can now be expressed as

$$||I_{\Delta_1} h^\nu - h^\nu||_{H^0_{\rm pol}} \le C|\Delta_1|^{o_1}||h^\nu||_{H^{o_1}_{\rm pol}}, \tag{4.6}$$

by using (A.5) of Theorem A.1. Because of the norm equivalence between $||\cdot||_{H^\gamma_{\rm pol}}$ and $||\cdot||_{H^\gamma_{\rm lp}}$, the analogous estimate of (4.6) holds for $||\cdot||_{H^\gamma_{\rm lp}}$. This, in combination with the triangle inequality, implies that

$$\begin{aligned}||h^\nu_* - h^\nu||_{H^0_{\rm lp}} &= ||I_{\Delta_2} I_{\Delta_1} h^\nu - h^\nu||_{H^0_{\rm lp}} \le ||I_{\Delta_2} I_{\Delta_1} h^\nu - I_{\Delta_1} h^\nu||_{H^0_{\rm lp}} + ||I_{\Delta_1} h^\nu - h^\nu||_{H^0_{\rm lp}} \\ &\le C|\Delta_2|^{o_2}||I_{\Delta_1} h^\nu||_{H^{o_2}_{\rm lp}} + C|\Delta_1|^{o_1}||h^\nu||_{H^{o_1}_{\rm lp}}.\end{aligned}$$

Since I_{Δ_1} is bounded, it follows that

$$||h^\nu_* - h^\nu||_{H^0_{\rm lp}} \le C(|\Delta_2|^{o_2}||h^\nu||_{H^{o_2}_{\rm lp}} + |\Delta_1|^{o_1}||h^\nu||_{H^{o_1}_{\rm lp}}).$$

Hence, for the partial back-projection

$$||\mathcal{R}^\#(h^\nu_* - h^\nu)||_{H^0_{\rm lp}} \le \beta C(|\Delta_2|^{o_2}||h^\nu||_{H^{o_2}_{\rm lp}} + |\Delta_1|^{o_1}||h^\nu||_{H^{o_1}_{\rm lp}}),$$

since both h^ν and h^ν_* are zero outside a θ-interval of length β. The total error from one partial back-projection is then bounded by

$$\begin{aligned}||I^3\mathcal{R}^\# h^\nu_* - \mathcal{R}^\# h^\nu)||_{H^0_{\rm lp}} &\le ||I^3\mathcal{R}^\# h^\nu_* - \mathcal{R}^\# h^\nu_*)||_{H^0_{\rm lp}} + ||\mathcal{R}^\# h^\nu_* - h^\nu)||_{H^0_{\rm lp}} \\ &\le C|\Delta_3|^{o_3}||\mathcal{R}^\# I_{\Delta_2} I_{\Delta_1} h^\nu||_{H^{o_3}_{\rm lp}} + \beta C(|\Delta_2|^{o_2}||h^\nu||_{H^{o_2}_{\rm lp}} + |\Delta_1|^{o_1}||h^\nu||_{H^{o_1}_{\rm lp}}) \\ &\le C(|\Delta_3|^{o_3}||h^\nu||_{H^{o_3}_{\rm lp}} + |\Delta_2|^{o_2}||h^\nu||_{H^{o_2}_{\rm lp}} + |\Delta_1|^{o_1}||h^\nu||_{H^{o_1}_{\rm lp}}),\end{aligned}$$

and hence also the total back-projection. We summarize the analysis above in the following theorem.

THEOREM 4.1. *The reconstruction error made in the back-projection step of Algorithm* 4.1 *satisfies*

$$||(\mathcal{R}^\#_{\rm lp} - \mathcal{R}^\#)h||_{H^0_{\rm lp}} \le C(|\Delta_3|^{o_3}||h||_{H^{o_3}_{\rm lp}} + |\Delta_2|^{o_2}||h||_{H^{o_2}_{\rm lp}} + |\Delta_1|^{o_1}||h||_{H^{o_1}_{\rm lp}}).$$

We end with a quantitative remark on the estimates above. The frequency content of quasi interpolators is, generally speaking, one at a neighborhood of zero and dies out for higher frequencies; cf. [21]. In principle they mimic sinc-interpolation, where the frequency content of the kernel is a box function centered at the origin. Thus, the error introduced by replacing the operator $\mathcal{R}^{\#}$ with $I_{\Delta_3}\mathcal{R}^{\#}I_{\Delta_2}$ is mainly due to what happens with Radon data far away from the origin. Now, consider the back-projection kernel of Figure 3.1. Apart from along the line $\mu = 0$, most of its energy lies in a neighborhood of the origin. As pointed out above, if data (limited to $s \in [\ln(a_m), 1]$) is smooth when represented in polar coordinates, it is smooth also in the log-polar representation. The fact that data is band-limited with respect to s in combination with the fast decay in the ω-direction should therefore keep the errors relatively small in practice.

4.3. Time complexity and implementation. Let us analyze Algorithm 4.1 in slightly more detail. The first filtering step is the same as in other filtered back-projection algorithms and can be implemented by p FFT operations of length $2q$ in time $\mathcal{O}(pq\log(q))$. At the first interpolation step, each interpolation $(I_{\Delta_1}h^{\nu})(\theta_j, \rho_i)$ consists of a weighted sum of $h(\theta_j, s_l)$ at points where s_l is within a kernel length distance from ρ_i. Hence this step is $\mathcal{O}(d_1pq') = \mathcal{O}(d_1p\kappa q)$, with d_1 the kernel length of I_{Δ_1}. Similarly, each interpolation from log-polar coordinates to a Cartesian $q \times q$-grid is made by a weighted sum, in both θ and ρ directions, giving a time complexity of $\mathcal{O}(d_3q^2)$, where d_3 is the kernel area of I_{Δ_3}. Assuming both d_1 and d_3 to be relatively small, and using the optimal relation between p and q in parallel beam geometry, $p = \frac{\pi q}{2}$ (cf. [14]), we arrive at a time complexity of $\mathcal{O}(q^2)$ for both interpolation steps.

Note that both for the interpolation weights above as well as for the computation of the kernel Z defined in (4.3), only geometry matters; i.e., these can be precomputed to save time. What remains are then the two-dimensional FFT steps. Taking into account the needed zero-padding, $2m$ two-dimensional FFT operations of size $(\frac{2p}{m} \times q')$ are required. This is performed in time $\mathcal{O}(pq\log(pq)) = \mathcal{O}(q^2\log(q))$, i.e. at the same complexity as other fast reconstruction methods. However, in comparison to, e.g., Fourier slice-reconstruction on a $q \times q$ grid, the constant of the leading term is worse. In principle this is due to oversampling factors of κ and 2 in the s and θ directions, respectively, and to the fact that both transformation and inversion are used. Together, this causes a worsening by a factor of about ten. However, that is in comparison with the most simple Fourier slice-reconstruction, with its well-known severe drawbacks in quality and without the effort of speeding up our proposed method (it is possible to use the zero-padded structure to decrease needed calculations). It should be added that more sophisticated slice-reconstruction schemes also require oversampling for accurate results.

Throughout section 4.1 we worked with quasi interpolators. The structure of Algorithm 4.1 easily incorporates usage of the prefiltering required by, e.g., spline interpolators discussed in section A. Required prefiltering for I_{Δ_1} can thus be included in w_b of (4.1), and prefiltering needed for I_{Δ_2} and I_{Δ_3} can be included in Z, defined by (4.3). This allows the usage of spline interpolators without increasing the computational cost, and thus allows higher interpolation orders than the ones achieved by interpolators of the same kernel length.

4.4. Simulations. Finally, let us look at some numerical results. In these simulations we have used cubic spline interpolators for I_{Δ_1}, I_{Δ_2}, and I_{Δ_3}; $m = 4$ partial reconstructions; $m_\beta = 3$; and $\kappa = 2$. The number of parallels and directions

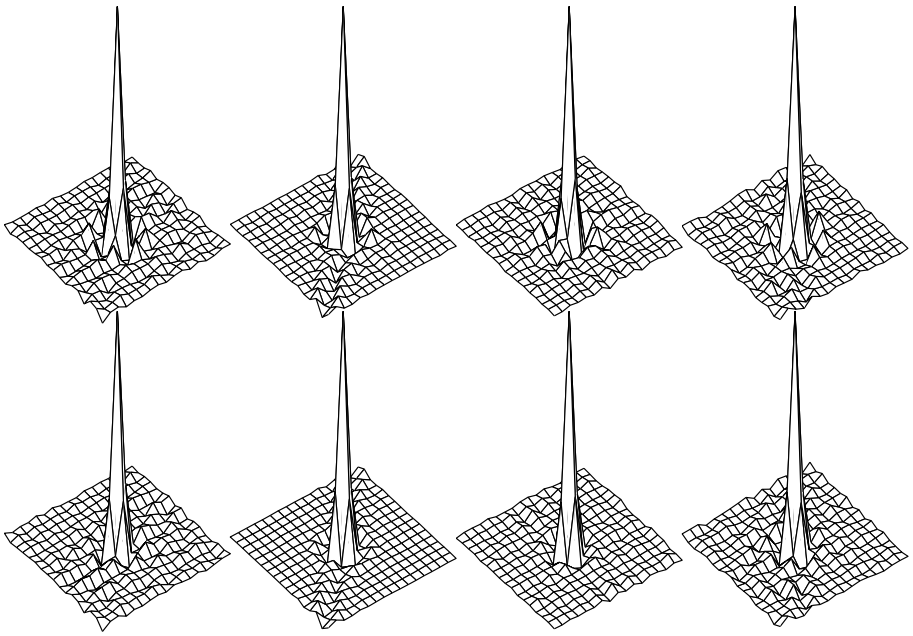


FIG. 4.1. *Impulse responses for log-polar (above) versus classical (below) back-projection at points* $(0.1233, 0.3816)$, $(0.0998, 0.0998)$, $(0.8004, -0.0020)$ *and* $(-0.3229, -0.2368)$, *respectively, from left to right.*

are $q = 512$ and $p = 768$, respectively. Furthermore, to limit the regularization impact from the choice of filter in the filtering step, the *Ram–Lak* filter suggested by Ramachandran and Lakshminarayanan in [17] is used. One of the most commonly used filters in CT is the *Shepp–Logan* filter. For purposes of noise-reduction, high frequency content is suppressed when using this filter, and the reconstructions obtained are somewhat smoothed. In the simulations presented below there is no noise present. We want to compare our proposed reconstruction method with the standard one, and the most honest way of doing this is without any smoothing present. The tests performed with the Ram–Lak filter are thus more unmasking than tests with the Shepp–Logan filter.

In, e.g., [11], analytic expressions are derived for the Radon transform of functions being characteristic functions of elliptical discs. We use such functions as reconstructions objects, since they allow us to sample the Radon data exactly.

To test the performance of our proposed method, we consider two types of test objects. First we use the characteristic function of small circles as approximations of delta functions, and then apply the *Shepp–Logan* phantom, described in [11], to analyze an artificial slice of a head built up by characteristic functions of ellipses.

In Figure 4.1 are shown impulse responses at four different points for log-polar and classically computed back-projection. The reconstructions show quite high resemblance to one another. Reconstruction on a Cartesian grid makes the impulse response vary slightly, depending on the placement of the impulse. Recall that the responses usually are wider and smoother when the ordinary filters (such as the Shepp–Logan filter) of the filtering step are used.

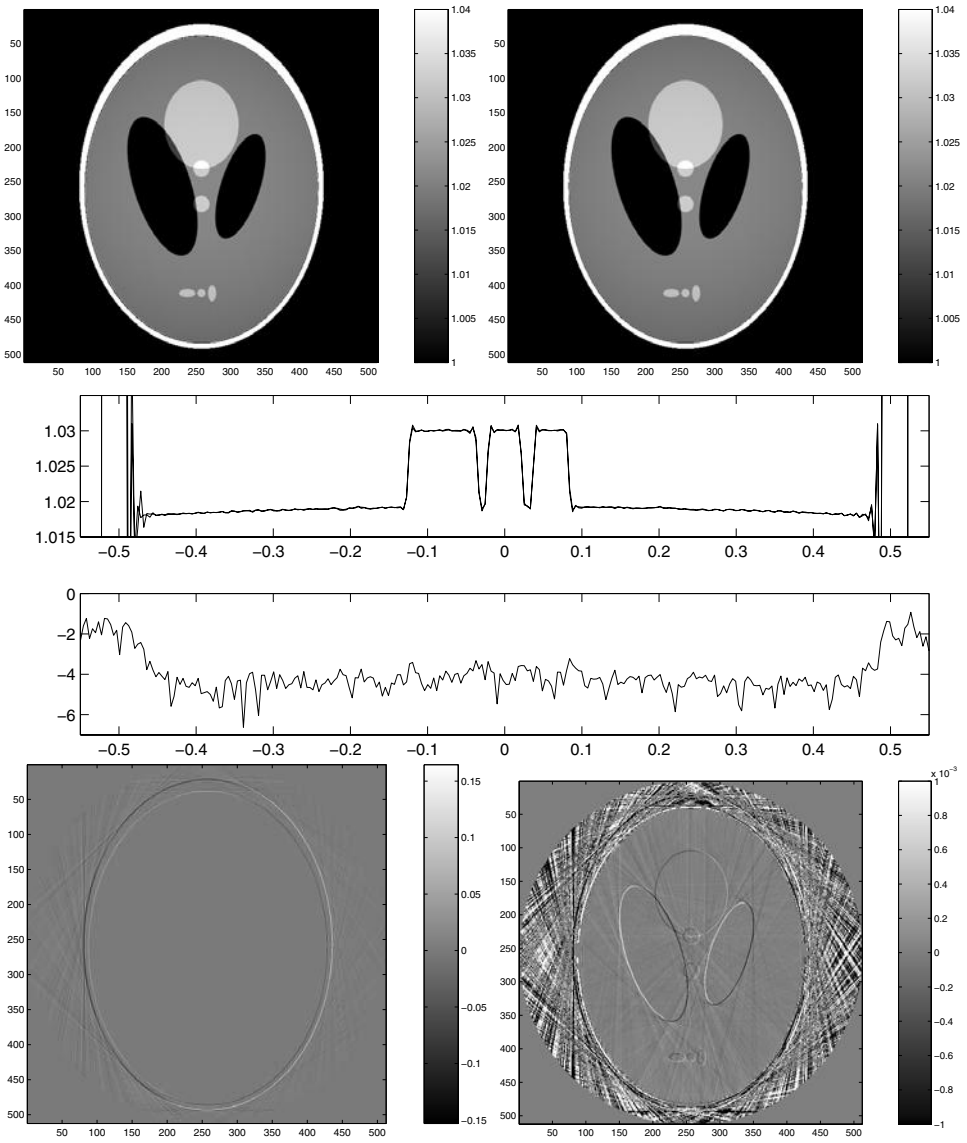


FIG. 4.2. *Upper left shows reconstruction using log-polar coordinates, upper right classical reconstruction. The two plots at center show comparisons between the two reconstructions along a horizontal line, where the upper one show their values, and the lower one their difference in* $\log_{10}$*-scale. The two lower images show the difference between the two reconstructions at different scales.*

Next we turn our attention to the Shepp–Logan phantom; cf. Figure 4.2. The true head phantom carries an intensity of 1.02 at its inner part, a border of intensity 2.0, "eyes" at 1.0, and additional ellipses at 1.03 (except at intersections). The two top images of Figure 4.2 show reconstruction by log-polar (right) and traditional (left) back-projection. Visually they appear to be identical, and it is hard to draw further conclusions. For a better perspective, we have chosen to plot the two reconstructions along a horizontal line through the "mouth" part of the phantom. This is shown

in the upper plot of Figure 4.2. More precisely the values of the two reconstructions along row 410 have been plotted. Except for minor disparities at the points of discontinuity, visually the two reconstructions agree completely. The (absolute) difference between the two reconstructions has therefore been plotted in $\log_{10}$-scale in the lower of the plots in Figure 4.2. This illustrates the high accuracy of our proposed method, compared to standard methods of higher complexity. For comparison, similar plots of the "mouth" part are made in [6], but there, differences between the different reconstruction methods investigated can be seen in the (linear scaled) plots.

For further comparison, the differences between the two reconstructions are shown as images below in Figure 4.2, with different intensity scales. The left panel displays differences in the reconstruction of the high intensity border around the skull. At right we can also see differences at the other discontinuity parts, but note the low intensity span. This image also clearly shows a slightly inhomogeneous assembling of lines outside the head, an effect of the different techniques used.

5. Conclusion. In this paper we have made use of the invariance properties of the Radon transform and its dual to construct a method of inversion based on log-polar representations. An analysis of the continuous case as well as analytical expressions for the kernels have been presented. To deal with the nonuniformity of the discrete case, the concept of partial reconstructions was introduced. The nonuniformity also enforces the discrete reconstruction developed to rely on interpolation. Fortunately, these can be invoked in the reconstruction procedure, for construction of a fast and accurate back-projection algorithm, based on use of the (uniform) two-dimensional FFT. The algorithm presented has a time complexity of $\mathcal{O}(q^2 \log(q))$, the same as those of other fast reconstruction algorithms. From the tests presented in this paper, the accuracy of the method seems to be of the same order as the traditional back-projection algorithms, having time complexity $\mathcal{O}(q^3)$.

Appendix. Some properties of convolution interpolators. According to the Whittaker–Shannon sampling theorem, for any b-band-limited function f, i.e., a function with a Fourier transform $\hat{f}(\xi)$ vanishing for $|\xi| > b$, it is possible to reconstruct f given the uniformly sampled values of f at $x_k = kT$, where $T = \pi/b$, $k \in \mathbb{Z}$. The reconstruction formula reads as

$$f(x) = \sum_{k\in\mathbb{Z}} f(x_k)\mathrm{sinc}(x - x_k). \tag{A.1}$$

A drawback of this *sinc*-based interpolation is the slow decay of the sinc-function. The sinc-interpolation is an example of convolution-based interpolation for uniformly sampled data:

$$I_T f(x) = \sum_{j\in\mathbb{Z}} c_j[f]\varphi\left(\frac{x}{T} - j\right), \tag{A.2}$$

a semidiscrete convolution of a set of coefficients and an interpolation kernel φ. The coefficients $c_j[f]$ are generally obtained from

$$c_j^T[f] = \mu\left(f\left(\frac{\cdot + j}{T}\right)\right), \tag{A.3}$$

where μ is a linear functional. For an overview of convolution-based interpolation, see [13]. The operator I_T is said to be an *interpolator* (sometimes referred to as cardinal interpolator) if $I_T f$ coincides with f at the sample points.

In some cases it is of more interest if I_T manages to reconstruct polynomials of a certain order correctly. To this end, I_T is said to be a *quasi interpolator* of order p if it successfully interpolates all polynomials of degree $p-1$. In order to construct a quasi interpolator of order p from an interpolation kernel φ, φ must satisfy the *Strang–Fix Conditions*; cf. [20]. Of course, there are additional requirements on the coefficients $c_j[f]$.

In this paper, the error estimates in approximating a function by (A.2) are carried out in Sobolev-norms, where $H^\gamma(\mathbb{R}^n)$ or H^γ of real order γ consists of the functions f such that

$$||f||_{H^\gamma}^2 = \sum_{|\alpha|\leq\gamma} ||D^\alpha f||_{L_2(\mathbb{R}^n)}^2 \quad \text{is finite,} \tag{A.4}$$

where

$$D^\alpha f = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n}, \quad \alpha = (\alpha_1, \ldots, \alpha_n).$$

The following theorem is a somewhat weaker version of Theorem 2.2 in [5].

THEOREM A.1. *Let I_T be a quasi interpolator of order p, defined by (A.2), where φ satisfies the Strang–Fix conditions and, in addition, φ and μ of (A.3) satisfy the conditions given in Theorem* 2.1 *in* [5]. *Then for $f \in H^p$ the following estimate holds:*

$$||f - I_T f||_{H^0} \leq C|T|^p ||f||_{H^p}. \tag{A.5}$$

An important type of interpolators are the spline interpolators [18], [19]. The B-spline kernel of order m is defined by

$$\varphi_{B^m} = \left(\frac{1}{2}\chi_{[-\frac{1}{2},\frac{1}{2}]} + \frac{1}{2}\chi_{(-\frac{1}{2},\frac{1}{2})}\right) * \cdots * \left(\frac{1}{2}\chi_{[-\frac{1}{2},\frac{1}{2}]} + \frac{1}{2}\chi_{(-\frac{1}{2},\frac{1}{2})}\right) \quad (m+1 \text{ factors}).$$

It is readily verified that φ_{B^m} consists of piecewise polynomials of degree m, and that it is supported in $(-\frac{m+1}{2}, \frac{m+1}{2})$.

A quasi interpolator of order m can be constructed with φ_{B^m} as interpolation kernel, and with

$$c_k[f] = \sum_{j\in\mathbb{Z}} b_k f(x_{j-k}), \tag{A.6}$$

where

$$\sum_{k\in\mathbb{Z}} b_k e^{-i\xi k} = \frac{1}{\sum_{k\in\mathbb{Z}} \varphi_{B^m}(k) e^{-i\xi k}}, \quad \xi \in \mathbb{R}.$$

This follows, e.g., from [12], [21]. The construction of the coefficients $c_k[f]$ in (A.6) can be interpreted as a prefiltering step, which preferably is computed by means of FFT. Note that since the support of the B-spline kernel is relatively short, the number of terms needed in (A.2) are relatively few. Hence, an accurate interpolation procedure can be constructed which, in contrast to the sinc-interpolation of (A.1), requires data in only a small neighborhood around the point of evaluation.

REFERENCES

[1] S. Alliney, *Digital reconstruction of images from their projections on polar coordinates*, Signal Process., 3 (1981), pp. 135–145.

[2] G. Beylkin, *On the fast Fourier transform of functions with singularities*, Appl. Comput. Harmon. Anal., 2 (1995), pp. 363–381.

[3] M. L. Brady, *A fast discrete approximation algorithm for the Radon transform*, SIAM J. Comput., 27 (1998), pp. 107–119.

[4] A. Brandt, J. Mann, M. Brodski, and M. Galun, *A fast and accurate multilevel inversion of the Radon transform*, SIAM J. Appl. Math., 60 (1999), pp. 437–462.

[5] H. G. Burchard and J. Lei, *Coordinate order of approximation by functional-based approximation operators*, J. Approx. Theory, 82 (1995), pp. 240–256.

[6] P. Danielsson and M. Ingerhed, *Backprojection in $\mathcal{O}(N^2 \log N)$ time*, IEEE Nucl. Sci. Symp., 2 (1997), pp. 1279–1283.

[7] P. P. B. Eggermont, *Tomographic reconstruction on a logarithmic polar grid*, IEEE Trans. Med. Imag., MI-2 (1983), pp. 40–48.

[8] K. Fourmont, *Non-equispaced fast Fourier transforms with applications to tomography*, J. Fourier Anal. Appl., 9 (2003), pp. 431–450.

[9] W. A. Götz and H. J. Druckmüller, *A fast digital Radon transform—An efficient means for evaluating the Hough transform*, Pattern Recognition, 28 (1995), pp. 1985–1992.

[10] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, 6th ed., Academic Press, San Diego, CA, 2000 (translated from the Russian; translation edited by A. Jeffrey and D. Zwillinger).

[11] A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*, Classics in Appl. Math. 33, SIAM, Philadelphia, PA, 2001.

[12] E. Maeland, *On the comparison of interpolation methods*, IEEE Trans. Med. Imag., 7 (1988), pp. 213–217.

[13] E. Meijering, *A chronology of interpolation. From ancient astronomy to modern signal and image processing*, Proc. IEEE, 90 (2002), pp. 319–342.

[14] F. Natterer, *The Mathematics of Computerized Tomography*, B. G. Teubner, Stuttgart, 1986.

[15] S. Nilsson, *Applications of Fast Backprojection Techniques for Some Inverse Problems of Integral Geometry*, Ph.D. thesis, Department of Mathematics, Linköping University, Sweden, 1997.

[16] D. Potts and G. Steidl, *A new linogram algorithm for computerized tomography*, IMA J. Numer. Anal., 21 (2001), pp. 769–782.

[17] G. N. Ramachandran and A. V. Lakshminarayanan, *Three-dimensional reconstruction from radiographs and electron micrographs: Application of convolutions instead of Fourier transforms*, Proc. Natl. Acad. Sci. USA, 68 (1971), pp. 2236–2240.

[18] I. J. Schoenberg, *Contributions to the problem of approximation of equidistant data by analytic functions. Part A. On the problem of smoothing or graduation. A first class of analytic approximation formulae*, Quart. Appl. Math., 4 (1946), pp. 45–99.

[19] I. J. Schoenberg, *Contributions to the problem of approximation of equidistant data by analytic functions. Part B. On the problem of osculatory interpolation. A second class of analytic approximation formulae*, Quart. Appl. Math., 4 (1946), pp. 112–141.

[20] G. Strang and G. Fix, *A Fourier analysis of the finite element variational method*, in Constructive Aspects of Functional Analysis II, G. Geymonat, ed., Cremonese, Rome, 1973, pp. 793–840.

[21] M. Unser, A. Aldroubi, and M. Eden, *Polynomial spline signal approximations: Filter design and asymptotic equivalence with Shannon's sampling theorem*, IEEE Trans. Inform. Theory, 38 (1992), pp. 95–103.

[22] A. I. Zayed, *Advances in Shannon's Sampling Theory*, CRC Press, Boca Raton, FL, 1993.

SIAM J. APPL. MATH.
Vol. 65, No. 3, pp. 838–857

ON THE DESIGN AND ANALYSIS OF INFLATED MEMBRANES: NATURAL AND PUMPKIN SHAPED BALLOONS*

FRANK BAGINSKI†

Abstract. Large scientific balloons are used by NASA and the space agencies of many countries to carry out research in the upper stratosphere. Such a balloon typically consists of a thin plastic shell with several external caps. Load tendons run the length of the balloon from top fitting to bottom fitting, dividing the balloon into identical regions called gores. The gores are made from flat panels of 20–30 μm polyethylene film that are sealed edge-to-edge to form the complete shape. A typical fully inflated shape can be over 120 meters in diameter and over 1 million cubic meters in volume. To date, the workhorse of NASA's balloon program has been the zero-pressure natural shape balloon, an axisymmetric onion-like design that dates back to the 1950s. The equilibrium equations at float for a natural shape balloon lead to a nonlinear boundary value problem that can be solved to determine the design shape. In recent years, demand for long duration midlatitude balloon flights has led to a design concept known as the pumpkin balloon. A number of ad hoc approaches based on crude approximations of equilibrium have been put forth that lead to pumpkin-like balloon shapes. In this paper, we derive equilibrium equations for a pumpkin balloon. We also present a brief review of balloon models that follow from the axisymmetric membrane theory. Numerical solutions are included.

Key words. high altitude balloons, inflated membranes

AMS subject classification. 74K15

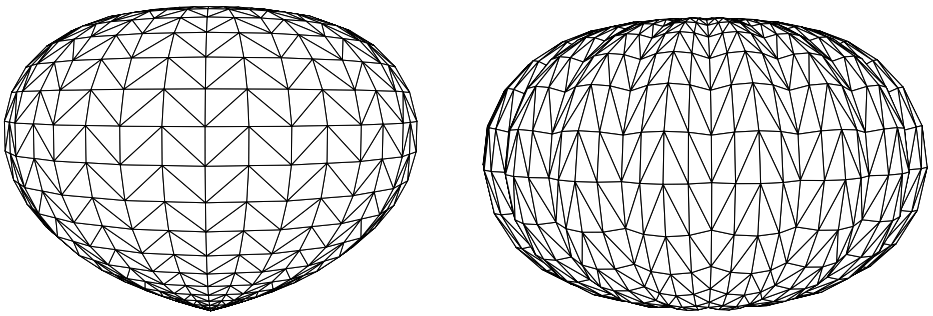
DOI. 10.1137/S0036139903438478

1. Introduction.

Balloons play an important role in NASA's current scientific investigations, including upper atmosphere research, high energy astrophysics, stratospheric composition, meteorology, and astronomy. With the development of the ultra long duration balloon and the possible uses of balloons in the exploration of planets in our solar system, balloons will play an important role in NASA's future scientific endeavors.

A high altitude large scientific balloon is normally designed to carry a payload of instruments to a certain altitude and then to maintain constant altitude while the science is carried out. The theoretical shape of a fully inflated high altitude balloon at float altitude is called the design shape. The design shape is normally modeled as an inextensible membrane. Quantities such as film weight density w_f, payload L, specific buoyancy of lifting gas b, circumferential stress σ_c, and differential pressure at the base of the balloon p_0 are parameters that affect the final shape. An axisymmetric model commonly used for applications to large scientific balloons is the *natural shape* model, which assumes that $\sigma_c = 0$, i.e., that all tension is meridional. When $p_0 = 0$ in the natural shape model, we have the so called *zero-pressure natural shape* (ZPNS) balloon (see Figure 1.1(a)), a design that NASA flies regularly in its scientific balloon program. Venting ducts open to the atmosphere are located at the base of a ZPNS balloon. During daylight, when the balloon envelope and lifting gas are heated, excess gas is vented through these ducts. At night, when the gas cools and the volume

*Received by the editors December 7, 2003; accepted for publication (in revised form) June 1, 2004; published electronically February 25, 2005. This work was supported in part by the National Aeronautics and Space Administration Award NAG5-5353.
http://www.siam.org/journals/siap/65-3/43847.html
†Department of Mathematics, The George Washington University, Washington, DC 20052 (baginsk@gwu.edu).



(a) Zero-pressure natural shape (b) Pumpkin shape

FIG. 1.1. *A* 20*-gore zero-pressure natural (onion-)shape balloon and a* 20*-gore pumpkin balloon.*

of the balloon decreases, ballast is dropped to maintain altitude. The number of diurnal cycles limits ZPNS balloons to midlatitude flights of a few days due to ballast limitations. ZPNS flights of several weeks have been achieved in Antarctica where day/night diurnal cycles can be avoided during certain times of the year. One way to avoid carrying significant ballast is to design a balloon that can hold enough gas to maintain a positive differential pressure during the night and is strong enough to hold the overpressure caused by solar heating during the day. Thus, by reducing the volume variation over diurnal cycles, altitude excursions will be reduced and less ballast will be required during the life of the mission. This approach led to a design that has come to be known as the *pumpkin balloon* (see Figure 1.1(b)), a concept that goes back to the 1970s. While there have been a number of pumpkin-like balloon models proposed, many are based on ad hoc analysis and approximate equilibrium conditions. In this paper, we develop equilibrium equations for a pumpkin balloon based on a simplification of a *unishell* theory. As defined by Libai and Simmonds [17, p. 343], a unishell is a shell in which there is a system of coordinates (σ, τ) on the reference surface in which the components of the deformed metric and curvature tensors are independent of τ. In the case of a pumpkin balloon, we will develop a system of equilibrium equations that depend only on the arc-length of the generating curve for a tubular surface corresponding to one lobe of the pumpkin shape. We formulate the equilibrium equations for a pumpkin balloon in the deformed configuration and derive a set of nonlinear ordinary differential equations (ODEs). The ODEs can be solved to determine the generating curve for the tubular surface defining the pumpkin balloon.

The original work on high altitude plastic balloons was carried out in the 1950s at the University of Minnesota, where the term natural shape balloon emerged (see [1]). In the 1960–70s, Justin Smalley did extensive work on axisymmetric balloon shapes (see e.g., [22], [23], [24]). He extended the natural shape model to handle a number of related axisymmetric cases, and implemented these mathematical models on a digital computer. Even though many axisymmetric balloon models are discussed in the literature, most follow from the natural shape equations. For the convenience of the reader, we will present a brief review of axisymmetric membrane models. By choosing parameters in the axisymmetric model in a certain way, we are led to various balloon models, including the ZPNS balloon, the super-pressure natural shape balloon, the Σ-shape, the Euler-elastica, and others. For an axisymmetric shape with $\sigma_c \neq 0$, we present one model whose solutions give rise to Delaunay surfaces (the spherical balloon is a special case).

The idea behind the pumpkin balloon is to use a light film as a gas barrier together with strong reinforcing tendons for both pressure confinement and to carry the weight of the balloon system. Roughly speaking, the shape of the pumpkin gore transfers much of the film stress to the load tendons. The term pumpkin balloon was coined by Smalley in his original work on the *e-balloon* (a model whose governing equations are the same as those for the Euler–Bernoulli elastica; "e" is for Euler). Smalley's original idea was to use the straining in an inflated gore to form a doubly curved surface. When fully pressurized, the balloon would take on the appearance of a pumpkin-like shape (see [25]). Shortly after Smalley's work on the e-balloon, the French carried out work on pumpkin balloons (see, e.g., [20]). There have been many variants on the pumpkin design. One notable effort in the 1980s was the *Endeavour* of Julian Nott in the race for the first circumnavigation of the globe. The *Endeavour* was a 64 gore *constant bulge shape design*; i.e., the angular width of the circular arc of a lobe was constant. There were some deployment problems related to this design (see section 3 for further details). The Japanese have studied pumpkin balloons (see, e.g., [26]), introducing cinching techniques in sewing to achieve a pressurized pumpkin shape. Using a *constant bulge radius design* and a tendon-based theory, Schur was the first to give a more analytical treatment of the pumpkin balloon (see [21]).

In ad hoc approaches to pumpkin balloons, it is usual to assume that the tendon profiles are known a priori from axisymmetric models (like those discussed in section 2). While this may be adequate as a rough first approximation, it is straightforward to model the pumpkin gore as a tubular surface and derive the generating curve for the pumpkin gore as the solution of a set of equilibrium equations. The formulation of a membrane model for a pumpkin gore is presented in section 3. Our formulation introduces the number of gores n_g and the bulge radius r_B as additional design parameters in the model equations.

It should be noted that a number of underlying assumptions made to determine shapes are sometimes violated in the real balloon. For example, the assumption of zero circumferential stress cannot hold in a real ZPNS balloon that is elastic: once it is pressurized, the hoop and meridional stresses will be nonzero and roughly the same order of magnitude in the upper portion of the balloon. Moreover, the ZPNS model assumes an axisymmetric surface of revolution, while the real balloon is constructed from long flat panels of film. This is not a problem, due in part to the compliant nature of polyethylene. Decades of successful ZPNS flights demonstrate the effectiveness of the ZPNS balloon.

The real ZPNS balloon has a curved unstrained reference configuration (bend a flat panel so that a developable surface spans the region between adjacent tendons). There is no such analogous unstrained/unwrinkled state of a three dimensional pumpkin balloon that is constructed from flat panels of film. This poses some complications in the construction process and is discussed in section 3. In this paper, our analysis assumes that the material is inextensible. However, we refer the reader to the literature for more on the subject of strained elastic balloons (see, [3], [6], [9], and the references therein).

The remainder of this paper is divided into four sections. In section 2, we provide some historical background on models that have been used in the design of scientific balloons. The most common models are derived from the equations for an axisymmetric membrane, and we establish connections between a number of special models. In section 3, we propose a model for a pumpkin balloon that reduces to a system of ODEs. In section 4, we present numerical solutions and also discuss some aspects of

TABLE 1.1
Design parameters.

Description	Variable	Value
Bulge radius (pumpkin only)	r_B	0.1066 m
Weight of top fitting	L_1	8.45 N
Radius of top/bot fittings	r_1	0.10 m
Number of gores	n_g	96
Tendon weight density	w_t	0.03 N/m
Film weight density	w_f	0.344 N/m^2
Payload	L_0	283 N
Buoyancy at design altitude	b_d	3.441 N/m^3
Buoyancy at sea level	b_0	10.258 N/m^3

the various balloon shape models. Table 1.1 contains design parameters that were used to generate shapes presented in this paper.

2. Axisymmetric membrane models for a balloon. First, it is instructive to consider the problem of "designing" an inextensible spherical balloon. Suppose that we wish to determine the radius R of a spherical balloon that must carry a weight L_0 at a constant design altitude. The density of the lifting gas ρ_g and the density of the atmosphere ρ_a are assumed to be known at the design altitude. Archimedes' Principle states that a balloon enclosing a gas volume V of density ρ_g will exert a net upward force (lift) that is equal in magnitude to the difference between the weight of the displaced air and the weight of the lifting gas, i.e.,

$$\text{Lift} = g\rho_a V - g\rho_g V.$$

The specific buoyancy of the lifting gas is $b = \frac{\text{Lift}}{V} = g(\rho_a - \rho_g)$, where b has the dimensions of force per volume. In our example, the balloon is assumed to be made of a single layer of film with weight density w_f. We assume that n load tendons of weight density w_t run from top to bottom of the balloon. From Archimedes' Principle, we know that a spherical balloon of radius R in equilibrium must satisfy

$$\frac{4}{3}b\pi R^3 = 4w_f\pi R^2 + nw_t\pi R + L_0. \tag{2.1}$$

Solving for R, we find the spherical design shape. It is not difficult to show that there is exactly one solution of (2.1) with $R > 0$ (see [7]). Finding the natural shape or a pumpkin shape is more involved. From force balance equations at float, a system of ODEs can be derived. Solving the ODEs yields the design shape.

2.1. Equilibrium equations for an axisymmetric membrane. The equations for a natural balloon shape were first derived by researchers at the University of Minnesota (see [1]). These equations and other related models are presented in [15, section V]. In [8], design and ascent shapes for axisymmetric membranes are considered for constant film weight density (i.e., the Σ-shape equations). In [7], the film weight density included contributions from external reinforcing caps and load tapes. For convenience, we present a synopsis of the equilibrium equations for an axisymmetric membrane. We choose a coordinate system such that the nadir of the balloon is located at the origin.

Since we seek axisymmetric solutions, we need to find a generating curve $s \to (z(s), r(s))$, where s is the arc-length. Let ℓ_d be the total length of the generating curve. The surface of the balloon $\mathcal{S}$ can be parametrized by $\mathbf{x}(s,\phi)$, where

$$\begin{aligned} \mathbf{x}(s,\phi) &= r(s)\mathbf{e}_1(\phi) + z(s)\mathbf{k}, \quad 0 < s < \ell_d, 0 \le \phi < 2\pi, \\ \mathbf{e}_1(\phi) &= \cos\phi\mathbf{i} + \sin\phi\mathbf{j}, \quad 0 \le \phi \le 2\pi. \end{aligned} \tag{2.2}$$

The set $\{\mathbf{e}_1(\phi), \mathbf{e}_2(\phi), \mathbf{k}\}$ is an orthonormal basis for $\mathbb{R}^3$, where $\mathbf{e}_2(\phi) = \mathbf{k} \times \mathbf{e}_1(\phi) = -\sin\phi\mathbf{i} + \cos\phi\mathbf{j}$. The balloon's shape is defined by $\mathcal{S} = \{\mathbf{x}(s,\phi) \mid s \in [0,\ell_d], \phi \in [0,2\pi]\}$. At each point along the curve, $s \to \mathbf{x}(s,\phi)$, the tangent vector is given by

$$\mathbf{t}(s,\phi) = \frac{\partial \mathbf{x}}{\partial s}(s,\phi) = \sin\theta(s)\mathbf{e}_1(\phi) + \cos\theta(s)\mathbf{k},$$

where $\theta(s)$ is the angle between $\mathbf{t}$ and $\mathbf{k}$, $z'(s) = \cos\theta(s)$, and $r'(s) = \sin\theta(s)$. The inward normal is $\mathbf{b}(s,\phi) = \mathbf{t} \times \mathbf{e}_2(\phi) = -\cos\theta\mathbf{e}_1(\phi) + \sin\theta\mathbf{k}$. Because the balloon is modeled as a membrane, we can ignore all stress couples. Under the assumption of axisymmetry, we can write the contact forces as

$$\begin{aligned} \mathbf{n}_1(s,\phi) &= \sigma_m(s)\mathbf{t}(s,\phi), \\ \mathbf{n}_2(s,\phi) &= \sigma_c(s)\mathbf{e}_2(\phi), \end{aligned}$$

where σ_m is the meridional stress resultant and σ_c the circumferential stress resultant. The forces acting on a curvilinear patch $\mathcal{A} \subset \mathcal{S}$ are the internal forces, $\mathbf{n}_1(s,\phi)$ and $\mathbf{n}_2(s,\phi)$, and the external forces,

$$\mathbf{f}(s,\phi) = -p(s)\mathbf{b}(s,\phi) - w(s)\mathbf{k},$$

where p is hydrostatic differential pressure and w is the balloon film weight density. All external forces are measured per unit area in the deformed configuration. Balancing the forces that act on $\mathcal{A}$, we are led to the following equilibrium equations (see [7]):

$$\frac{\partial}{\partial s}(r\sigma_m\mathbf{t}) - \sigma_c\mathbf{e}_1(\phi) + r\mathbf{f} = \mathbf{0}. \tag{2.3}$$

If we carry out the differentiation in (2.3) and project the result onto $\mathbf{t}$ and $\mathbf{b}$, we obtain

$$(r\sigma_m)\frac{d\theta}{ds} = \sigma_c\cos\theta - rw\sin\theta - rp, \tag{2.4}$$

$$\frac{d}{ds}(r\sigma_m) = \sigma_c\sin\theta + rw\cos\theta, \tag{2.5}$$

respectively. Equations (2.4)–(2.5) are in agreement with those in [1, Vol. I, p. 3-2, (6)–(7)].

The hydrostatic pressure can be written as $p = bz + p_0$, where p_0 is the pressure at the base of the balloon ($z = 0$) and $b = g(\rho_a - \rho_g)$. While b is assumed to be known as a function of altitude, it depends on a number of other parameters such as air and gas temperatures. A natural shape balloon derived under the assumption $p_0 = 0$ is called a *ZPNS* balloon. If $p_0 > 0$, we refer to the corresponding shape as a

superpressure natural shape balloon (SPNS balloon). At the bottom of a zero-pressure balloon are venting ducts which are open to the atmosphere. A superpressure balloon is sealed.

2.2. The natural shape model equations ($\sigma_c = 0$). In this section, we formulate the natural shape model (i.e., $\sigma_c = 0$). It is convenient to define the total film load T in the meridional direction, i.e., $T = 2\pi r\sigma_m$. Multiplying (2.4)–(2.5) by 2π, substituting T for $2\pi r\sigma_m$, and setting $\sigma_c = 0$, we have

$$\begin{aligned}\theta'(s) &= \frac{-2\pi r\,(w\sin\theta + p)}{T},\\ T'(s) &= 2\pi r w\cos\theta.\end{aligned}$$

For applications considered here, we will assume that a load L_0 is suspended from the base of the balloon. In addition, we assume that the balloon is attached to a circular disk of radius r_0 at the bottom and disk of radius r_1 at the top. In a superpressure balloon, a force of magnitude $p_0\pi r_0^2$ acts on the bottom end-cap. Thus,

$$\cos\theta_0 T(0) = L_0 + p_0\pi r_0^2, \tag{2.6}$$

where $p_0 \geq 0$, $\theta(0) = \theta_0$, and θ_0 is one-half the "cone-angle" at the base of the balloon. Note that θ_0 is not known a priori and must be computed based on certain parameter values and boundary conditions.

The governing differential equations for a natural shape balloon in terms of T are

$$\theta'(s) = \frac{-2\pi r(w\sin\theta + p)}{T}, \tag{2.7}$$

$$T'(s) = 2\pi r w\cos\theta, \tag{2.8}$$

$$z'(s) = \cos\theta, \tag{2.9}$$

$$r'(s) = \sin\theta. \tag{2.10}$$

The initial conditions for (2.7)–(2.10) are

$$\begin{aligned}\theta(0) &= \theta_0,\\ T(0) &= \frac{(L_0 + \pi r_0^2 p_0)}{\cos\theta_0},\\ z(0) &= 0,\\ r(0) &= r_0.\end{aligned} \tag{2.11}$$

For a given pair (θ_0, ℓ), (2.7)–(2.10) can be integrated over the interval $0 < s < \ell$ beginning with the initial conditions in (2.11). Using a shooting method, one determines $(\theta_0, \ell) = (\theta_d, \ell_d)$ and the functions $\theta(s;\theta_d,\ell_d)$, $T(s;\theta_d,\ell_d)$, $z(s;\theta_d,\ell_d)$, and $r(s;\theta_d,\ell_d)$ that satisfy (2.7)–(2.10), and

$$\begin{aligned}\cos\theta_1 T(\ell_d;\theta_d,\ell_d) &= -L_1 + \pi r_1^2 p_1,\\ r(\ell_d;\theta_d,\ell_d) &= r_1,\end{aligned} \tag{2.12}$$

where L_1 is the weight of the top fitting, $\theta_1 = \theta(\ell_d;\theta_d,\ell_d)$, $z_1 = z(\ell_d;\theta_d,\ell_d)$, and $p_1 = bz_1 + p_0$. Note that when $L_1 > 0$, we expect $-\pi < \theta_1 < -\frac{1}{2}\pi$. With $p(z) = bz + p_0$, such a solution automatically satisfies Archimedes Principle. If we project (2.3) onto the **k** direction, multiply by 2π, integrate from $s = 0$ to $s = \ell_d$, and use the identity

$$r(bz + p_0)\frac{dr}{ds} = \frac{1}{2}\frac{d}{ds}\left(r^2(bz + p_0)\right) - \frac{1}{2}br^2\frac{dz}{ds},$$

we obtain $T(\ell_d)\cos\theta_1 - T(0)\cos\theta_0 = 2\pi\int_0^{\ell_d} wrds - b\mathcal{V}$. After rearranging terms and using (2.11) and (2.12), we observe that Archimedes Principle is satisfied; i.e.,

$$L_1 + L_0 + \pi(p_1 r_1^2 - p_0 r_0^2) + W_B = b\mathcal{V}, \tag{2.13}$$

where W_B is the total weight of the balloon system (balloon shell, caps, and load tapes).

In the natural shape model presented in [8], the weight of the caps and load tapes were added to the payload. The model presented here follows the approach in [7], where a variable film thickness and tendons are introduced. Here, we assume that the balloon system includes one cap that is modeled as an added thickness. The length of the cap is denoted by c_1 and is assumed to be known. For a given ℓ, let $s_1 = \ell - c_1$. This means that $w(s_1+0) \neq w(s_1-0)$ and $d\theta/ds$ and dT/ds are discontinuous at s_1. The number of gores in a complete shape is n_g.

For shape determination, we assume that the tendon weight is distributed uniformly over the circumference of the balloon and is incorporated by modifying w appropriately. We divide $n_g w_t$ by $2\pi r(s)$ for $0 < s < \ell$ to get the average weight density of load tape (tendon) with respect to area. We define the *effective film weight density*,

$$w(s) = w_f(s) + w_t(s), \quad 0 < s < \ell, \tag{2.14}$$

where $w_t(s) = n_g w_{tape}/(2\pi r(s))$, $0 < s < \ell$,

$$w_f(s) = \begin{cases} w_s, & 0 \le s \le s_1, \\ w_s + w_c, & s_1 < s \le \ell, \end{cases}$$

w_s is the weight density of a single layer of balloon film, and w_{tape} is the load tape weight density. Additional caps with different weight densities can be incorporated easily. As a structural element, a load tendon and load tape behave very differently. However, for a statically determinate balloon, weight is the only consideration that is relevant.

In previous work (see, e.g., [7]), we handled the discontinuity in the effective film weight density by using a parallel shooting method to break the integration of (2.7)–(2.10) into subintervals. Doing so, we see that all functions are continuously differentiable on each subinterval. Existence of solutions obtained by the shooting method follows from standard arguments (see, e.g., [14]).

2.3. Axisymmetric balloon models related to the natural shape. In the following subsections we present a number of balloon models that are based on the natural shape equations (2.7)–(2.10). See [4] for further discussions on balloon models.

2.3.1. ZPNS balloons.

Constant film weight density (Σ-shapes). In the early days of scientific ballooning and before digital computers were readily available, balloon designers relied on the Σ-shape model. Assuming $p_0 = 0$ and $w(s) = w_s$ in (2.7)–(2.10), one finds that the resulting equations can be rescaled in a convenient way using parameters λ and Σ (see [15, p. V-8]), where $\lambda^3 = L_0/b$, $\Sigma = (2\pi/L_0)^{1/3}(\mathcal{V}_d/G)^{2/3} w_s$, G is the gross weight of the system, and $\mathcal{V}_d$ is the volume of the balloon at float. For this reason, the resulting shapes were called Σ-shapes. See [8] for a discussion on Σ-shapes and related axisymmetric ascent shapes for which the volume $\mathcal{V} \le \mathcal{V}_d$. See [1] and [15] for more on Σ-shapes. The Σ-shape equations were first solved using an analog

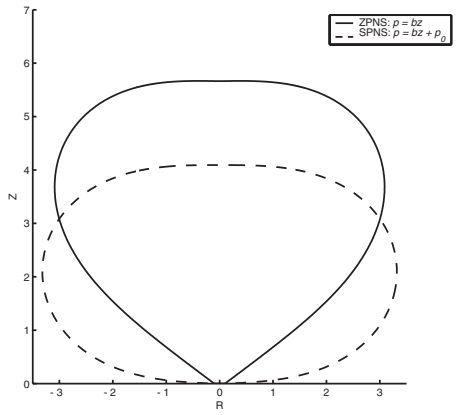


FIG. 2.1. *Comparison of ZPNS and SPNS balloon profiles.*

computer (see [1]). Using a set of tables with various values of λ and Σ, the balloon designer could look up the appropriate value of θ_0 and determine the gore length in terms of λ.

Nonconstant film weight density. When $w(s) = w_f(s) + w_t(s)$ is the effective film weight density, this model will yield a shape that, in general, is slightly shorter and fatter than the one determined by the Σ-shape model (see [7]). When $w(s)$ is not constant, then the basic equations cannot be rescaled as was done for the Σ-shapes. Smalley was the first to use a digital computer in the analysis of balloon shapes, and carried out extensive research and development of other related models (see, e.g., [22], [23], [24], [25]). Variants of this model are used in the design and manufacture of high altitude zero-pressure balloons.

2.3.2. Superpressure balloons. High altitude balloons are very sensitive to temperature changes that are experienced during a day/night cycle. A ZPNS balloon will lose significant altitude during the evening when the lifting gas cools and its volume contracts. On the other hand, the lifting gas will expand during the day when the balloon is heated by the sun. To reduce volume fluctuations, one idea is to overpressurize the balloon in such a way that the differential pressure at the nadir is positive for all conditions (day/night, clear/cloudy, etc.). While the axisymmetric inextensible membrane theory suggests that a SPNS balloon might work, a stress analysis of an elastic SPNS balloon shows that when p_0 is too large, the corresponding stress resultants for currently available films would be outside the safe operating range (see [3]). In some ad hoc approaches to the pumpkin balloon, the tendons are assumed to lie along the trace of the generating curve of a SPNS balloon, and so for completeness we include results on SPNS balloons.

SPNS balloons. In a superpressurized natural shape balloon, we include $p_0 > 0$ as an additional design parameter and solve (2.7)–(2.10) to compute the balloon shape. Compared to the zero-pressure design, the superpressure natural shape balloon is, in general, shorter and wider (see Figure 2.1).

Euler elastica and the e-balloon. In this model, it is supposed that $bz_1 \ll p_0$ near float. We neglect film and load tape weight contributions in the ODEs, and set $p = p_0$ for shape determination. Weight contributions due to the load tapes, film, and caps are added to the payload. The length of the generating curve must be chosen so

that $b\mathcal{V}$ is equal to the weight of the balloon system. In particular, setting $b = 0$ and $w = 0$ in (2.7)–(2.10), we find that T must be constant and $\theta'(s) = -2\pi r p_0/T$. Setting $T = T_0$ and differentiating this result, we find $\theta''(s) = -2\pi p_0 r'(s)/T_0$ and

$$\theta'' + 2\tau \sin\theta = 0, \tag{2.15}$$

where $\tau = \pi p_0/T_0$. In this model, T_0 is the total tension. Typical boundary conditions that are used in conjunction with (2.15) are

$$\theta'(0) = \theta'(\ell) = 0. \tag{2.16}$$

For each ℓ, there is a one parameter family of shapes parametrized by θ_0. Equation (2.15) can be solved explicitly in terms of elliptic functions. Following Smalley's discussion in [25], we note that the radial component of the generating curve for the e-balloon is

$$r^2(\theta) = \tau^{-1} \sin\left(\frac{1}{2}\pi - \theta\right). \tag{2.17}$$

An expression for z can be found by integrating the appropriate differential equation or by using elliptic functions (see [16]). In [18], the problem of constructing a mylar balloon from two circular disks is considered. While this approach is not practical for large balloons, the shapes of [18] are not unrelated to the Euler-elastica balloon discussed here and the Delaunay surfaces discussed in the following section.

In Figure 2.1, we present axisymmetric balloon designs based on the ZPNS and the SPNS models. Design parameters and related quantities are presented in Tables 1.1 and 2.1. Numerical results will be discussed in more detail in section 4.

TABLE 2.1
Design values.

(a) Axisymmetric shapes

Description	ZPNS	SPNS	Elastica
Nadir pressure (Pa)	0	200	200
Base angle (deg)	53.21	87.74	87.23
Height (m)	5.664	4.091	4.153
Diameter (m)	6.178	6.638	6.690
Tendon length (m)	8.988	8.547	8.628
Volume (m^3)	102.07	101.91	104.73
Surface area (m^2)	107.39	109.40	111.32
Skin weight (N)	36.91	37.60	0
Tape weight (N)	25.89	24.62	0
Load (N)	280	280	353

(b) Pumpkin shape: $r_B = 0.1066$ m and $p_0 = 200$ Pa

Description	$\sigma_c = 0$ N/m	$\sigma_c = 10$ N/m
Base angle (deg)	87.75	85.94
Height (m)	4.178	4.234
Diameter (m)	6.691	6.694
Volume (m^3)	104.15	104.46
Surface area (m^2)	133.50	133.7
Skin weight (N)	45.89	45.96
Tendon weight (N)	24.48	24.52
Tendon length (m)	8.501	8.513
Flat Tendon length (m)	8.662	8.674
Mid gore length (m)	8.659	8.668

2.4. Axisymmetric balloons with $\sigma_c \neq 0$. Smalley considered balloons with nonzero circumferential stress in [23]. A spherical balloon under a buoyant load is treated in [15, pp. V-19–V-21]. Balloons of this type have not been used as extensively as the natural shape for high altitude balloons.

If we eliminate s in the problem formulation and use z to parametrize the generating curve (i.e., $(r, z) = (r(z), z)$), we find that the curvature $\theta'(s)$ can be expressed as $r''/(\sqrt{1+r'^2})^3$, and we can replace (2.4)–(2.5) by the following:

$$\frac{rr''\sigma_m}{(\sqrt{1+r'^2})^3} = \frac{\sigma_c - wrr'}{\sqrt{1+r'^2}} - pr, \tag{2.18}$$

$$\frac{d}{dz}(r\sigma_m) = \sigma_c r'(z) + wr,$$

where $r'(z) = dr/dz$ (see [1, Vol. I, p. 3-2]). If the weight of the film and buoyancy are negligible (i.e., $w = 0$ and $b = 0$), and we set $p = p_0$, with σ_m and σ_c constant, then (2.19) is satisfied when $\sigma_c = \sigma_m$. Thus, if we let σ_f denote the common film stress resultant, we find that (2.19) can be expressed as

$$2\mathcal{H}r(\sqrt{1+r'^2})^3 = -rr'' + 1 + r'^2. \tag{2.19}$$

Equation (2.19) is the equation for a Delaunay surface with $\mathcal{H} = p_0/(2\sigma_f)$ (see [13, p. 107]). A Delaunay surface (see [10]) is a surface of revolution with constant mean curvature $\mathcal{H}$. Thus (2.19) should not have been a complete surprise. Any soap bubble, axisymmetric or otherwise, displays equal tension in all directions, and so the equation of normal equilibrium immediately leads to the condition that the mean curvature must be constant. For a sphere of radius R with constant pressure p_0, we find $\sigma_f = \frac{1}{2}p_0R$. If we set $w = 0$ and $\sigma_f = \frac{1}{2}Rp_0$, then we see $\theta(s) = \frac{1}{2}\pi - s/R, r(s) = R\cos\theta, z(s) = R - R\sin\theta$ satisfy (2.7)–(2.10), defining a sphere of radius R. A thorough discussion of Delaunay's surfaces can be found in [19, p. 115]. A number of interesting geometrical characterizations of Delaunay surfaces can be found in [12]. See also [18] for a related discussion on small mylar balloons.

3. Pumpkin shape.

3.1. Background. As a first approximation to a pumpkin-like shape, it is not uncommon to make ad hoc assumptions to simplify finding the shape (see, e.g., [25], [26]). For example, in a constant bulge radius approach the author in [21] assumes that a tendon follows the trace of the generating curve for a related SPNS balloon. If the number of gores is sufficiently high and the bulge radius is chosen appropriately, it is not unreasonable to suppose that the region between adjacent load tendons is spanned by circular arcs, forming a pumpkin-like balloon. In the constant bulge radius approach, one assumes that the tendon profiles are known. The tangent to the tendon profile is orthogonal to a plane containing a circle of radius r_B. The angular width of the circular arc is then determined. In the constant bulge shape approach, the angular width is assumed to be a constant value independent of s. Since it is straightforward to formulate the equations for a pumpkin gore, assuming a tubular surface geometry, we can avoid these ad hoc assumptions and obtain a model that is a better representation of the three dimensional pumpkin gore. We will discuss the merits and flaws of these various designs in section 2.3.1, but first we derive the ODEs for a pumpkin balloon.

3.2. Equilibrium equations for a pumpkin balloon. In this section, we derive the equilibrium equations for a pumpkin shape balloon as a membrane with the geometry of a tubular surface (see [17, Chapter VI] for a related discussion on unishells). While tubular surfaces with boundary are considered in other works (see, e.g., [2]), the balloon problem is somewhat simpler because we can ignore bending moments. We will assume that the pumpkin shape is made up of n_g symmetric pumpkin gores.

We begin with a curve, $\boldsymbol{\Upsilon}(s) = R(s)\mathbf{i} + Z(s)\mathbf{k} \in \mathbb{R}^3$, that we call the generator of the pumpkin gore. A priori, $\boldsymbol{\Upsilon} \in \mathbb{R}^3$ is unknown and must be derived from equilibrium conditions. The generator is parametrized by arc length s, i.e., $R'(s)^2 + Z'(s)^2 = 1$. Let $\mathbf{t}$ denote the unit tangent of $\boldsymbol{\Upsilon}$, $\mathbf{b}$ its inward unit normal ($\mathbf{j} = \mathbf{t} \times \mathbf{b}$), $\theta = \theta(s)$ is the angle between $\mathbf{t}$ and $\mathbf{k}$, and

$$\begin{aligned} \mathbf{t}(s) &= \sin\theta\mathbf{i} + \cos\theta\mathbf{k}, \\ \mathbf{b}(s) &= -\cos\theta\mathbf{i} + \sin\theta\mathbf{k}. \end{aligned}$$

Typically, $\theta(s)$ is decreasing in superpressure applications. The set $\{\mathbf{b}, \mathbf{t}, \mathbf{j}\}$ gives a right-hand curvilinear basis for $\mathbb{R}^3$. Since $\boldsymbol{\Upsilon}$ is a plane curve, its torsion is zero, and the Frenet equations reduce to

$$\begin{aligned} \mathbf{t}'(s) &= \kappa(s)\mathbf{b}(s), \\ \mathbf{b}'(s) &= -\kappa(s)\mathbf{t}(s), \end{aligned}$$

where κ is the curvature of $\boldsymbol{\Upsilon}$ (see [11, section 1.5]). We define a tubular surface in the following manner. Let

$$\mathbf{x}(s,v) = \boldsymbol{\Upsilon}(s) + r_B\left(-\mathbf{b}(s)\cos v + \mathbf{j}\sin v\right), \quad -\pi < v < \pi, \quad 0 < s < L_d, \tag{3.1}$$

and $x(s,v) = \mathbf{x}(s,v)\cdot\mathbf{i}$, $y(s,v) = \mathbf{x}(s,v)\cdot\mathbf{j}$, and $z(s,v) = \mathbf{x}(x,v)\cdot\mathbf{k}$.

In the following, we denote partial differentiation using subscript notation, e.g., $\mathbf{x}_s = \partial\mathbf{x}/\partial s$. By direct calculation, we have

$$\begin{aligned} \mathbf{x}_s(s,v) &= (1 + r_B\kappa(s)\cos v)\mathbf{t}(s), \\ \mathbf{x}_v(s,v) &= r_B\left(\mathbf{b}(s)\sin v + \mathbf{j}\cos v\right), \\ \mathbf{x}_s \times \mathbf{x}_v &= r_B(1 + r_B\kappa(s)\cos v)\left(\mathbf{b}(s)\cos v - \mathbf{j}\sin v\right), \end{aligned}$$

and $dA = r_B(1 + r_B\kappa\cos v)dsdv$. A unit vector normal to the tubular surface is

$$\mathbf{N}(s,v) = \frac{\mathbf{x}_s \times \mathbf{x}_v}{|\mathbf{x}_s \times \mathbf{x}_v|} = \mathbf{b}(s)\cos v - \mathbf{j}\sin v,$$

and the triple $\{\mathbf{x}_s, \mathbf{x}_v, \mathbf{N}\}$ gives a right-hand basis for $\mathbb{R}^3$.

By direct calculation, we have $\mathbf{N}_s(s,v) = -\kappa(s)\mathbf{t}(s)\cos v$ and $\mathbf{N}_v(s,v) = -\mathbf{b}(s)\sin v + \mathbf{j}\cos v$. The principal curvatures of the tubular surface are

$$\begin{aligned} \kappa_1(s,v) &= -\frac{\mathbf{N}_s\cdot\mathbf{x}_s}{\mathbf{x}_s\cdot\mathbf{x}_s} = \frac{\kappa\cos v}{1 + r_B\kappa\cos v}, \\ \kappa_2(s,v) &= -\frac{\mathbf{N}_v\cdot\mathbf{x}_v}{\mathbf{x}_v\cdot\mathbf{x}_v} = \frac{1}{r_B}. \end{aligned}$$

The tubular surface is sufficiently smooth if we assume (see [11, p. 399])

$$r_B\kappa_0 < 1, \quad \text{where} \quad \kappa_0 < \max_{0\le s\le L_d} |\kappa(s)|. \tag{3.2}$$

We find that the condition in (3.2) is met for our work on pumpkin balloons.

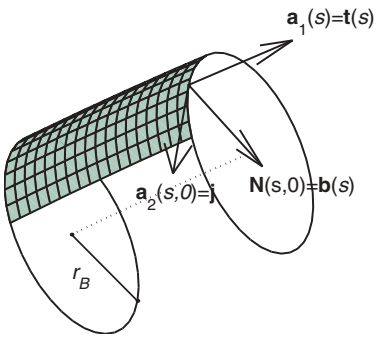


FIG. 3.1. *A patch of a pumpkin gore with* $\mathbf{a}_1, \mathbf{a}_2, \mathbf{N}$.

For the problems considered here, (3.2) is satisfied. A unit tangent to the curve $s \to \mathbf{x}(s, v)$ is

$$\mathbf{a}_1(s, v) = \frac{\mathbf{x}_s(s, v)}{|\mathbf{x}_s(x, v)|} = \mathbf{t}(s),$$

and a unit tangent to the curve $v \to \mathbf{x}(s, v)$ is

$$\mathbf{a}_2(s, v) = \frac{\mathbf{x}_v(s, v)}{|\mathbf{x}_v(s, v)|} = \mathbf{b}(s) \sin v + \mathbf{j} \cos v.$$

Note that $\partial \mathbf{a}_2 / \partial v = \mathbf{b}(s) \cos v - \mathbf{j} \sin v = \mathbf{N}$. In Figure 3.1, the geometry of a pumpkin gore patch is illustrated with the vectors $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{N}\}$ when $v = 0$. Arc length in the tubular surface along a curve parallel to the generator $s \to \mathbf{x}(s, v)$ is $\bar{s}$, where $d\bar{s} = (1 + r_B \kappa(s) \cos v) ds$. We formulate the equilibrium equations for the tubular surface using (s, v) as parameters. We assume the meridional stress resultant is independent of v; i.e., $\sigma_m(\bar{s}(s), v) = \sigma_m(s)$ and the circumferential stress resultant is $\sigma_c(\bar{s}(s), v) = \sigma_c(s)$. The quantities $\sigma_m(s)$ and $\sigma_c(s, v)$ are measured per unit length in the tubular surface. We define

$$\mathbf{n}_1(s, v) = \sigma_m(s) \mathbf{a}_1(s),$$
$$\mathbf{n}_2(s, v) = \sigma_c(s) \mathbf{a}_2(s, v).$$

A pumpkin gore will be a subset of a tubular surface. We assume that the pumpkin gore is situated symmetrically with respect to the xz plane and interior to the wedge defined by the half-planes $y = \pm \tan(\pi / n_g) x$ with $x \geq 0$. We will refer to r_B as the *bulge radius* of the pumpkin gore. The curve traced by $v \to \boldsymbol{\Upsilon}(s) + r_B(-\mathbf{b}(s) \cos v + \mathbf{j} \sin v)$ is a circle lying in the plane with normal $\mathbf{t}(s)$. To find the length of the segment of the circle that forms a circumferential arc of the pumpkin gore, we need to find the values of v where this arc intersects the planes $y = \pm \tan(\pi / n_g) x$. For fixed s, we find that v must satisfy the condition

$$y(s, v) = \tan\left(\frac{\pi}{n_g}\right) \; x(s, v).$$

This leads us to the equation

$$A(s) + B(s) \cos v + C \sin v = 0, \tag{3.3}$$

where $A(s) = -R(s)\tan(\pi/n_g)$, $B(s) = -r_B\cos\theta(s)\tan(\pi/n_g)$, $C = r_B$. Now (3.3) can be solved for v, yielding

$$v = v_B = \arccos\left(\frac{-AB + C\sqrt{C^2 + B^2 - A^2}}{B^2 + C^2}\right).$$

Since A and B are functions of s and other parameters, so is $v_B = v_B(s, n_g, R(s), \theta(s))$. The parameter dependence will be clear from context, and so we write v_B or $v_B(s)$ for convenience. By symmetry, the solution corresponding to the plane $y = -\tan(\pi/n_g)x$ is $v = -v_B$. We define the theoretical three dimensional pumpkin gore $\mathcal{G}$ to be the set

$$\mathcal{G} = \{\mathbf{x}(s, v), \quad -v_B(s) < v < v_B(s), \quad 0 < s < L_d\}.$$

A complete shape $\mathcal{S}$ has cyclic symmetry and is made up of n_g copies of $\mathcal{G}$. Although $\mathcal{G}$ is not axisymmetric, in practice n_g is large, and so we ignore shear stress resultants in our formulation. A priori, the generating curve $\boldsymbol{\Upsilon}(s)$ for the pumpkin gore is unknown. To determine $\boldsymbol{\Upsilon}(s)$, we must first consider the equilibrium equations for an arbitrary patch on the tubular surface. From these, we are led to a system of ODEs that can be solved to determine $\boldsymbol{\Upsilon}$ and then $\mathcal{G}$.

We begin by deriving equilibrium equations for a patch, $\mathcal{A} = \{\mathbf{x}(\xi, v) \mid s_0 < \xi < s, v_0 < v < v_1\}$, where $-v_B(\xi) < v_0 < v < v_1 < v_B(\xi)$ for $s_0 < \xi < s$. Balancing forces acting on $\mathcal{A}$, we find

$$\begin{aligned}\vec{0} = &\int_{v_0}^{v_1} \sigma_m(s)\mathbf{a}_1(s) r_B\, dv - \int_{v_0}^{v_1} \sigma_m(s_0)\mathbf{a}_1(s_0) r_B\, dv \\ &+ \int_{s_0}^{s} \sigma_c(\xi)\mathbf{a}_2(s, v_1)(1 + r_B\kappa\cos v_1)\, d\xi - \int_{s_0}^{s} \sigma_c(\xi)\mathbf{a}_2(s, v_0)(1 + r_B\kappa\cos(v_0))\, d\xi \\ &+ \int_{\mathcal{A}} \mathbf{f}(\xi, \psi) r_B(1 + r_B\kappa(\xi)\cos\psi)\, d\psi\, d\xi, \end{aligned} \tag{3.4}$$

where $\mathbf{f}$ is the external force. Thus, we can rewrite (3.4) in the form

$$\begin{aligned}\vec{0} = \int_{s_0}^{s}\int_{v_0}^{v_1} \Big\{ &\frac{\partial}{\partial s}\left[\sigma_m(\xi)\mathbf{a}_1(\xi) r_B\right] + \frac{\partial}{\partial \psi}\left[\sigma_c(\xi)\mathbf{a}_2(\xi, \psi)(1 + r_B\kappa(\xi)\cos\psi)\right] \\ &+ \mathbf{f}(\xi, \psi) r_B(1 + r_B\kappa(\xi)\cos\psi)\Big\} d\psi\, d\xi, \end{aligned} \tag{3.5}$$

where

$$\begin{aligned}&\mathbf{f}(s, v) = -p(z(s, v))\mathbf{N}(s, v) - w(s)\mathbf{k}, \\ &z(s, v) = Z(s) - r_B\sin\theta\cos v.\end{aligned}$$

By differentiating (3.5), one is led to a system of partial differential equations that determines the pumpkin gore. This system is complicated, and since n_g is typically large, it is tempting to use a small angle approximation to justify dropping terms involving $2\pi/n_g$ to obtain a simpler system. However, we can eliminate the dependence on v by integrating and retain terms that influence the design profile. In particular, if we set $v_0 = -v_B(s)$ and $v_1 = v_B(s)$ in (3.5) and integrate with respect to ψ,

we obtain

$$\begin{aligned}\vec{0} = \int_{s_0}^{s} \Bigg\{ & \frac{\partial}{\partial s}\left(\sigma_m(\xi)\mathbf{a}_1(\xi)2r_B v_B(\xi)\right) \\ & + \sigma_c(\xi)\left(\mathbf{a}_2(\xi, v_B(\xi)) - \mathbf{a}_2(\xi, -v_B(\xi))\right)\left(1 + r_B\kappa(\xi)\cos v_B\right) \\ & + \int_{-v_B(\xi)}^{v_B(\xi)} \left(-p(z(\xi,\psi))\mathbf{N}(\xi,\psi) - w(\xi)\mathbf{k}\right) r_B(1 + r_B\kappa(\xi)\cos\psi)d\psi \Bigg\} d\xi. \end{aligned} \tag{3.6}$$

To add the tendon weight density to (3.6), we note that $d\bar{s} = (1+r_B\kappa\cos v_B(s))ds$ is an arc-length measure along the tendon. The weight of a segment $d\bar{s}$ is $w_t d\bar{s}$, and (3.6) becomes

$$\begin{aligned}\vec{0} = \int_{s_0}^{s} \Bigg\{ & \frac{\partial}{\partial s}\left(\sigma_m(\xi)\mathbf{a}_1(\xi)2r_B v_B(\xi)\right) \\ & + \sigma_c(\xi)\left(\mathbf{a}_2(\xi, v_B(\xi)) - \mathbf{a}_2(\xi, -v_B(\xi))\right)\left(1 + r_B\kappa(\xi)\cos v_B(\xi)\right) \\ & - w_t\mathbf{k}(1 + r_B\kappa(\xi)\cos v_B(\xi)) \\ & + \int_{-v_B(\xi)}^{v_B(\xi)} \left(-p(z(\xi,\psi))\mathbf{N}(\xi,\psi) - w(\xi)\mathbf{k}\right) r_B(1 + r_B\kappa(\xi)\cos\psi)d\psi \Bigg\} d\xi. \end{aligned} \tag{3.7}$$

In balloon design, it is not uncommon to make some simplifying assumptions about the form of σ_c. In the axisymmetric natural shape models, one assumes that $\sigma_c = 0$. In the Delaunay surfaces, one assumes that $\sigma_c = \sigma_m$ is constant. The assumption $\sigma_c(\xi) = \sigma_m(\xi)$ could be handled without difficulty, but for convenience of exposition we will assume that σ_c is a nonnegative constant. Using the properties of $\mathbf{a}_2$ and assuming $\sigma_c(\xi, v) = \sigma_c$, we find $\mathbf{a}_2(\xi, v) - \mathbf{a}_2(\xi, -v) = 2\mathbf{b}(\xi)\sin v$ and

$$\sigma_c(\xi)(\mathbf{a}_2(\xi, v_B(\xi)) - \mathbf{a}_2(\xi, -v_B(\xi))) = 2\sigma_c\mathbf{b}(\xi)\sin v_B(\xi). \tag{3.8}$$

Using (3.7) and (3.8), we obtain the following:

$$\begin{aligned}\vec{0} = \int_{s_0}^{s} \Bigg\{ & \frac{\partial}{\partial s}\left(\sigma_m(\xi)\mathbf{a}_1(\xi)2r_B v_B(\xi)\right) \\ & + \left[2\sigma_c\mathbf{b}(\xi)\sin v_B(\xi) - w_t\mathbf{k}\right]\left(1 + r_B\kappa(\xi)\cos v_B(\xi)\right) \\ & + \int_{-v_B(\xi)}^{v_B(\xi)} \left(-p(z(\xi,\psi))\mathbf{N}(\xi,\psi) - w(\xi)\mathbf{k}\right) r_B(1 + r_B\kappa(\xi)\cos\psi)d\psi \Bigg\} d\xi. \end{aligned} \tag{3.9}$$

From the last term in (3.9), we find

$$-\int_{-v_B(\xi)}^{v_B(\xi)} p\mathbf{N}(\xi,\psi)\; r_B(1 + r_B\kappa\cos\psi)d\psi = -\mathbf{b}(\xi)\left(\kappa Q_1 + Q_0\right),$$

where

$$\begin{aligned} Q_1(b, p_0, r_B, v_B, Z, \theta) = {} & \frac{1}{12}r_B^2\big(12p_0 v_B + 12bv_B Z - br_B\cos(\theta - 3v_B) \\ & - 9br_B\cos(\theta - v_B) + 9br_B\cos(\theta + v_B) + br_B\cos(\theta + 3v_B) \\ & + 6p_0\sin(2v_B) + 6bZ\sin(2v_B)\big), \\ Q_0(b, p_0, r_B, v_B, Z, \theta) = {} & \frac{1}{4}r_B\big(-br_B\cos(\theta - 2v_B) + br_B\cos(\theta + 2v_B) \\ & - 4br_B v_B\sin\theta + 8p_0\sin v_B + 8bZ\sin v_B\big). \end{aligned}$$

Equation (3.9) can be expressed in the form

$$\begin{aligned}\vec{0} = \int_{s_0}^{s} \Big\{ & \frac{\partial}{\partial \xi} \left(2 v_B(\xi) r_B \sigma_m(\xi) \mathbf{t}(\xi)\right) \\ & + \mathbf{b}(\xi) \left\{2\sigma_c \sin v_B (1 + \kappa r_B \cos v_B) - \kappa Q_1 - Q_0\right\} \\ & - 2\mathbf{k} r_B w_f \left\{v_B + r_B \kappa \sin v_B\right\} - \mathbf{k} w_t (1 + r_B \kappa \cos v_B) \Big\} \, d\xi. \end{aligned} \tag{3.10}$$

Differentiating (3.10) with respect to s and simplifying, we obtain

$$\begin{aligned}\vec{0} = & \left(\sigma_m(s) 2 r_B v_B(s)\right)' \mathbf{t}(s) + \left(\sigma_m(s) 2 r_B v_B(s)\right) \left(-\theta'(s)\right) \mathbf{b}(s) \\ & + \mathbf{b}(s) \left\{2\sigma_c \sin v_B (1 + \kappa r_B \cos v_B) - \kappa Q_1 - Q_0\right\} \\ & - 2\mathbf{k} r_B w_f \left\{v_B + r_B \kappa \sin v_B\right\} - \mathbf{k} w_t (1 + r_B \kappa \cos v_B). \end{aligned} \tag{3.11}$$

The inner product of (3.11) with $\mathbf{j}$ is zero. Taking the inner product of (3.11) with $\mathbf{t}$ and $\mathbf{b}$, we obtain the respective equations

$$\left(\sigma_m(s) 2 r_B v_B(s)\right)' = 2 r_B w_f \left(v_B + r_B \kappa \sin v_B\right) \cos\theta + w_t (1 + r_B \kappa \cos v_B) \cos\theta$$

and

$$\begin{aligned} 0 = & -\sigma_m 2 r_B v_B \theta'(s) + 2\sigma_c \sin v_B (1 + \kappa r_B \cos v_B) - \kappa Q_1 - Q_0 \\ & - 2 r_B w_f \left(v_B + r_B \kappa \sin v_B\right) \sin\theta - w_t (1 + r_B \kappa \cos v_B) \sin\theta. \end{aligned}$$

Using the definition $\kappa = -\theta'(s)$, we find

$$\begin{aligned} 0 = \theta'(s) & \left\{-\sigma_m 2 r_B v_B + Q_1 + 2 r_B^2 w_f \sin v_B \sin\theta - \sigma_c r_B \sin(2 v_B) + w_t r_B \cos v_B \sin\theta\right\} \\ & + 2\sigma_c \sin v_B - Q_0 - 2 r_B w_f v_B \sin\theta - w_t \sin\theta. \end{aligned}$$

Defining the total meridional tension per gore,

$$T_m(s) = 2 r_B v_B(s) \sigma_m(s),$$

and replacing κ with $-\theta'(s)$, we find

$$\cos\theta \left(2 r_B^2 w_f \sin v_B + w_t r_B \cos v_B\right) \theta'(s) + T_m'(s) = 2 r_B v_B w_f \cos\theta + w_t \cos\theta,$$

$$\begin{aligned} & \left(T_m(s) - Q_1 - 2 r_B^2 w_f \sin v_B \sin\theta - w_t r_B \cos v_B \sin\theta + \sigma_c r_B \sin(2 v_B)\right) \theta'(s) \\ & \qquad = 2(\sigma_c \sin v_B - \tfrac{1}{2} Q_0 - r_B w_f v_B \sin\theta) - w_t \sin\theta. \end{aligned}$$

We are finally led to a system of ODEs in s that the generating curve for the pumpkin gore must satisfy. Setting $w^*(s) = 2 r_B v_B(s) w_f$ and

$$\hat{\kappa}(\theta, T_m, Z, Q_0, Q_1, r_B, v_B, \sigma_c, w^*, w_t)$$

$$= -\frac{2\sigma_c \sin v_B - Q_0 - (w^* + w_t) \sin\theta}{T_m - Q_1 - r_B (w^* \sin v_B / v_B + w_t \cos v_B) \sin\theta + \sigma_c r_B \sin(2 v_B)},$$

we obtain

$$\theta'(s) = -\hat{\kappa}(\theta, T_m, Z, Q_0, Q_1, r_B, v_B, \sigma_c, w^*, w_t), \tag{3.12}$$

$$T_m'(s) = \left(w^* \left(1 + \frac{r_B \hat{\kappa} \sin v_B}{v_B}\right) + w_t (1 + r_B \hat{\kappa} \cos v_B)\right) \cos\theta, \tag{3.13}$$

$$R'(s) = \sin\theta, \tag{3.14}$$

$$Z'(s) = \cos\theta. \tag{3.15}$$

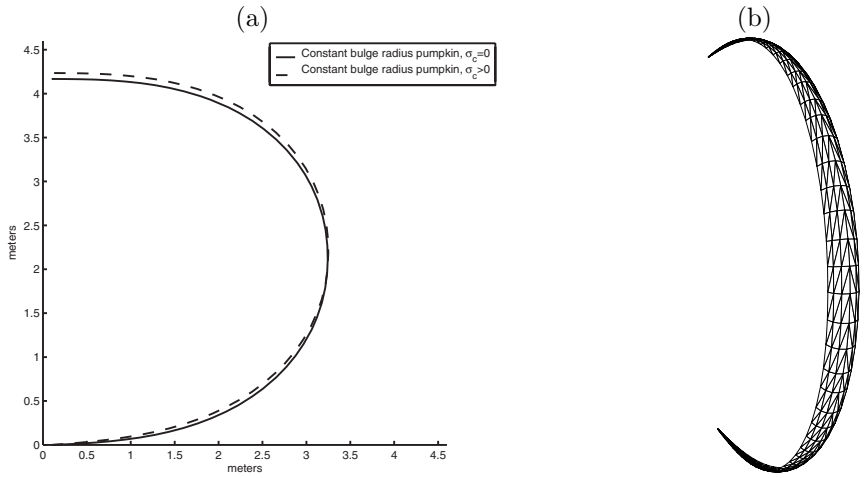


FIG. 3.2. (a) *Tendon profiles for pumpkin balloon models* $\sigma_c = 0$ *N/m and* $\sigma_c = 10$ *N/m;* $p_0 = 200$ *Pa and* $r_B = 0.1066$ *m.* (b) *Three dimensional pumpkin gore.*

The ODEs (3.12)–(3.15) are supplemented by initial conditions

$$\begin{aligned} \theta(0) &= \theta_0, \\ T_m(0) &= \frac{(L_0 + \pi R_0^2 p_0)}{(n_g \cos\theta_0)}, \\ R(0) + r_B \cos\theta(0) &= R_0, \\ z(0) &= Z_0, \end{aligned} \tag{3.16}$$

and auxiliary equations

$$\begin{aligned} n_g \cos\theta_1 T_m(\ell) &= L_1 - \pi R_1^2 p_1, \\ R(\ell) + r_B \cos\theta(\ell) &= R_1. \end{aligned} \tag{3.17}$$

Equations (3.12)–(3.17) are solved via a shooting method, and the generating curve $\boldsymbol{\Upsilon}$ is determined. In Figure 3.2(a), we present the tendon profiles for pumpkin balloons with $\sigma_c = 0$ and $\sigma_c > 0$. The models will be discussed in more detail in the next section. In Figure 3.2(b), we present a discretization of the theoretical pumpkin gore.

As a check that Archimedes' Principle is satisfied at float, we take the dot product of (3.9) with $\mathbf{k}$, multiply by n_g, and integrate from $s = 0$ to $s = L_d$; after simplifying the result and using

$$2n_g r_B v_B(L_d)\sigma_m(L_d)\cos\theta(L_d) - 2n_g r_B v_B(0)\sigma_m(0)\cos\theta_d + W_c + W_{tapes} + W_{film} - b\mathcal{V} = 0,$$

$$W_c = 2\int_0^{L_d} \sigma_c(\xi)\mathbf{b}(\xi)\cdot\mathbf{k}\sin v_B(\xi)(1+\kappa(\xi)\cos v_B(\xi))d\xi,$$

we find

$$L_0 + L_1 - \pi(p_1 R_1^2 - p_0 R_0^2) + W_{tapes} + W_c + W_{film} = bV. \tag{3.18}$$

Note, if $\sigma_c \neq 0$, that the boundary conditions along the edge of the gore contribute a net force W_c in the $\mathbf{k}$ direction, which is reflected in (3.18). For the case considered in Table 2.1(b), we find that W_c is approximately 2 N.

4. Numerical results. In this section, we present numerical results for solutions of the respective shape-finding problems for balloon models discussed in previous sections. The design parameters used in our demonstration cases are those given in Table 1.1. For each balloon type, we calculated the corresponding shape (ZPNS, SPNS, and pumpkin) that would maintain a payload of 280 N at an altitude of 10 km. Each balloon was constructed with 96 gores and no caps. We use a 1.5 mil polyethylene skin for our designs. Relevant values (height, diameter, volume, etc.) of the specific designs are presented in Table 2.1(a) for axisymmetric shapes and in Table 2.1(b) for pumpkin shapes.

Normally, load tapes are used in the construction of natural shape balloons, and (heavier) load tendons are used in the construction of pumpkin balloons. This distinction is important to the balloon designer. For example, since a load tape can contain as much as 3% slackness and is less stiff than a load tendon, the balloon film will bear a significant portion the weight of the balloon system in certain regions. In a pumpkin balloon, the balloon film functions primarily as a gas barrier, locally transmitting loads to the tendons. Because we are not concerned with straining of the balloon film here, we use the same weight densities for tendons and load tapes.

While the Euler-elastica model utilizes some crude assumptions, it does have the advantage of a closed form solution. We note that the maximum radius for an Euler elastica profile is given by $1/\sqrt{\tau}$ (see (2.17)). For the Euler-elastica shape related to the designs presented in Figure 3.2(a) with $p_0 = 200$ Pa, we find $T_0 = (274\text{ N})/\cos(1.526) = 6.134$ kN and $\tau = \pi p_0/T_0 = 0.1024\text{ m}^{-2}$, and the maximum radius is $\sqrt{1/\tau} = 3.1246$ m. The maximal radius in the SPNS model is 3.0267 m. The elastica model gives a 3.0% relative error.

It is important to keep in mind that in reality, the shape-finding process as described in sections 2–3 is only one part of the process in the construction of a balloon. In a ZPNS, once (θ_d, ℓ_d) are determined and $r(s)$ and $z(s)$ are known, the theoretical natural shape gore is given by

$$\mathcal{N} = \{(x, y, z) = \mathbf{x}(s, \phi), 0 < s < \ell_d, -\pi/n_g < \phi < \pi/n_g\}.$$

In reality, the true gore $\mathcal{N}$ is approximated by a developable surface,

$$\hat{\mathcal{N}} = \left\{(x, y, z) = (r(s), y, z(s)), 0 < s < \ell_d, -r(s)\tan\left(\frac{\pi}{n_g}\right) < y < r(s)\tan\left(\frac{\pi}{n_g}\right)\right\}.$$

The developable surface $\hat{\mathcal{N}}$ can be rolled into a plane without straining, thus defining the lay-flat gore pattern to be cut. The complete balloon is then constructed by seaming together n_g individual gores. While a circumferential fiber in $\mathcal{N}$ has a length of $2\pi r(s)/n_g$, when n_g is large $\tan(\pi/n_g) \approx \pi/n_g$ and we see that $\hat{\mathcal{N}}$ is a good approximation of $\mathcal{N}$. For a large balloon, r is comparatively large (away from the end-caps), and we see that the curvature κ of the curve $s \to (r(s), 0, z(s))$ is small. The other principal curvature is $-\cos\theta(s)/r(s)$, and so, away from the end-caps, it is also small because r is large. For superpressure balloons, θ is near $\pm\pi/2$ in the vicinity of the end-caps. ZPNS balloons flown by NASA typically have between 100 and 200 gores. When a real ZPNS balloon is fully pressurized, the balloon will assume a shape that is very nearly a surface of revolution.

The curvature of the pumpkin gore generator $\boldsymbol{\Upsilon}$ is of the same order of magnitude as the curvature of the generator of the corresponding SPNS. However, the curvature in the hoop direction is r_B^{-1} and in general is not small. If s_c is arc-length measured

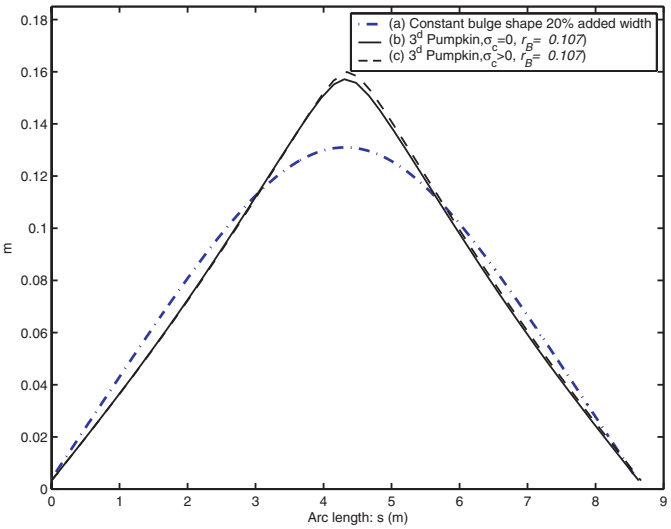


FIG. 4.1. *Normalized flat half-gore patterns.*

down the center of the pumpkin gore, then $ds_c = (1 + r_B\kappa(s))ds$, and we define

$$\tilde{\mathcal{G}} = \{(u,t) \mid t = s_c(s), -r_B v_B(s) < u < r_B v_B(s), 0 < s < L_d\}.$$

While $\tilde{\mathcal{G}}$ can be thought of as a flat panel approximation to $\mathcal{G}$, even when fully pressurized, the balloon will have difficulty achieving the desired doubly curved shape. The length of an edge of $\tilde{\mathcal{G}}$ is much longer than the length of the corresponding theoretical gore $\mathcal{G}$. To achieve the desired shape in a real balloon, the material along the edge of $\tilde{\mathcal{G}}$ is gathered, and a tendon of length $L_t = \int_0^{L_d} ds_t$ is attached. Note that the gathering is nonuniform. (It is maximal near the equatorial latitudes and decreases as one approaches the end-caps.) In any case, there will be wrinkling in the inflated pumpkin balloon. If flat warped panels of film are used in the construction of a pumpkin gore, the amount of excess material and wrinkling can be reduced (see [5]).

In Figure 4.1, we present three flat gore patterns: a constant bulge shape design, a constant bulge radius design with $\sigma_c = 0$, and a constant bulge radius design with $\sigma_c = 10$ N/m. The first design is based on the constant bulge shape approach, as was utilized in the *Endeavour* design. In the constant bulge shape approach, it is assumed that the load tendons followed the trace of the Euler-elastica curve and the region between tendons is spanned by circular arcs, assuming a constant angular width. In this approach, a node on the tendon in the $\phi = -\pi/n_g$ plane is identified with the corresponding node in the $\phi = \pi/n_g$ plane. The chord length $2y(s)$ is replaced by a circular arc of length $1.2 \times 2y(s)$ (20% "added width"). It is clear from Figure 4.1 that the constant bulge shape approach has significantly more material away from the equatorial region than $\tilde{\mathcal{G}}$ for a comparable pumpkin gore. The *Endeavor* was built with 64 constant bulge shape gores and, when fully pressurized, it assumed a severely distorted configuration that did not resemble a pumpkin shape. In hindsight, this should not have been a complete surprise, since the *Endeavor* gore pattern did not follow from equilibrium equations of the fully inflated shape. Only when four gores were removed was the *Endeavor* able to deploy into a shape one might characterize

as pumpkin-like (see [16]).

5. Concluding remarks. In this paper, we reviewed a number of axisymmetric balloon models including the zero-pressure natural shape balloon, the superpressure natural shape balloon, the Euler-elastica, and the sphere. Based on a tubular surface membrane theory, we developed a model for a three dimensional pumpkin gore. The pumpkin model that we present is an improvement over previous models in that it introduces a circumferential stress parameter into the model and leads to a more accurate representation of the three dimensional gore. We also discussed some of the practical issues that the manufacturer must deal with when trying to construct large doubly curved gores from flat sheets of balloon film.

Acknowledgment. The author would like to thank the referees for their comments.

REFERENCES

[1] Anon., *Research Development in the Field of High Altitude Plastic Balloons*, NONR–710(01a) Reports, Department of Physics, University of Minnesota, Minneapolis, 1951–1956.
[2] E. L. Axelrad and F. A. Emmerling, *Elastic tubes*, Appl. Mech. Rev., 37 (1984), pp. 891–897.
[3] F. Baginski and W. Schur, *Structural analysis of pneumatic envelopes: A variational formulation and optimization-based solution process*, AIAA J., 41 (2003), pp. 304–311.
[4] F. Baginski and J. Winker, *The natural shape balloon and related models*, Adv. Space Res., 33 (2004), pp. 1617–1622.
[5] F. Baginski and W. Schur, *Design Issues for Large Scientific Balloons*, in Proceedings of the 3rd Annual Aviation Technology, Integration, and Operations (ATIO) Technical Forum, Denver, CO, 2003, paper AIAA-2003-6788.
[6] F. Baginski and W. Collier, *Modeling the shapes of constrained partially inflated high altitude balloons*, AIAA J., 39 (2001), pp. 1–11.
[7] F. Baginski, Q. Chen, and I. Waldman, *Designing the shape of a large scientific balloon*, Appl. Math. Modeling, 25 (2001), pp. 953–966.
[8] F. Baginski, W. Collier, and T. Williams, *A parallel shooting method for determining the natural shape of a large scientific balloon*, SIAM J. Appl. Math., 58 (1998), pp. 961–974.
[9] W. Collier, *Estimating stresses in a partially inflated high altitude balloon using a relaxed energy*, Quart. Appl. Math., 61 (2003), pp. 17–40.
[10] C. Delaunay, *Sur la Surface de Revolution dont la Courbure Moyenne est Constante*, J. Math. Pures Appl., 6 (1841), pp. 309–320.
[11] M. P. do Carmo, *Differential Geometry of Curves and Surfaces*, Prentice–Hall, Englewood Cliffs, NJ, 1976.
[12] H. Eells, *The surfaces of Delaunay*, Math. Intell., 9 (1987), pp. 53–57.
[13] C. Isenberg, *The Science of Soap Films and Soap Bubbles*, Dover, New York, 1992.
[14] H. B. Keller, *Numerical Two-Point Boundary-Value Problems*, Dover, New York, 1982.
[15] A. L. Morris, ed., *Scientific Ballooning Handbook*, Technical report NCAR-TN-99, National Center for Atmospheric Research, Boulder, CO, 1975.
[16] B. A. Lennon and S. Pellegrino, *Stability of lobed inflatable structures*, AIAA-2000-1728, Proceedings of the 41st AIAA/ASME/ASCE/AHS/ASC SDM Conference, Atlanta, GA, 2000.
[17] A. Libai and J. G. Simmonds, *The Nonlinear Theory of Elastic Shells*, 2nd ed., Cambridge University Press, Cambridge, UK, 1998.
[18] I. Mladenov and J. Oprea, *The mylar balloon revisited*, MAA Monthly, 110 (2003), pp. 761–784.
[19] J. Oprea, *Differential Geometry*, Prentice–Hall, Upper Saddle River, NJ, 1997.
[20] M. Rougeron, *Up to date CNES Balloon Study*, in Proceedings of the 10th AFCRL Scientific Balloon Symposium, 19783, paper AFCRL-TR-79-005.
[21] W. Schur, *Analysis of load tape constrained pneumatics envelopes*, Proceedings of the 40th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference and Exhibit, St. Louis, MO, 1999, Collection of Technical Papers, Vol. 4 (A99-24601 05-39).

[22] J. H. Smalley, *Determination of the shape of a free balloon*, Theoretical Development, Technical report AFCRL-65-68, Air Force Cambridge Research Labs, 1964.

[23] J. H. Smalley, *Determination of the shape of a free balloon—Balloons with superpressure, subpressure and circumferential stress and capped balloons*, Technical report AFCRL-65-72, Air Force Cambridge Research Labs, 1964.

[24] J. H. Smalley, *Balloon shapes and stresses below the design altitude*, Technical report NCAR-TN-25, Minneapolis, Minnesota, 1966.

[25] J. H. Smalley, *Development of the e-balloon*, Technical report AFCRL-70-0543, National Center for Atmospheric Research, Boulder, CO, 1970.

[26] N. Yajima, *A survey of balloon design problems and prospects for large super-pressure balloons in the next century*, Adv. Space Res., 30 (2002), pp. 1183–1192.

SIAM J. APPL. MATH.
Vol. 65, No. 3, pp. 858–881

MODELING AND DESIGN OF COATED STENTS TO OPTIMIZE THE EFFECT OF THE DOSE*

MICHEL C. DELFOUR†, ANDRÉ GARON‡, AND VITO LONGO§

Abstract. Stents are used in interventional cardiology to keep a diseased vessel open. New stents are coated with a medicinal agent to prevent early reclosure due to the proliferation of smooth muscle cells. It is recognized that it is the dose of the agent that effectively controls the cells in the wall of the vessel. This paper focuses on the effect of the number of struts and the ratio between the coated area of the struts and the targeted area of the vessel on the design problem under set therapeutic bounds on the dose. It introduces mathematical models of the dose for a zero-thickness periodic stent and an asymptotic stent that will play a central role in our analysis. Theoretical and numerical results are presented along with their impact on the design process.

Key words. stenosis, restenosis, atherosclerosis, bioactive material, dose, modeling, design, numerical simulation, coated stent, interventional cardiology, medical applications, asymptotic behavior of solutions, Neumann sieve, shape optimization, sensitivity analysis

AMS subject classifications. 92C50, 35B27, 35B40

DOI. 10.1137/S0036139902411600

1. Introduction.

The use of stents (see Figure 1.1) in *interventional cardiology* is fairly recent.[1]

> An electrical engineer, Wiktor underwent open heart surgery to correct an aortic dissection in 1984. Following the procedure, he wondered why such a vascular repair couldn't be done with less surgical trauma, and began to read about angioplasty. He came up with a variety of stent designs and signed a consulting agreement with Medtronic in 1988.
>
> The "Wiktor Stent," an intravascular stent (U.S. patent No. 4,886,062) provides an important solution to coronary artery reconstruction and recanalization. The stent keeps a diseased vessel open and prevents reclosure. Made of tantalum, a noncorrosive and malleable metal which is easily seen by the cardiologist during fluoroscopy, the stent is extremely easy to handle, deliver and deploy, which is of the utmost importance in emergency and routine situations.
>
> In the case of the Wiktor Stent, the delivering catheter is inflated to expand and deploy the stent to maintain the opening. The balloon is then deflated and the catheter removed. Within a month, the stent becomes incorporated into the artery wall. Today, Medtronics' Wiktor stent has a 20 percent market share in Europe.

*Received by the editors July 18, 2002; accepted for publication (in revised form) April 21, 2004; published electronically February 25, 2005. This research was supported by National Sciences and Engineering Research Council of Canada discovery grants and by an FQRNT grant from the Ministère de l'Éducation du Québec.
http://www.siam.org/journals/siap/65-3/41160.html

†Centre de recherches mathématiques et Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal, QC, Canada H3C 3J7 (delfour@crm.umontreal.ca).

‡Département de Génie mécanique, École Polytechnique de Montréal, C.P. 6128, succ. Centre-ville, Montréal, QC, Canada H3C 3J7 (andre.garon@polymtl.ca).

§Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal, QC, Canada H3C 3J7 (longo@dms.umontreal.ca).

[1]New Jersey Hall of Fame: Dominik M. Wiktor (Bellcore, Morristown), Inventor of the Year, 1996 (http://www.njinvent.njit.edu/).

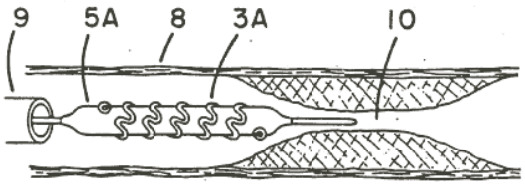


FIG. 1.1. *Wiktor stent as drawn in U.S. patent No.* 4,886,062.

More than 40 percent of patients treated for atherosclerosis present restenosis within six months of the operation. Indeed, implanting a stent generally leads to complications, thrombosis and proliferation of smooth-muscle cells being the root causes of restenosis. It is important to understand that stenosis and restenosis develop by different mechanisms and over very different time scales: 30 years for stenosis, 6 months for restenosis. A great deal of research has been conducted on medicinal agents designed to have an antithrombotic effect or to limit the proliferation of smooth-muscle cells. This type of medication can be delivered in two ways: systematically or locally. The primary disadvantage of systemic delivery is generally its greater toxicity in the body, since, to be effective, a much higher level of the medication is a priori required than for local delivery. For this reason, the local solution is the preferred choice. For instance, a typical system consists of a stent coated with a thin layer of polymer which has been impregnated with a molecule that has an effect on the proliferation of smooth muscle cells [1, 9, 10].

Modern stents are designed to be less invasive from the hemodynamic point of view. Engineers are introducing new shapes with minimal strut diameters to minimize induced perturbations that would theoretically increase stresses on the fluid and recirculation. Yet it is recognized that this strategy alone cannot prevent restenosis and considerable efforts are currently aimed at developing new stent coating to reduce or stop the proliferation of smooth muscle cells. However, to the authors' knowledge, the distribution of the coating (mass and contact surface) are not part of the design and the optimal selection of the shape parameters of the stent (coating is added to an already mechanically designed stent). An optimal coated stent (that would produce minimal restenosis) would be the result of a shape optimization that would simultaneously account for the hemodynamic and mass transfer processes.

It is the purpose of this paper to broaden the knowledge and improve the craft of the stent designer by providing new mathematical and numerical tools to help in the preliminary shape selection of stents. Mechanically the stent has to be strong enough to exert pressure on the wall of the vessel sufficient to keep it open and to restore a normal flow of blood. Once the purely mechanical parameters are set (choice of the material, contact surface, and thickness of the struts), there are a number of parameters left to control the delivery of the molecule to the wall: the total mass of the molecule and the distribution of the contact surface.

It is generally accepted that it is the effect of the concentration of the molecule over time that effectively controls the proliferation of the smooth muscle cells (cf. [2, 9, 10]). This naturally leads to the notion of a *dose* in each point of the artery. Mathematically the dose is defined as the integral of the concentration over all times ranging from 0 to ∞. The first key finding of this paper is the high sensitivity of the dose to the geometric distribution of the contact surface. This leads to the introduction of an *asymptotic stent* that describes the limit behavior of a family of

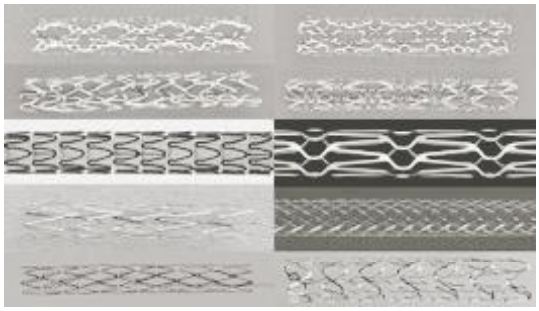


FIG. 1.2. *Typical patterns of stents.*

periodic stents as the number of struts increases, while maintaining a constant contact surface (and hence a constant distributed force on the wall to keep the vessel open). Given a bioactive material or molecule and the appropriate therapeutic bounds on the dose (a lower one to control the smooth muscle cells and an upper one to limit the toxicity of the molecule), the asymptotic stent provides a first estimate of the minimum total amount of bioactive material required. The second key finding is the behavior of the dose as the number of struts increases: It is almost periodic, and its amplitude decreases as the number of struts increases. It is then possible to select a realistic number of struts to achieve the design objectives and deliver the proper amount of bioactive material to the wall while respecting the upper therapeutic bound. Obviously this design would be followed by three-dimensional simulations to confirm the complete behavior of the device.

In this paper we have limited our investigation to the first two issues since they steer the overall process and contribute to limiting the number of three-dimensional simulations, thus reducing the cost of the design. For that purpose, we use a crude yet operational model of the lumen and the wall of the vessel: two concentric cylinders long enough not to affect the design in the *target area* where the stent will be deployed. For simplicity, the stent is chosen to be periodic and made up of a finite number of rings of radius R with no thickness and uniform width. However, our modeling and our analysis (in sections 4 and 5) and their consequences readily extend to more complex periodic stent structures (in section 5.3) of the type shown in Figure 1.2. In its undeployed state we assume that all the struts of the stent are side by side without space between them. When the stent is deployed to fill the target area, space is created between the struts. We call ρ the ratio between the contact surface occupied by the struts and the surface of the target area; it is a number between 0 and 1. When the struts are periodically spaced, $1 - \rho$ is a measure of the percentage of the area of the interface lumen/wall in the target area which is open to chemical/biological exchanges. The impact of such exchanges on the stent has not yet been fully investigated in the literature.

In our analysis we assume that the mass per unit area of the molecule impregnated in the thin layer on the stent is constant. We study the distribution of the dose in the artery with respect to the number of struts and the ratio ρ while keeping the surface of the target area constant. For a fixed ρ, as the number of struts increases, their width and the space between two adjacent struts decrease. So we are naturally led to introduce the notion of an *asymptotic stent*. We obtain the equations for the dose of the *asymptotic stent* and study the distribution of the dose in the wall of

the artery. One of the technical difficulties is to determine the right *transmission conditions* at the wall/lumen interface in the target area. This naturally comes out of our mathematical analysis without any ad hoc arguments: a variational formulation for the zero-thickness periodic stent leading to the identification of the variational equation of the asymptotic stent and a complete constructive proof of convergence of the dose. The constructive analysis is certainly one of the main contributions of this paper, since it clarifies and justifies the general mathematical modeling of the biophysical process at hand.

Section 2 presents the time-space diffusion-transport equations for the concentration of the product using appropriate conditions on the flow at both ends of the artery. Section 3 gives the equations for the *dose* and the variational model. Since the thickness of the layer of polymer is small compared to the other geometric parameters, we let the thickness of the layer go to zero in the variational model and obtain in section 4 a new variational model for the dose which is a reasonable approximation. The construction of the asymptotic stent[2] and the corresponding variational equation for the dose are given in section 5. The generalization to stents of arbitrary geometry via an arbitrary characteristic function is developed in section 5.3. The asymptotic model is related to the *Neumann sieve* studied in [13, 5, 7], where the plane surface is replaced by the interface lumen/wall in the lateral boundary of a cylinder and the possible apparition of a jump across the interface and a "strange term coming from nowhere," as in [3, 4, 12]. However, in our limiting process the surface of the holes does not go to zero, and we do not get a jump in the dose across the interface lumen/wall in the target area. Section 6 presents selected numerical simulations for a stent with 1, 6, 12, 24, 48, 96, 192, and 384 struts, and the asymptotic case for ρ equal to 0.1, 0.2, 0.5, and 0.9. A table of the integral of the dose in the wall of the artery is presented as a function of the ratio ρ and the number of struts N. Complete results and details on the numerical implementation will be available in a more specialized paper. This paper concentrates on the theory and the discussion of the numerical simulation. Section 7 is a concluding section that reviews the theoretical and numerical results and their impact on the design process and objectives.

2. Equations for the concentration of product. Consider a section of cylindric artery of length H where the stent will be deployed (cf. Figure 2.1). For simplicity assume that the artery is made up of two homogeneous regions: the *lumen* and the *wall*. More realistic multilayer models of the wall can be considered [11], but this will be sufficient for our purposes. Before the insertion of the stent, the lumen is assumed to be the open cylinder

$$C_R \stackrel{\text{def}}{=} \left\{(x_1, x_2, z) \,:\, x_1^2 + x_2^2 < R^2, \quad 0 < z < H\right\} \tag{2.1}$$

of radius R and length H. The wall is the open domain $C_{R+E}\backslash\overline{C}_R$ between the closed cylinder $\overline{C}_R$ and the open cylinder

$$C_{R+E} \stackrel{\text{def}}{=} \left\{(x_1, x_2, z) \,:\, x_1^2 + x_2^2 < (R+E)^2, \quad 0 < z < H\right\} \tag{2.2}$$

of radius $R + E$ and length H, where E is the radial thickness of the wall.

A stent of zero thickness will be deployed in the *target area*

$$\tilde{\Sigma} \stackrel{\text{def}}{=} \left\{(x_1, x_2, z) \,:\, x_1^2 + x_2^2 = R^2, \quad \frac{H - L_s}{2} \le z \le \frac{H + L_s}{2}\right\} \tag{2.3}$$

[2]When the number of struts goes to infinity and the width of each strut goes to zero, while the surface of the target area, the ratio ρ, and the mass per unit area of product are kept fixed.

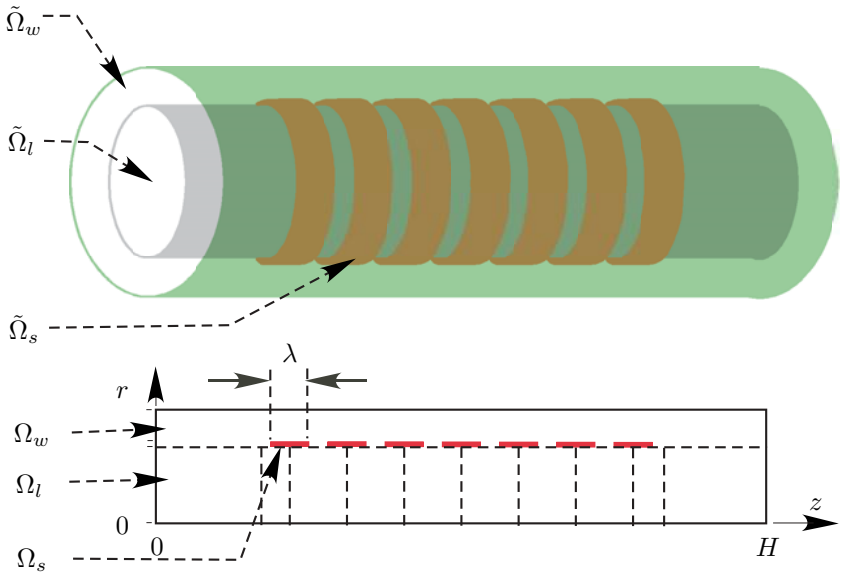


FIG. 2.1. *The lumen $\tilde{\Omega}_l$, the stent $\tilde{\Omega}_s$, and the wall $\tilde{\Omega}_w$ in $\mathbf{R}^3$ and the associated two-dimensional generating surfaces Ω_l for the lumen, Ω_s for the stent, and Ω_w for the wall.*

of the interface between the lumen and the wall. The length of the target region is $L_s < H$. The *actual stent* will be characterized by a closed subset $\tilde{\Sigma}_s$ of the surface $\tilde{\Sigma}$ (cf., for instance, one of the periodic patterns in Figure 1.2). By construction, the region $\tilde{\Sigma}$ is centered in $H/2$ at equal distance

$$z_0 \stackrel{\text{def}}{=} \frac{H - L_s}{2} > 0 \tag{2.4}$$

from the boundaries of C_{R+E} in $z = 0$ and $z = H$, which are *artificial boundaries* introduced for the analysis of the problem. The length H of the section of the artery is assumed to be enough longer than L_s that the effect of introducing an artificial boundary in $z = 0$ and $z = H$ is negligible. It also means that the region $\tilde{\Sigma}$ does not touch the boundaries of the cylinder C_{R+E} in $z = 0$ and $z = H$.

The zero-thickness stent is coated with a polymer. Coating can exist on both sides of the stent. The regions occupied by the polymer are denoted

$$\tilde{\Omega}_s^+ \stackrel{\text{def}}{=} \left\{ (x_1, x_2, z) \,:\, \left(R\frac{(x_1, x_2)}{\sqrt{x_1^2 + x_2^2}}, z \right) \in \tilde{\Sigma}_s \text{ and } R < \sqrt{x_1^2 + x_2^2} < R + e^+ \right\},$$

$$\tilde{\Omega}_s^- \stackrel{\text{def}}{=} \left\{ (x_1, x_2, z) \,:\, \left(R\frac{(x_1, x_2)}{\sqrt{x_1^2 + x_2^2}}, z \right) \in \tilde{\Sigma}_s \text{ and } R - e^- < \sqrt{x_1^2 + x_2^2} < R \right\},$$

$$\tilde{\Omega}_s \stackrel{\text{def}}{=} \tilde{\Omega}_s^+ \cup \tilde{\Omega}_s^-,$$

where e^+ and e^- are the respective thicknesses of the *coating* on the upper and lower surfaces of $\tilde{\Sigma}_s$. Once the stent is deployed, the open regions $\tilde{\Omega}_l$ and $\tilde{\Omega}_w$ occupied by the lumen and the wall are

$$\tilde{\Omega}_l \stackrel{\text{def}}{=} C_R \backslash \overline{\tilde{\Omega}_s^-} \quad \text{and} \quad \tilde{\Omega}_w \stackrel{\text{def}}{=} C_{R+E} \backslash \overline{\tilde{\Omega}_s^+ \cup C_R}. \tag{2.5}$$

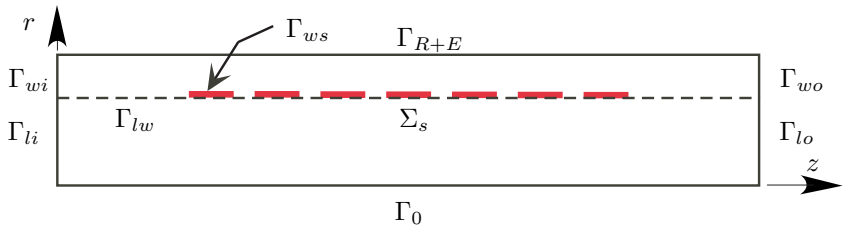


FIG. 2.2. *The two-dimensional boundaries and interfaces.*

Mathematically, $\tilde{\Omega}_l$, $\tilde{\Omega}_w$, and $\tilde{\Omega}_s^{\pm}$ are open domains in $\mathbf{R}^3$.

It is the design of the set $\tilde{\Sigma}_s$ which is the ultimate objective of the analysis. There are several aspects to this design. For instance, the stent has to be mechanically strong enough to keep the lumen open. In this paper we neglect this aspect and concentrate on the delivery of the product to the wall. For simplicity we start with a periodic stent (in the z-direction) with cylindric symmetry. Further assume that it is coated only on its upper surface, that is, $e^- = 0$, $\tilde{\Omega}_s^- = \varnothing$, and $\tilde{\Omega}_s = \tilde{\Omega}_s^+$ (cf. Figure 2.1). In addition, the stent is assumed to be a set of N equally spaced rings of width λ and zero thickness with coating of thickness $e = e^+$, that is, the region between the radius R and the radius $R + e$. The centerlines of the rings are located at coordinates

$$\{z_i \,:\, 1 \le i \le N\}, \quad z_0 < z_1 < z_2 < \cdots < z_N < z_0 + L_s, \tag{2.6}$$

along the z-axis as shown in the lower part of Figure 2.1 with $z_i + \lambda < z_{i+1}$. Thus all the rings are contained in the target region $\tilde{\Sigma}$. The domains $\tilde{\Omega}_l$, $\tilde{\Omega}_w$, and $\tilde{\Omega}_s$ and the interfaces $\tilde{\Sigma}$ and $\tilde{\Sigma}_s$ are generated by rotation of the two-dimensional open domains Ω_l, Ω_w, and Ω_s and the interfaces Σ and Σ_s (cf. Figure 2.1) around the axis Γ_0. In what follows, the tilded domains, boundaries, and interfaces will always denote the three-dimensional object generated by rotation around the axis Γ_0.

The boundary $\tilde{\Gamma}_l = \partial\tilde{\Omega}_l$ of the lumen $\tilde{\Omega}_l$ is made up of four parts:

- $\tilde{\Sigma}_s$, the interface between $\tilde{\Omega}_l$ and the region $\tilde{\Omega}_s$ occupied by the polymer;
- $\tilde{\Gamma}_{lw}$, the interface between $\tilde{\Omega}_l$ and the region $\tilde{\Omega}_w$ occupied by the wall;
- $\tilde{\Gamma}_{li}$, the part of the boundary of $\tilde{\Omega}_l$ where the blood flows in;
- $\tilde{\Gamma}_{lo}$, the part of the boundary of $\tilde{\Omega}_l$ where the blood flows out.

The corresponding parts of the generating surface are Σ_s, Γ_{lw}, Γ_{li}, Γ_{lo}, and the centerline or axis Γ_0 of the cylinders (cf. Figure 2.2).

The boundary $\tilde{\Gamma}_w = \partial\tilde{\Omega}_w$ of the wall $\tilde{\Omega}_w$ is made up of five parts:

- $\tilde{\Gamma}_{lw}$, the interface between $\tilde{\Omega}_w$ and the region $\tilde{\Omega}_l$ occupied by the lumen;
- $\tilde{\Gamma}_{ws}$, the interface between $\tilde{\Omega}_w$ and the region $\tilde{\Omega}_s$ occupied by the polymer;
- $\tilde{\Gamma}_{wi}$, the part of the boundary of $\tilde{\Omega}_w$ where $z = 0$;
- $\tilde{\Gamma}_{wo}$, the part of the boundary of $\tilde{\Omega}_w$, where $z = H$;
- $\tilde{\Gamma}_{R+E}$, the outer lateral boundary of the cylinder of radius $R + E$.

The corresponding parts of the generating surface are Γ_{lw}, Γ_{ws}, Γ_{wi}, Γ_{wo}, and the upper boundary Γ_{R+E} at $r = R + E$ (cf. Figure 2.2).

The boundary $\tilde{\Gamma}_s = \partial\tilde{\Omega}_s$ of the polymer $\tilde{\Omega}_s$ is made up of two parts:

- $\tilde{\Sigma}_s$, the interface between $\tilde{\Omega}_s$ and the region $\tilde{\Omega}_l$ occupied by the lumen;
- $\tilde{\Gamma}_{ws}$, the interface between $\tilde{\Omega}_s$ and the region $\tilde{\Omega}_w$ occupied by the wall.

The corresponding parts of the generating surface are Σ_s and Γ_{ws} (cf. Figure 2.2).

The fluid (here, the blood) in the lumen is assumed to be incompressible,

$$\operatorname{div} u = 0 \quad \text{in } \tilde{\Omega}_l, \tag{2.7}$$

where u is the *velocity of the fluid*. Further assume that

$$u \cdot n_l \le 0 \text{ on } \tilde{\Gamma}_{li} \quad \text{and} \quad u \cdot n_l \ge 0 \text{ on } \tilde{\Gamma}_{lo}, \tag{2.8}$$

$$u \cdot n_l = 0 \quad \text{or} \quad u = 0 \text{ on } \tilde{\Sigma}_s \cup \tilde{\Gamma}_{lw}. \tag{2.9}$$

Condition (2.8)–(2.9) means that the blood is coming *in* through the cross section $\tilde{\Gamma}_{li}$ and going *out* through the cross section $\tilde{\Gamma}_{lo}$. The velocity u and the pressure p will also verify the Navier–Stokes equation with the condition $u = 0$ on $\tilde{\Sigma}_s \cup \tilde{\Gamma}_{lw}$. Yet the diffusion-transport equations will still make sense under the weaker condition $u \cdot n_l = 0$ on $\tilde{\Sigma}_s \cup \tilde{\Gamma}_{lw}$. This would correspond to a non-Newtonian viscosity model, which is not the purpose of this paper. For instance, experimental data show the existence of a near-wall plasma layer that separates the blood corpuscles from the wall. The effect of this small layer is to account for a *slipping* of the flow over the wall.

Assume that the concentration $c(x,t)$ of product is given by the diffusion-transport equation (lumen) and diffusion equations (wall and the polymer):

$$\frac{\partial c}{\partial t} = \begin{cases} \operatorname{div}(D_w \nabla c) & \text{in } \tilde{\Omega}_w, \\ \operatorname{div}(D_s \nabla c) & \text{in } \tilde{\Omega}_s, \end{cases} \tag{2.10}$$

$$\frac{\partial c}{\partial t} + u \cdot \nabla c = \operatorname{div}(D_l \nabla c) \quad \text{in } \tilde{\Omega}_l, \tag{2.11}$$

where D_w, D_s, and D_l are the respective diffusion constants in the wall, the polymer, and the lumen. The *inner product* of two vectors $u = (u_1, u_2, u_3)$ and $v = (v_1, v_2, v_3)$ in $\mathbf{R}^3$ is denoted by

$$u \cdot v \stackrel{\text{def}}{=} \sum_{i=1}^{3} u_i \, v_i.$$

In view of the incompressibility condition (2.7), equation (2.11) can be rewritten

$$\frac{\partial c}{\partial t} = \operatorname{div}(D_l \nabla c - cu) \quad \text{in } \tilde{\Omega}_l, \tag{2.12}$$

since $\operatorname{div} u = 0$ implies $\operatorname{div}(cu) = \nabla c \cdot u + c \operatorname{div} u = \nabla c \cdot u$.

The boundary conditions on c are

$$\begin{aligned} &\text{wall} && \left[\frac{\partial c}{\partial n_w} = 0 \quad \text{on } \tilde{\Gamma}_{wi} \cup \tilde{\Gamma}_{wo} \cup \tilde{\Gamma}_{R+E}, \right. \\ &\text{lumen} && \left[\begin{array}{l} D_l \dfrac{\partial c}{\partial n_l} - u \cdot n_l \, c = 0 \quad \text{or} \quad c = 0 \text{ on } \tilde{\Gamma}_{li}, \\ \dfrac{\partial c}{\partial n_l} = 0 \text{ on } \tilde{\Gamma}_{lo}, \end{array} \right. \end{aligned} \tag{2.13}$$

where n_w, n_l, and n_s are the respective unit outward normals to $\tilde{\Omega}_w$, $\tilde{\Omega}_l$, and $\tilde{\Omega}_s$. The first boundary condition involving u at the entry $\tilde{\Gamma}_{li}$ of the lumen is a *transparent condition* similar to those used in [2]. It allows for some backward diffusion at the interface $\tilde{\Gamma}_{li}$. In that case the first condition (2.8) has to be strengthened to

$$\exists \beta > 0 \text{ such that } -u \cdot n_l \ge \begin{cases} 0 & \text{on } \tilde{\Gamma}_{li} \backslash \tilde{\gamma}_{li}, \\ \beta & \text{on } \tilde{\gamma}_{li} \subset \tilde{\Gamma}_{li}, \end{cases} \quad \text{and} \quad u \cdot n_l \ge 0 \text{ on } \tilde{\Gamma}_{lo}, \tag{2.14}$$

where $\tilde{\gamma}_{li}$ is some fixed subarea of the cross section $\tilde{\Gamma}_{li}$ around its center. The second case with $c = 0$ on $\tilde{\Gamma}_{li}$ corresponds to the assumption that $\tilde{\Gamma}_{li}$ is chosen sufficiently far from the region of the stent $\tilde{\Sigma}_s$ that the concentration c on $\tilde{\Gamma}_{li}$ can be taken as zero.

The conditions on c at the interfaces are

$$\begin{aligned} &\text{wall/polymer} && D_w \frac{\partial c}{\partial n_w} + D_s \frac{\partial c}{\partial n_s} = 0 \quad \text{on } \tilde{\Gamma}_{ws}, \\ &\text{wall/lumen} && D_w \frac{\partial c}{\partial n_w} + D_l \frac{\partial c}{\partial n_l} = 0 \quad \text{on } \tilde{\Gamma}_{lw}, \\ &\text{polymer/lumen} && \frac{\partial c^+}{\partial n_s} = 0 \quad \text{and} \quad \frac{\partial c^-}{\partial n_l} = 0 \quad \text{on } \tilde{\Sigma}_s. \end{aligned} \tag{2.15}$$

Recall that the lumen is isolated from the polymer and that there is an upper trace c^+ and a lower trace c^- of the concentration on the two sides of the interface $\tilde{\Sigma}_s$. $\tilde{\Sigma}_s$ is made up of the N ring-shaped cracks in the three-dimensional domain, and different boundary conditions are specified on each side.

The initial condition is

$$c(0,x) = \begin{cases} c_0(x) & \text{in } \tilde{\Omega}_s, \\ 0 & \text{in } \tilde{\Omega}_w \cup \tilde{\Omega}_l, \end{cases} \tag{2.16}$$

for some positive function $c_0(x) \geq 0$ representing the initial concentration of the product at time 0 in the polymer.

3. Mathematical models for the dose. The *dose* is the cumulative concentration integrated over time from 0 to infinity in a given position x, that is,

$$q(x) \stackrel{\text{def}}{=} \int_0^\infty c(t,x)\, dt. \tag{3.1}$$

3.1. Equations for the dose. Since all our equations are linear, the equations, boundary conditions, and interface conditions for q are readily obtained from those for c. The equations for the dose $q(x)$ are

$$\operatorname{div}(D_w \nabla q) = 0 \quad \text{in } \tilde{\Omega}_w, \tag{3.2}$$

$$\operatorname{div}(D_s \nabla q) = -c_0 \quad \text{in } \tilde{\Omega}_s, \tag{3.3}$$

$$\operatorname{div}(D_l \nabla q - qu) = 0 \quad \text{in } \tilde{\Omega}_l. \tag{3.4}$$

The boundary conditions are

$$\begin{aligned} &\text{wall} && \left[\frac{\partial q}{\partial n_w} = 0 \quad \text{on } \tilde{\Gamma}_{wi} \cup \tilde{\Gamma}_{wo} \cup \tilde{\Gamma}_{R+E}, \right. \\ &\text{lumen} && \left[\begin{aligned} &D_l \frac{\partial q}{\partial n_l} - u \cdot n_l\, q = 0 \quad \text{or} \quad q = 0 \text{ on } \tilde{\Gamma}_{li}, \\ &\frac{\partial q}{\partial n_l} = 0 \quad \text{on } \tilde{\Gamma}_{lo}. \end{aligned} \right. \end{aligned} \tag{3.5}$$

Again the choice of the first transparent condition on $\tilde{\Gamma}_{li}$ requires the stronger condi-

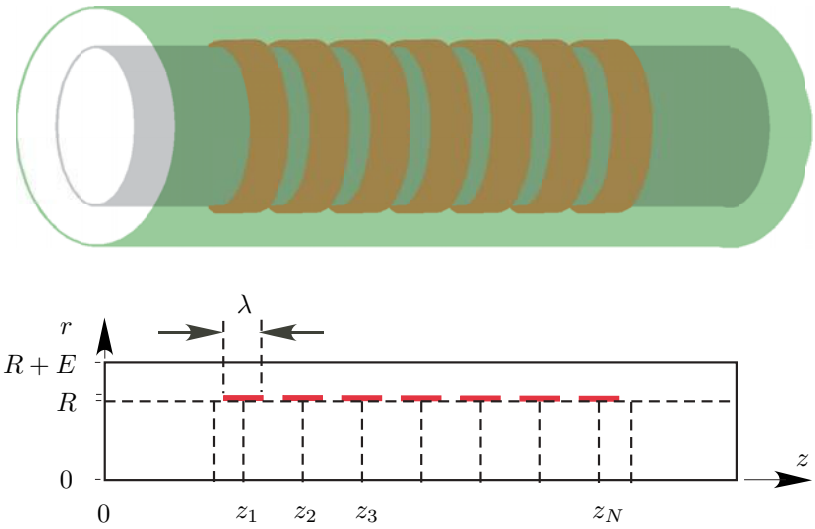


FIG. 3.1. *Schematic representation of the lumen, the polymer, and the wall (upper panel) and its* 2*-dimensional generating surface (lower panel).*

tion (2.14) on u. The conditions at the interfaces are

$$\begin{aligned}
&\text{wall/polymer} && D_w \frac{\partial q}{\partial n_w} + D_s \frac{\partial q}{\partial n_s} = 0 \quad \text{on } \tilde{\Gamma}_{ws},\\
&\text{wall/lumen} && D_w \frac{\partial q}{\partial n_w} + D_l \frac{\partial q}{\partial n_l} = 0 \quad \text{on } \tilde{\Gamma}_{lw},\\
&\text{polymer/lumen} && \frac{\partial q^+}{\partial n_s} = 0 \quad \text{and} \quad \frac{\partial q^-}{\partial n_l} = 0 \quad \text{on } \tilde{\Sigma}_s.
\end{aligned} \tag{3.6}$$

3.2. Variational equation for the dose. In this section we construct a variational formulation of the equations of the dose over the domain (cf. Figure 3.1)

$$\tilde{\Omega} \stackrel{\text{def}}{=} \left\{(x_1, x_2, z) \,:\, |x_1|^2 + |x_2|^2 < (R+E)^2,\, 0 < z < H\right\} \setminus \tilde{\Sigma}_s, \tag{3.7}$$

that is, $C_{R+E} \setminus \tilde{\Sigma}_s$. It is a bounded connected open domain with *two-dimensional cracks* along the polymer/lumen interfaces $\tilde{\Sigma}_s$. *This is not a Lipschitzian domain.* The associated space of solution is

$$V(\tilde{\Omega}) \stackrel{\text{def}}{=} \begin{cases} H^1(\tilde{\Omega}) & \text{with condition (2.14)–(2.9) on } u,\\ \{v \in H^1(\tilde{\Omega}) \,:\, v|_{\tilde{\Gamma}_{li}} = 0\} & \text{with condition (2.8)–(2.9) on } u. \end{cases} \tag{3.8}$$

We introduce the following bilinear form:

$$\begin{aligned} a(q,v) \stackrel{\text{def}}{=} &\int_{\tilde{\Omega}_w} D_w \nabla q \cdot \nabla v \, dx + \int_{\tilde{\Omega}_s} D_s \nabla q \cdot \nabla v \, dx\\ &+ \int_{\tilde{\Omega}_l} (D_l \nabla q - q u) \cdot \nabla v \, dx + \int_{\tilde{\Gamma}_{lo}} u \cdot n_l \, q \, v \, d\Gamma. \end{aligned}$$

Then $q \in V(\tilde{\Omega})$ must verify the variational equation

$$\boxed{\forall v \in V(\tilde{\Omega}), \quad a(q,v) = \int_{\tilde{\Omega}_s} c_0 \, v \, dx.} \tag{3.9}$$

The bilinear form can be rewritten as

$$a(q,v) = \int_{\tilde{\Omega}} D\,\nabla q \cdot \nabla v\,dx - \int_{\tilde{\Omega}_l} q\,u \cdot \nabla v\,dx + \int_{\tilde{\Gamma}_{lo}} u \cdot n_l\,q\,v\,d\Gamma,$$

by introducing the space-dependent diffusion defined almost everywhere on $\tilde{\Omega}$:

$$D(x) \stackrel{\text{def}}{=} \begin{cases} D_w & \text{if } x \in \tilde{\Omega}_w, \\ D_s & \text{if } x \in \tilde{\Omega}_s, \\ D_l & \text{if } x \in \tilde{\Omega}_l. \end{cases}$$

The bilinear form a is not symmetrical, but it is coercive on $V(\tilde{\Omega})$ under the two boundary conditions (2.8) and (2.9) on the velocity field u and $q = 0$ on $\tilde{\Gamma}_{li}$ and under the two boundary conditions (2.14) and (2.9) on the velocity field u for the transparent condition on q. Indeed, let $\alpha > 0$ be the minimum of D_w, D_s, and D_l. Then, using the fact that $\operatorname{div} u = 0$ and condition (2.9) and (2.8), we get

$$a(q,q) \geq \alpha \int_{\tilde{\Omega}} |\nabla q|^2\,dx - \frac{1}{2}\int_{\tilde{\Gamma}_{li}} u \cdot n_l\,|q|^2\,d\Gamma + \frac{1}{2}\int_{\tilde{\Gamma}_{lo}} u \cdot n_l\,|q|^2\,d\Gamma \geq \alpha \int_{\tilde{\Omega}} |\nabla q|^2\,dx.$$

Using condition (2.9) and (2.14), we get

$$\begin{aligned} a(q,q) &\geq \alpha \int_{\tilde{\Omega}} |\nabla q|^2\,dx - \frac{1}{2}\int_{\tilde{\Gamma}_{li}} u \cdot n_l\,|q|^2\,d\Gamma + \frac{1}{2}\int_{\tilde{\Gamma}_{lo}} u \cdot n_l\,|q|^2\,d\Gamma \\ &\geq \alpha \int_{\tilde{\Omega}} |\nabla q|^2\,dx + \frac{\beta}{2}\int_{\tilde{\gamma}_{li}} q^2\,d\tilde{\Sigma} \geq \min\left\{\alpha, \frac{\beta}{2}\right\}\left\{\int_{\tilde{\Omega}} |\nabla q|^2\,dx + \int_{\tilde{\gamma}_{li}} q^2\,d\tilde{\Sigma}\right\}. \end{aligned}$$

However, the last term on the right-hand side is an equivalent norm on $H^1(\tilde{\Omega})$ and a is coercive on it. Therefore, by the Lax–Milgram theorem (cf. [8]), there exists a unique $q \in V(\tilde{\Omega})$ solution of the variational equation (3.9).

4. Equations for the dose as the thickness of the polymer goes to zero. In this section we construct new equations for the dose as the thickness of the polymer goes to zero, while keeping the total mass of product constant in the polymer. They are obtained from the stent made up of N identical equally spaced flat rings, with $L > 0$ the distance between the center of two consecutive struts; λ, $0 < \lambda < L$, the width of the rings; and $e > 0$ the thickness of the polymer.

4.1. Parametrization of the thickness. Start from the zero-thickness stent

$$\tilde{\Sigma}_s \stackrel{\text{def}}{=} \bigcup_{i=1}^{N} \left\{ (x_1,x_2,z) \,:\, \begin{array}{l} |x_1|^2 + |x_2|^2 = R^2 \\ 0 < z < H \text{ and } |z - z_i| \leq \lambda/2 \end{array} \right\} \tag{4.1}$$

made up of N identical equally spaced flat rings, with $L > 0$ the distance between the centerlines of two consecutive struts and λ, $0 < \lambda < L$, the width of the rings (cf. Figure 3.1). Let ε, $0 < \varepsilon \leq e$, be the variable thickness of the polymer, and define the new domain occupied by the polymer as

$$\tilde{\Omega}_s^{\varepsilon} \stackrel{\text{def}}{=} \bigcup_{i=1}^{N} \left\{ (x_1,x_2,z) \,:\, \begin{array}{l} R^2 < |x_1|^2 + |x_2|^2 < (R+\varepsilon)^2 \\ 0 < z < H \text{ and } |z - z_i| < \lambda/2 \end{array} \right\},$$

where z_i is the position of the ith strut along the z-axis. This induces a new domain for the wall

$$\tilde{\Omega}_w^\varepsilon \stackrel{\text{def}}{=} \left\{(x_1,x_2,z) \,:\, \begin{array}{l} R^2 < |x_1|^2+|x_2|^2 < (R+E)^2 \\ 0<z<H \end{array}\right\} \setminus \bigcup_{i=1}^N \left\{(x_1,x_2,z) \,:\, \begin{array}{l} R^2 \le |x_1|^2+|x_2|^2 \le (R+\varepsilon)^2 \\ 0<z<H \text{ and } |z-z_i| \le \lambda/2 \end{array}\right\}.$$

For $\varepsilon = 0$, $\tilde{\Omega}_s^0 = \varnothing$ and

$$\tilde{\Omega}_w^0 = \left\{(x_1,x_2,z) \,:\, R^2 < |x_1|^2+|x_2|^2 < (R+E)^2 \text{ and } 0<z<H\right\} = C_{R+E}\backslash\overline{C_R}.$$

Since, up to a set of zero measure, $\tilde{\Omega}_w^\varepsilon \cup \tilde{\Omega}_s^\varepsilon = \tilde{\Omega}_w^0$ and $\tilde{\Omega} = \tilde{\Omega}_w^\varepsilon \cup \tilde{\Omega}_s^\varepsilon \cup \tilde{\Omega}_l = \tilde{\Omega}_w^0 \cup \tilde{\Omega}_l$, it will be convenient to define the new space-dependent diffusion D^ε almost everywhere on $\tilde{\Omega}$ as

$$\boxed{D^\varepsilon(x) \stackrel{\text{def}}{=} \begin{cases} D_w & \text{if } x \in \tilde{\Omega}_w^\varepsilon, \\ D_s & \text{if } x \in \tilde{\Omega}_s^\varepsilon, \\ D_l & \text{if } x \in \tilde{\Omega}_l, \end{cases}}$$

and the new bilinear form

$$\boxed{a^\varepsilon(q,v) \stackrel{\text{def}}{=} \int_{\tilde{\Omega}} D^\varepsilon \,\nabla q \cdot \nabla v \, dx - \int_{\tilde{\Omega}_l} q\, u \cdot \nabla v \, dx + \int_{\tilde{\Gamma}_{lo}} u \cdot n_l \, q \, v \, d\Gamma,}$$

which turns out to be coercive with the same constant (independent of ε) as for the bilinear form $a(q,v)$ on $V(\tilde{\Omega})$. Here, to make the connection with the notation of the previous sections, the bilinear form $a(q,v)$ is now equal to $a^e(q,v)$. Note that the parameter ε occurs only in the definition of the diffusion coefficient D^ε and not in the domains over which the integrals are defined.

The initial linear right-hand side

$$\ell(v) \stackrel{\text{def}}{=} \int_{\tilde{\Omega}_s} c_0 \, v \, dx \tag{4.2}$$

(for the thickness e) has to be adjusted in order to deliver the same mass of product for a thickness ε. Assume that the initial concentration c_0 is constant in $\tilde{\Omega}_s$; that is, the total mass of product in the polymer is

$$\begin{aligned} m &\stackrel{\text{def}}{=} c_0 \int_{\tilde{\Omega}_s} dx = c_0 \sum_{i=1}^N \int_{z_i-\lambda/2}^{z_i+\lambda/2} dz \int_R^{R+e} 2\pi\, r \, dr \\ &= c_0 \, N \, \lambda\pi \, \left[(R+e)^2 - R^2\right] = \boxed{c_0 \, N \, \lambda\pi \, e \, (2R+e).} \end{aligned}$$

Define the new concentration c_0^ε such that the total mass remains m over the new domain $\tilde{\Omega}_s^\varepsilon$; that is,

$$\begin{aligned} m &= \int_{\tilde{\Omega}_s^\varepsilon} c_0^\varepsilon \, dx = \sum_{i=1}^N \int_{z_i-\lambda/2}^{z_i+\lambda/2} dz \int_R^{R+\varepsilon} c_0^\varepsilon \, 2\pi \, r \, dr \\ &= c_0^\varepsilon \, N \, \lambda\pi \, \left[(R+\varepsilon)^2 - R^2\right] = \boxed{c_0^\varepsilon \, N \, \lambda \, \pi \, \varepsilon \, (2R+\varepsilon).} \end{aligned}$$

In the domain $\tilde{\Omega}_s^\varepsilon$ occupied by the polymer, choose the new concentration

$$\boxed{c_0^\varepsilon \stackrel{\text{def}}{=} \frac{1}{\varepsilon}\left(\frac{1}{\lambda\,(2R+\varepsilon)\,\pi}\right)\frac{m}{N}}$$

and the corresponding linear right-hand side

$$\boxed{\ell^\varepsilon(v) \stackrel{\text{def}}{=} \int_{\tilde{\Omega}_s^\varepsilon} c_0^\varepsilon\, v\, dx.} \tag{4.3}$$

The new variational problems indexed by ε, $0 < \varepsilon \le e$, are

$$\boxed{\exists q^\varepsilon \in V(\tilde{\Omega}) \quad \text{such that } \forall v \in V(\tilde{\Omega}), \quad a^\varepsilon(q^\varepsilon, v) = \ell^\varepsilon(v),} \tag{4.4}$$

where $\tilde{\Omega}$ is the open cylinder C_{R+E} minus the stent $\tilde{\Sigma}_s$, as defined in (3.7).

4.2. Limiting process. The next step is to determine the limit q^0 of the dose q^ε as ε goes to zero, and to show that it is a solution of a new variational equation.

THEOREM 4.1. *As $\varepsilon > 0$ goes to zero, the solution $q_\varepsilon \in V(\tilde{\Omega})$ of (4.4) weakly converges to the solution $q^0 \in V(\tilde{\Omega})$ of the variational equation*

$$\boxed{\forall v \in V(\tilde{\Omega}), \quad a^0(q^0, v) = \ell^0(v),} \tag{4.5}$$

where

$$\boxed{\ell^0(v) \stackrel{\text{def}}{=} \int_{\tilde{\Sigma}_s} c_s\, v^+\, dx, \qquad c_s \stackrel{\text{def}}{=} \frac{1}{\lambda\, 2R\,\pi}\frac{m}{N},} \tag{4.6}$$

$$\boxed{a^0(q^0, v) \stackrel{\text{def}}{=} \int_{\tilde{\Omega}} \nabla q^0 \cdot D^0\, \nabla v\, dx - \int_{\tilde{\Omega}_l} q^0\, u \cdot \nabla v\, dx + \int_{\tilde{\Gamma}_{lo}} u \cdot n_l\, q^0\, v\, d\Gamma} \tag{4.7}$$

are the respective continuous linear and bilinear forms on $V(\tilde{\Omega})$, c_s is the surface density of the product *in kg/m^2, v^+ is the trace of v on the upper side of $\tilde{\Sigma}_s$, and D^0 is the diffusion defined almost everywhere in $\tilde{\Omega}$:*

$$\boxed{D^0(x) \stackrel{\text{def}}{=} \begin{cases} D_w & \text{if } x \in \tilde{\Omega}_w^0, \\ D_l & \text{if } x \in \tilde{\Omega}_l. \end{cases}} \tag{4.8}$$

Proof. The proof follows by standard arguments. ∎

4.3. Equations for q^0. From the variational equation (4.5) for q^0 we get the following set of equations for the dose $q^0(x,t)$:

$$\operatorname{div}(D_w \nabla q^0) = 0 \quad \text{in } \tilde{\Omega}_w^0, \tag{4.9}$$

$$\operatorname{div}(D_l \nabla q^0 - q^0 u) = 0 \quad \text{in } \tilde{\Omega}_l. \tag{4.10}$$

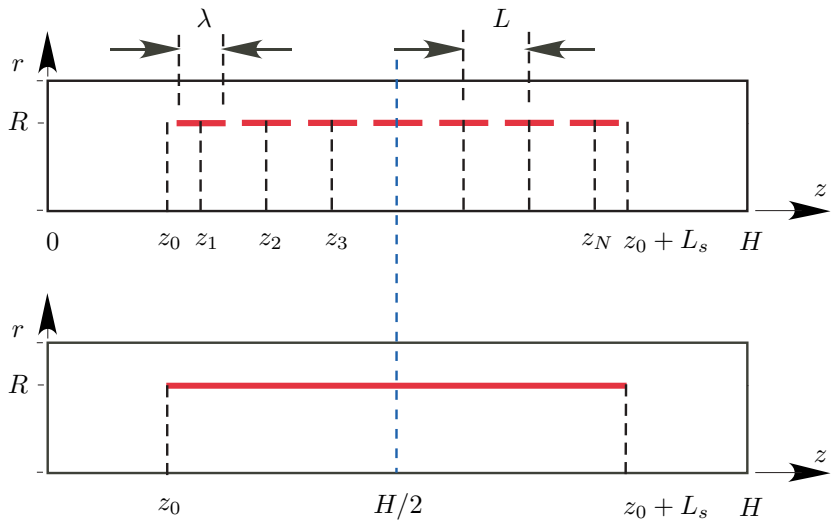


FIG. 5.1. *Stent/lumen and stent/wall interfaces for N struts (upper panel) and as N goes to infinity (lower figure).*

The boundary conditions are

$$
\text{(4.11)}\qquad
\begin{aligned}
&\text{wall} \quad \begin{cases} \dfrac{\partial q^0}{\partial n_w} = 0 & \text{on } \tilde{\Gamma}_{wi} \cup \tilde{\Gamma}_{wo} \cup \tilde{\Gamma}_{R+E}, \\ D_w \dfrac{\partial q^{0+}}{\partial n_w} = c_s & \text{on } \tilde{\Sigma}_s, \end{cases} \\
&\text{lumen} \quad \begin{cases} D_l \dfrac{\partial q^0}{\partial n_l} - u \cdot n_l \, q^0 = 0 \quad \text{or} \quad q^0 = 0 & \text{on } \tilde{\Gamma}_{li}, \\ \dfrac{\partial q^0}{\partial n_l} = 0 & \text{on } \tilde{\Gamma}_{lo}, \\ \dfrac{\partial q^{0-}}{\partial n_l} = 0 & \text{on } \tilde{\Sigma}_s. \end{cases}
\end{aligned}
$$

The condition at the interface is

$$
\text{(4.12)}\qquad \text{wall/lumen} \quad D_w \frac{\partial q^0}{\partial n_w} + D_l \frac{\partial q^0}{\partial n_l} = 0 \quad \text{on } \tilde{\Gamma}_{lw}.
$$

5. Asymptotic stent. In the design of the stent, we are left with several parameters: the surface density of product $c_s = m/(2\pi RN\lambda)$, the total length of the space occupied by the stent $L_s = NL$, the ratio $\rho = N\lambda/NL = \lambda/L$ between the width of a strut λ and the distance L between two successive struts, and the total number N of struts. Recall that the stent (see Figure 5.1) is specified by the set

$$
\tilde{\Sigma}_s^N \overset{\text{def}}{=} \bigcup_{i=1}^{N} \left\{ (x_1, x_2, z) : \begin{array}{l} x_1^2 + x_2^2 = R^2 \\ z_i - \dfrac{\lambda}{2} \le z \le z_i + \dfrac{\lambda}{2} \end{array} \right\}, \quad z_i = z_0 + \left(i - \frac{1}{2} \right) L, \ 1 \le i \le N,
$$

where the superscript emphasizes the dependence on N. Recall that

$$
z_0 = \frac{H - L_s}{2} > 0.
$$

Therefore the N-strut stent and the asymptotic stent will be centered in $[0, H]$:

$$[z_0, z_0 + L_s] = [z_0, z_0 + NL] \subset]0, H[\quad \text{and} \quad z_0 + \frac{NL}{2} = \frac{H}{2}.$$

In this section we construct an asymptotic model for the dose $q_N^0 = q^0$ as the number of struts goes to infinity, while keeping constant the length L_s, the ratio ρ, and the surface density of the product c_s. Again the superscript on q_N^0 emphasizes the dependence of q^0 on N. The main technical difference between this and the asymptotic analysis in the previous section is that the space of the solution will also depend on N. In general, even if we can find a uniform bound in a large enough function space independent of N, we will not be able to use test functions in the fixed larger space unless the projection onto the N-dependent solution spaces is strongly continuous. This asymptotic problem is very similar to the *Neumann sieve* studied by [13, 5, 7], where the plane surface is replaced by the lateral boundary of the cylinder of radius R. Fortunately here the total surface of the *holes* is constant and different from zero in the limiting process, and there will be no discontinuity of the trace of the asymptotic solution. The proof will use the weak convergence of a sequence of characteristic functions associated with the N-stent strut as in [6].

5.1. Construction of the asymptotic problem. Recall that for fixed N the dose q_N^0 is the solution in the space

$$V(\tilde{\Omega}^N) \stackrel{\text{def}}{=} \begin{cases} H^1(\tilde{\Omega}^N), & \text{condition (2.14)–(2.9) on } u, \\ \{v \in H^1(\tilde{\Omega}^N) \,:\, v|_{\tilde{\Gamma}_{li}} = 0\}, & \text{condition (2.8)–(2.9) on } u, \end{cases} \tag{5.1}$$

$$\tilde{\Omega}^N \stackrel{\text{def}}{=} \{(x_1, x_2, z) \,:\, |x_1|^2 + |x_2|^2 < (R+E)^2,\, 0 < z < H\} \backslash \tilde{\Sigma}_s^N \tag{5.2}$$

of the variational equation

$$\forall v \in V_N, \quad a^0(q_N^0, v) = \ell_N^0(v). \tag{5.3}$$

The linear form can now be rewritten in terms of the following characteristic function on $[0, H]$,

$$\chi_N(z) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } z \in \displaystyle\bigcup_{i=1}^{N} \left[z_i - \frac{\lambda}{2}, z_i + \frac{\lambda}{2} \right], \\ 0 & \text{otherwise,} \end{cases} \tag{5.4}$$

$$\begin{aligned} \ell_N^0(v) = \int_{\tilde{\Sigma}_s^N} c_s\, v^+\, dx &= \sum_{i=1}^{N} \int_{z_i - \frac{\lambda}{2}}^{z_i + \frac{\lambda}{2}} dz\, R \int_0^{2\pi} d\theta\, c_s\, v(R^+, \theta, z) \\ &= c_s \int_0^H dz\, \chi_N(z)\, R \int_0^{2\pi} d\theta\, v(R^+, \theta, z), \end{aligned} \tag{5.5}$$

or in terms of the characteristic function $\chi_{\tilde{\Sigma}_s^N}$, defined on the target area $\tilde{\Sigma}$ in the lateral boundary of the cylinder C_R,

$$\ell_N^0(v) = \int_{\tilde{\Sigma}} c_s \chi_{\tilde{\Sigma}_s^N}\, v^+\, d\tilde{\Sigma}, \quad \chi_{\tilde{\Sigma}_s^N}(x) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } x \in \tilde{\Sigma}_s^N, \\ 0 & \text{otherwise.} \end{cases}$$

The bilinear form

$$\boxed{a^0(w,v) = \int_{\tilde{\Omega}^0_w} D_w\, \nabla w \cdot \nabla v\, dx + \int_{\tilde{\Omega}_l} (D_l\, \nabla w - w\, u) \cdot \nabla v\, dx + \int_{\tilde{\Gamma}_{lo}} u \cdot n_l\, w\, v\, d\Gamma} \tag{5.6}$$

is independent of N. Assuming that there exist constants L_s, $0 < L_s < H$, and ρ, $0 < \rho < 1$, such that, as N goes to infinity, $LN = L_s$ and $\lambda N = \rho L_s$, it is readily seen that the sequence $\{\chi_N\}$ is uniformly bounded in $L^2(0,H)$ and weakly convergent:

$$\begin{aligned} \forall p \geq 1, \quad & \int_0^H (\chi_N)^p\, dz = \int_0^H \chi_N\, dz = N\lambda = \rho\, L_s = \text{constant} \\ & \Rightarrow \chi_N \rightharpoonup \rho\, \chi_{[z_0, z_0+L_s]} \quad L^2(0,H)\text{-weak}, \\ \forall \varphi \in L^2(0,H), \quad & \int_0^H \chi_N\, \varphi\, dz \to \rho \int_{z_0}^{z_0+L_s} \varphi\, dz, \end{aligned}$$

where

$$\chi_{[z_0,z_0+L_s]}(z) \stackrel{\text{def}}{=} \begin{cases} 1, & z \in [z_0, z_0+L_s], \\ 0 & \text{otherwise.} \end{cases} \tag{5.7}$$

Equivalently,

$$\chi_{\tilde{\Sigma}^N_s} \rightharpoonup \rho\, \chi_{\tilde{\Sigma}} \quad L^2(\tilde{\Sigma})\text{-weakly} \quad \text{and} \quad \forall \varphi \in L^2(\tilde{\Sigma}), \quad \int_{\tilde{\Sigma}} \chi_{\tilde{\Sigma}^N_s}\, \varphi\, d\Sigma \to \rho \int_{\tilde{\Sigma}} \varphi\, d\Sigma,$$

where $\tilde{\Sigma}$ is the target area as defined in (2.3).

Since $L_s = NL$, $\lambda N = \rho L_s$, and $\rho = \lambda / L$ are constant in the limiting process, we get for all v in $L^2(\tilde{\Sigma})$

$$\begin{aligned} \ell^0_N(v) &= c_s \int_{\tilde{\Sigma}} \chi_{\tilde{\Sigma}^N_s}\, v\, d\Sigma = c_s \int_0^H dz\, \chi_N(z)\, R \int_0^{2\pi} d\theta\, v(R^+, \theta, z) \\ &\to \ell^0_\infty(v) \stackrel{\text{def}}{=} c_s\, \rho \int_{z_0}^{z_0+L_s} dz\, R \int_0^{2\pi} d\theta\, v(R^+, \theta, z) = c_s\, \rho \int_{\tilde{\Sigma}} v^+\, d\Gamma. \end{aligned}$$

This suggests introducing the new domain $\tilde{\Omega}^\infty$,

$$\tilde{\Omega}^\infty \stackrel{\text{def}}{=} \tilde{\Omega}^0 \backslash \tilde{\Sigma}, \tag{5.8}$$

in the open cylinder C_{R+E},

$$\tilde{\Omega}^0 \stackrel{\text{def}}{=} C_{R+E} = \left\{(x_1, x_2, z) \,:\, |x_1|^2 + |x_2|^2 < (R+E)^2,\, 0 < z < H\right\}, \tag{5.9}$$

along with the new larger space of solution

$$\boxed{V(\tilde{\Omega}^\infty) \stackrel{\text{def}}{=} \begin{cases} H^1(\tilde{\Omega}^\infty), & \text{condition (2.14)–(2.9) on } u, \\ \left\{v \in H^1(\tilde{\Omega}^\infty) \,:\, v|_{\tilde{\Gamma}_{li}} = 0\right\}, & \text{condition (2.8)–(2.9) on } u, \end{cases}} \tag{5.10}$$

and the smaller space

$$\boxed{V(\tilde{\Omega}^0) \stackrel{\text{def}}{=} \begin{cases} H^1(\tilde{\Omega}^0), & \text{condition (2.14)–(2.9) on } u, \\ \left\{v \in H^1(\tilde{\Omega}^0) \,:\, v|_{\tilde{\Gamma}_{li}} = 0\right\}, & \text{condition (2.8)–(2.9) on } u. \end{cases}} \tag{5.11}$$

Observe that for all $N \geq 1$, $V(\tilde{\Omega}^0) \subset V(\tilde{\Omega}^N) \subset V(\tilde{\Omega}^\infty)$, and recall that the linear term acts only on the upper part of the new crack $\tilde{\Sigma}$, that is,

$$\boxed{\ell^0_\infty(v) \stackrel{\text{def}}{=} \int_{\tilde{\Sigma}} \rho\, c_s\, v^+\, d\Gamma.} \tag{5.12}$$

It is easy to show that the solutions q^0_N are uniformly bounded in the norm of $V(\tilde{\Omega}^\infty)$, that there is a subsequence which weakly converges to some $q^0_\infty \in V(\tilde{\Omega}^\infty)$, and that q^0_∞ is a solution of the variational equation

$$\boxed{\forall v \in V(\tilde{\Omega}^0), \quad a^0(q^0_\infty, v) = \ell^0_\infty(v).} \tag{5.13}$$

However this equation is incomplete since the test function belongs to the smaller space $H^1(\tilde{\Omega}^0)$ which does not see the crack $\tilde{\Sigma}$.

THEOREM 5.1. *Assume that there exist constants L_s, $0 < L_s < H$, and ρ, $0 < \rho < 1$, such that, as N goes to infinity,*[3] *$LN = L_s$ and $\lambda N = \rho L_s$. The sequence of solutions $q^0_N \in V(\tilde{\Omega}^N)$ converges $V(\tilde{\Omega}^\infty)$-weakly to q^0_∞, which is the solution in $V(\tilde{\Omega}^0)$ of the variational equation*

$$\boxed{\exists q^0_\infty \in V(\tilde{\Omega}^0) \quad \textit{such that}\, \forall v \in V(\tilde{\Omega}^0), \quad a^0(q^0_\infty, v) = \ell^0_\infty(v),} \tag{5.14}$$

where the continuous bilinear and the linear forms a^0 and ℓ^0_∞ are given by the expressions (5.6) *and* (5.12).

Remark 5.1. When the product is applied on the inner and outer surfaces of the stent, the asymptotic model predicts that the two sides contribute to increasing the total dose in the wall.

Proof. By the weak convergence of $\{q^0_N\}$ in $V(\tilde{\Omega}^\infty)$, the *jump* $[q^0_N] \stackrel{\text{def}}{=} \left(q^{0+}_N - q^{0-}_N\right)\big|_{\tilde{\Sigma}}$ of q^0_N across $\tilde{\Sigma}$ strongly converges:

$$\left[q^0_N\right] \to \left[q^0_\infty\right] \quad L^2(\tilde{\Sigma})\text{-strongly.}$$

By continuity of the trace of q^0_N across the region of the holes $\tilde{\Sigma} \backslash \tilde{\Sigma}^N_s$,

$$\forall \varphi \in L^2(\tilde{\Sigma}), \quad \int_{\tilde{\Sigma} \backslash \tilde{\Sigma}^N_s} [q^0_N]\, \varphi\, d\Gamma = 0.$$

Since we already have

$$\chi_N \rightharpoonup \rho\, \chi_{[z_0, z_0 + L_s]} \quad L^2(0,H)\text{-weakly} \quad \text{and} \quad \chi_{\tilde{\Sigma}^N_s} \rightharpoonup \rho\, \chi_{\tilde{\Sigma}} \quad L^2(\tilde{\Sigma})\text{-weakly,}$$

then for all $\varphi \in L^2(\tilde{\Sigma})$

$$\begin{aligned} 0 = \int_{\tilde{\Sigma} \backslash \tilde{\Sigma}^N_s} [q^0_N]\, \varphi\, d\Gamma &= \int_{\tilde{\Sigma}} \left(1 - \chi_{\tilde{\Sigma}^N_s}\right)\, [q^0_N]\, \varphi\, d\Gamma \\ &\to \int_{\tilde{\Sigma}} (1 - \rho\, \chi_{\tilde{\Sigma}})\, [q^0_\infty]\, \varphi\, d\Gamma = (1-\rho) \int_{\tilde{\Sigma}} [q^0_\infty]\, \varphi\, d\Gamma \\ &\Rightarrow\ \forall \varphi \in L^2(\tilde{\Sigma}), \quad (1-\rho) \int_{\tilde{\Sigma}} [q^0_\infty]\, \varphi\, d\Gamma = 0. \end{aligned}$$

Hence for $0 \leq \rho < 1$

$$[q^0_\infty] = 0 \text{ along } \tilde{\Sigma} \quad \Rightarrow\ q^0_\infty \in H^1(\tilde{\Omega}^0) \quad \Rightarrow\ q^0_\infty \in V(\tilde{\Omega}^0).$$

Combining $q^0_\infty \in V(\tilde{\Omega}^0)$ with (5.13) and the fact that a^0 is coercive on $V(\tilde{\Omega}^0)$, we conclude that q^0_∞ is the unique solution of the variational equation (5.14). ∎

[3] $L = L_s/N$ and $\lambda = \rho L_s/N$ both depend on N and go to zero as N goes to infinity.

5.2. Equations for the dose of the asymptotic stent. From the variational equation (5.14) for q^0_∞ we get the following set of equations for the dose $q^0_\infty(x,t)$:

$$\operatorname{div}(D_w \nabla q^0_\infty) = 0 \quad \text{in } \tilde{\Omega}^0_w, \tag{5.15}$$

$$\operatorname{div}(D_l \nabla q^0_\infty - q^0_\infty u) = 0 \quad \text{in } \tilde{\Omega}_l. \tag{5.16}$$

The boundary conditions are

$$\begin{aligned} &\text{wall} && \begin{cases} \dfrac{\partial q^0_\infty}{\partial n_w} = 0 & \text{on } \tilde{\Gamma}_{wi} \cup \tilde{\Gamma}_{wo} \cup \tilde{\Gamma}_{R+E}, \end{cases} \\ &\text{lumen} && \begin{cases} D_l \dfrac{\partial q^0_\infty}{\partial n_l} - u \cdot n_l \, q^0_\infty = 0 \text{ or } q^0_\infty = 0 & \text{on } \tilde{\Gamma}_{li}, \\ \dfrac{\partial q^0_\infty}{\partial n_l} = 0 & \text{on } \tilde{\Gamma}_{lo}. \end{cases} \end{aligned} \tag{5.17}$$

The condition at the interface is

$$\text{wall/lumen} \quad D_w \frac{\partial q^0_\infty}{\partial n_w} + D_l \frac{\partial q^0_\infty}{\partial n_l} = \begin{cases} 0 & \text{on } \tilde{\Gamma}^\infty_{lw}, \\ \rho \, c_s & \text{on } \tilde{\Sigma}, \end{cases} \tag{5.18}$$

where

$$\tilde{\Gamma}^\infty_{lw} = \left\{ (x_1, x_2, z) \,:\, x_1^2 + x_2^2 + z^2 = R^2 \text{ and } z \in]0, H[\setminus [z_0, z_0 + L_s] \right\}. \tag{5.19}$$

5.3. Extension to general periodic stents. The previous modeling and asymptotic analysis rests on the introduction of a periodic characteristic function (only a function of the axial variable z). Yet the modeling and the variational equations of section 4 remain true for an arbitrary characteristic function $\chi(z,\theta)$ of the axial variable z and the angular variable θ in cylindrical coordinates. Stents with a periodicity in a transverse direction (e.g., a spiral or helical shape) can now be considered and, once the pattern (cf. Figure 1.2) of the periodic stent has been selected, an asymptotic analysis similar to that of the previous subsections can be carried out.

Given an arbitrary characteristic function χ defined in the target area $\tilde{\Sigma}$,

$$\chi \in X(\tilde{\Sigma}) \stackrel{\text{def}}{=} \{\chi \in L^2(\tilde{\Sigma}) \,:\, (1-\chi)\chi = 0 \text{ a.e. in } \tilde{\Sigma}\}, \tag{5.20}$$

the zero-thickness stent is completely specified by

$$\tilde{\Sigma}_s(\chi) \stackrel{\text{def}}{=} \{x \in \tilde{\Sigma} \,:\, \chi(x) = 1\}, \tag{5.21}$$

the cracked open domain by

$$\tilde{\Omega}(\chi) \stackrel{\text{def}}{=} \{(x_1, x_2, z) \,:\, x_1^2 + x_2^2 < (R+E)^2 \text{ and } 0 < z < H\} \setminus \tilde{\Sigma}_s(\chi), \tag{5.22}$$

and the space of solutions by

$$V(\chi) = V(\tilde{\Omega}(\chi)) \stackrel{\text{def}}{=} \{v \in V(\tilde{\Omega}) \,:\, (1-\chi)\,[v] = 0 \text{ on } \tilde{\Sigma}\}, \tag{5.23}$$

where $[v]$ denotes the jump of v across the target surface $\tilde{\Sigma}$. The dose $q = q(\chi)$ is the solution of the variational equation

$$\boxed{\exists q \in V(\chi) \quad \text{such that } \forall v \in V(\chi), \quad a^0(q,v) = \ell^0(\chi; v),} \tag{5.24}$$

TABLE 6.1
Parameters.

	Artery and Blood		
Notation	Description	mm	a-dimensional
R	Radius of the lumen	1.5	0.5
E	Thickness of the wall of the artery	0.4	0.133
Re	Reynold's number		141.8
Pe_l	Peclet's number		10^8
	Wiktor stent		
Notation	Description	mm	x /3mm
N	Number of struts		24
$\lambda = 2r$	Diameter of the strut	0.15	0.05
L	Distance between two struts	0.7	0.233
ρ	Ratio = area of the stent/$2\pi\, R\, L_s$		0.214
L_s	Length of the target area	16.8	5.6
	Geometry		
Notation	Description	mm	a-dimensional
R	Radius of the lumen	1.5	0.5
E	Thickness of the artery wall	0.4	0.133
$R+E$	Radius of the artery	1.9	0.63333
L_s	Length of the target area	16.8	5.6
z_0	Length of the inlet section	16.8	5.6
z_0	Length of the outlet section	16.8	5.6
H	Length of the artery	50.4	16.8
ρ	Ratio = area of the stent/$2\pi\, R\, L_s$		0.1 to 0.9
	Parameters of the diffusion		
Notation	Description		a-dimensionsl
Pe_p	Diffusion in the wall (Peclet's number)		10^8
Pe_l	Diffusion in the lumen (Peclet's number)		10^8
c_s	Surface density of product		10^{-8}

where

$$a^0(w,v) \stackrel{\text{def}}{=} \int_{\tilde{\Omega}_p^0} D_w\, \nabla w \cdot \nabla v\, dx + \int_{\tilde{\Omega}_l} (D_l\, \nabla w - w\, u) \cdot \nabla v\, dx + \int_{\tilde{\Gamma}_{lo}} u \cdot n_l\, w\, v\, d\Gamma, \tag{5.25}$$

$$\ell^0(\chi; v) \stackrel{\text{def}}{=} \int_{\tilde{\Sigma}} c_s\, \chi\, v\, d\tilde{\Sigma}. \tag{5.26}$$

The notation emphasizes the fact that ℓ^0 and V both depend on χ. This generalizes the equation of the dose to a stent of arbitrary geometry.

6. Numerical experimentation. In this section we complete the theoretical results by extensive numerical simulations in order to get a feeling for the kind of phenomena involved. The velocity profile u of the flow in the lumen has been obtained by solving the incompressible Navier–Stokes equations (269,000 unknowns). The geometry and the equation of the dose have been scaled in order to work with dimensionless variables. The geometry of the artery, the parameters of the incompressible Navier–Stokes equation, and the characteristics of the diffusion-transport equation are given in Table 6.1 with the condition $q = 0$ on $\tilde{\Sigma}_{li}$. As a point of comparison, the parameters of the Wiktor stent were approximatively $L = 0.7$ mm for the distance between the centers of two struts, $\lambda = 0.15$ mm for the width, and 3–4 mm for the diameter. The number of struts was $N = 24$, which gives a target area of length $L_s = 16.8$ mm and a ratio $\rho = 0.15/0.7 \simeq 0.214$.

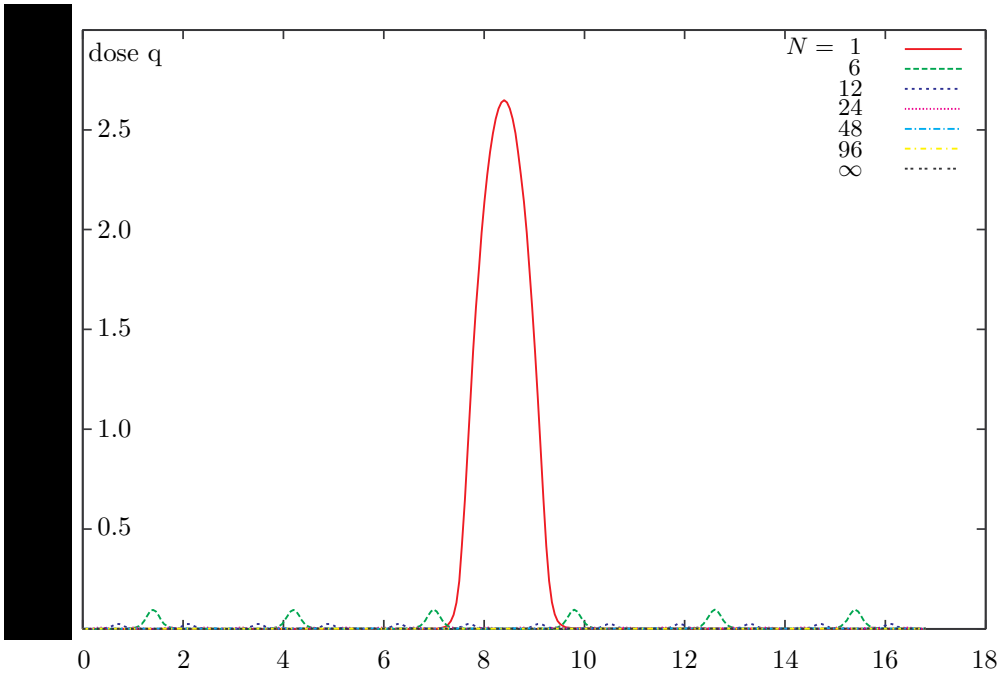


FIG. 6.1. *Dose versus the position z along the axis at the exterior radius $R+E=0.6333$ of the artery as a function of the number of struts N for $\rho=0.1$.*

We have made the computations for $\rho=0.1$, 0.2, 0.5, and 0.9 and for N from 1 to 382. We have chosen to display the dose at the outer radius $R+E$ of the artery for the case $\rho=0.1$ in order to avoid the large fluctuations at the lumen/wall interface. The graphs of the dose start at $z_0=6$ and extend to z_0+18, which goes slightly beyond the target region of length $L_s=16.8$ with the variable $z-z_0$ in abscissa. They are displayed for different values of the number of struts N. Because of the broad range of numerical values of the dose, three sets of graphs are used. Figure 6.1 is in the range of 0 to 2.7 for $N=1$, 6, 12, 24, 48, 96, ∞. At that scale, only the cases $N=1$, 6, and 12 are visible. Figure 6.2 has two sets of graphs: the first one in the range of 0 to 0.0065 for $N=24$, 48, 96, 192, 384, and $N=\infty$, and the second one in the range of 0 to 0.0020 for $N=48$, 96, 192, 384, and $N=\infty$.

As an indication of the highly irregular behavior of the dose in the vicinity of the stent, two figures give an (r,z)-plot of the dose at the upstream end of the target region occupied by the stent. For $\rho=0.1$, Figure 6.3 corresponds to $N=382$ struts and Figure 6.4 to the asymptotic stent. Notice the sharp spikes in the area of the struts and the sharp drop between two struts. It is also interesting that, even if the asymptotic theory predicts that the dose is continuous across the wall/lumen interface, a sharp drop is observed in the dose near the target area $\tilde{\Sigma}$ of the wall/lumen interface due to the high level of convection in the lumen. To get sharp and stable results, a very large number of variables was used in the numerical computations. Complete numerical tests and a description of the numerical method used will be reported in a subsequent paper. Finally, Table 6.2 gives the integral of the dose as a function of the number of struts N and the ratio ρ.

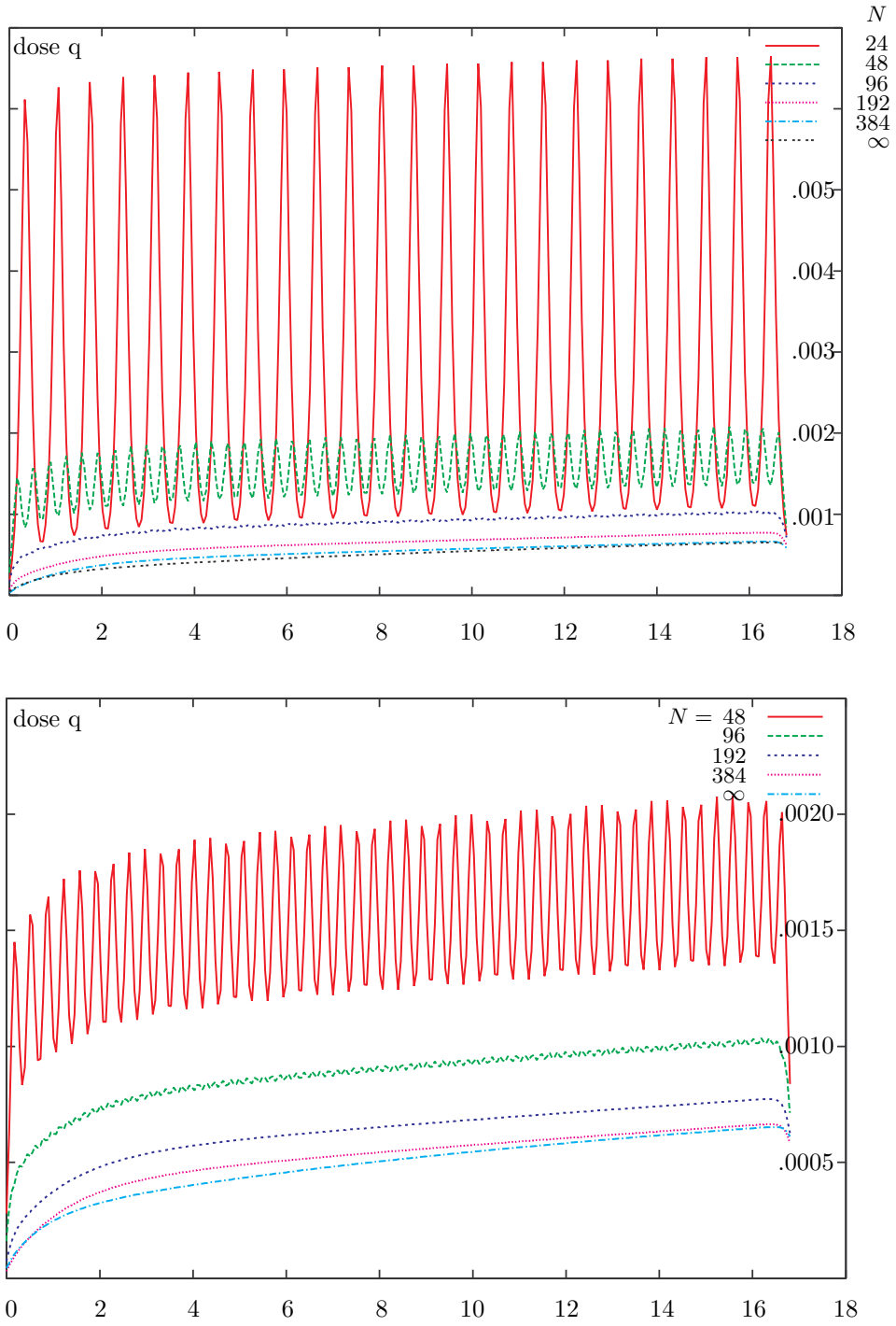


FIG. 6.2. *Dose versus the position z along the axis at the exterior radius $R+E=0.6333$ of the artery as a function of the number of struts N for $\rho=0.1$. The lower graph is a zoom of the upper one to show the dose for $N \geq 48$ in the range from 0 to .0020.*

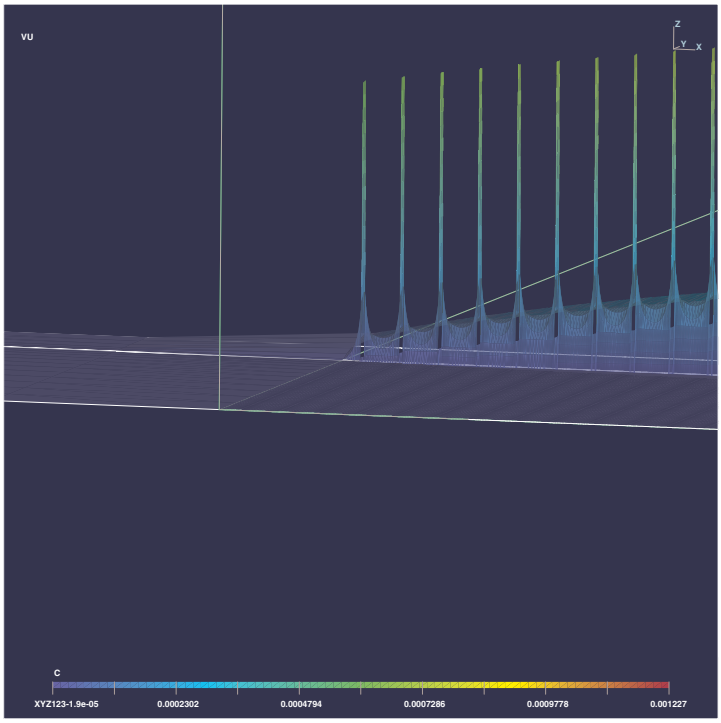


FIG. 6.3. *Dose as a function of* (r, z) *for* $N = 382$ *at the upstream end of* $\tilde{\Sigma}$.

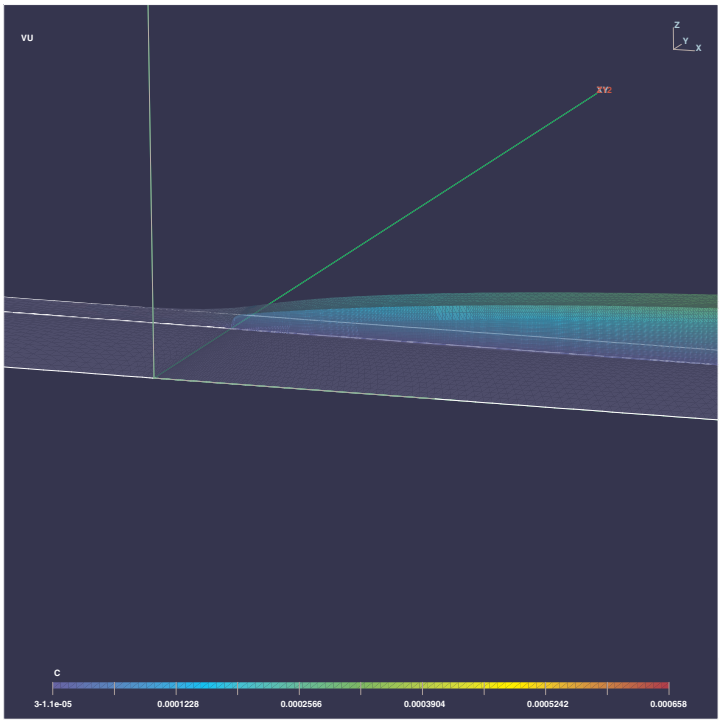


FIG. 6.4. *Dose as a function of* (r, z) *for* $N = \infty$ *at the upstream end of* $\tilde{\Sigma}$.

TABLE 6.2
Integral of the dose in the wall as a function of ρ and N (blanks indicate no computation).

N	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.9$
1	$2.44626\ 10^{-1}$	$1.76146\ 10^{+0}$	$2.57995\ 10^{+1}$	$1.47576\ 10^{+2}$
6	$1.57697\ 10^{-2}$	$7.96223\ 10^{-2}$	$8.84819\ 10^{-1}$	$4.64757\ 10^{+0}$
12	$7.37995\ 10^{-3}$	$3.14226\ 10^{-2}$	$2.79274\ 10^{-1}$	$1.40548\ 10^{+0}$
24	$3.82175\ 10^{-3}$	$1.45893\ 10^{-2}$	$1.04696\ 10^{-1}$	$5.27827\ 10^{-1}$
48	$2.10712\ 10^{-3}$	$7.40946\ 10^{-3}$	$4.83751\ 10^{-2}$	$2.41867\ 10^{-1}$
96	$1.25858\ 10^{-3}$	$3.80518\ 10^{-3}$	$2.47581\ 10^{-2}$	$1.168133\ 10^{-1}$
192	$9.4094\ 10^{-4}$			
384	$8.0285\ 10^{-4}$			
asymptotic	$7.79\ 10^{-4}$	$1.54234\ 10^{-3}$	$3.85586\ 10^{-3}$	$6.94056\ 10^{-3}$

7. Concluding remarks.

7.1. The onchocerciasis control problem. A problem closely related to the coating of stents is the onchocerciasis control program [2]. Black flies are known not only as a nuisance, causing economic losses in different areas of human activities, but also as transmitters of pathogens and parasites to man and animals. In some areas, black flies are vectors of a filarial worm (Onchocerca volvulus) which causes a serious endemic disease whose final stage is known as river blindness. The strategy chosen to combat this parasite was to break the chain of transmission by destroying the vector at its most vulnerable state, that is, the larval state. To control black fly larvae in running waters, special products have been developed with targeted toxic effects. Helicopters are used to periodically spray the rivers at prescribed sites over very large geographical areas. The objective was to minimize the total amount of larvicide used to treat a segment of river while maintaining a mortality rate P of the larvae at each point in the river. The control parameters were the number of injections, the amount injected, and the location of the injections. It turned out that the minimum total weight corresponds to an infinity of uniformly distributed injection points. Of course, this asymptotic solution is not practically implementable when the injections are made by helicopters crossing the river. In practice, with real parameters, there was a significant drop in the minimal weights going from 1 (166.9kg) to 2 (22.97kg) to 3 (13.12kg) injections, and 20 (5.9kg) injections gave a total weight close to the asymptotic solution (5.34kg). Experiments and measurements were possible in the laboratory, and a reliable relation between the mortality rate P of the larvae and the required dose was established.

TABLE 7.1
Approximate amplitude of the dose for $\rho = 0.1$ as a function of N.

N	1	6	24	48	96	192	384	asympt.
Dose	2.6000	0.1000	0.0068	0.0020	0.0010	0.00075	0.00065	0.00065

7.2. Analogies between the two problems. The injection points on the river correspond to the positions of our struts. What is required is a relatively uniform dose in the wall along the target area in the vicinity of the lumen. In Figure 6.1 the case $N = 1$ gives a peak of 2.6 concentrated at the center of the stent for $N = 1$. Figure 6.2 goes to smaller scales. Starting with $N = 6$, the dose is roughly periodic in the target area and its amplitude decreases as N increases, as shown in Table 7.1. As in the onchocerciasis control problem and with realistic parameters, we observe a very high sensitivity of the amplitude of the dose for N small, but after $N = 24$ we have

practically reached the asymptotic level. Recall that the original Wiktor stent had $N = 24$ struts. If the mass surface density c_s becomes a design parameter, we could require a minimum therapeutic dose q_{min} on the upper side of the target area $\tilde{\Sigma}$ or in a small region S' above $\tilde{\Sigma}$,

$$q(\chi)^+ \geq q_{min} \text{ on } \tilde{\Sigma} \quad \text{or} \quad q(\chi)^+ \geq q_{min} \text{ on } S', \tag{7.1}$$

$$S' \stackrel{\text{def}}{=} \left\{ (x_1, x_2, z) \,:\, \begin{array}{l} R^2 < x_1^2 + x_2^2 < (R + e')^2 \\ z_0 < z < z_0 + L_s \end{array} \right\}, \quad 0 < e' < E, \tag{7.2}$$

while minimizing c_s.

This strategy is probably not far from what is required for the design. Yet the wall of the artery is quite different from a river. In the wall there are biochemical effects and dynamics that are not fully understood or, more importantly, quantified. Following the implant of a stent, the endothelial wall is seriously damaged. It is necessary to delay the growth of smooth muscle cells to give time for the arterial wall to be recolonized by endothelial cells that contain chemical mediators responsible for orderly control of the growth process. Without such cells the growth process goes incorrectly. What is required is a relationship between the rate of growth of the smooth muscle cells (analgous to the larval mortality rate in [2]) and the dose in the vicinity of the wall/lumen interface. This could be obtained by ex vivo experiments and measurements to relate the rate of growth of the smooth muscle cells to the therapeutic dose.

As a side remark, clinical observations indicate that restenosis often takes place at the ends of the stent. This is corroborated by the fact that the dose sharply drops at both ends of the stent, as predicted in Figures 6.2, 6.3, and 6.4.

7.3. Impact on the design process. Using real parameters, our numerical simulations have revealed the high sensitivity of the dose to the number of struts N of a periodic stent. The space distribution of the dose is almost periodic in the target area: It is maximal at the strut and decreases to a minimal value between two adjacent struts. The results also show that the design parameter N effectively controls the amplitude and the uniformity of the dose in the target area and that the dose is always above the dose of the asymptotic stent (cf. Figure 6.2). Furthermore, the amplitude of the dose diminishes as the number of struts is increased.

As the *quantitative discussion* of section 7.2 indicates, even if an asymptotic stent cannot be constructed, a periodic stent (with a sufficiently large number of struts) can be designed to produce a distribution of the dose that is very close to that of the asymptotic stent. In view of this analysis, the design of a stent should start from the asymptotic analysis described in section 5. From a numerical point of view, this is simple and requires only two-dimensional simulations. Given a bioactive material, the *lower therapeutic bound* on the dose, a fixed target area, and a deployment ratio ρ, the asymptotic stent would yield an estimate of the minimum value of the required mass surface density c_s (that is, a minimum total mass of bioactive material). The coating thickness of this material is easily obtained from the required total mass and the area of the target surface. Using the parameters obtained from the asymptotic design, the next step is to choose the number of struts N of the periodic stent of section 4. The numerical experiments of Figure 6.2 show that, for the same mass surface density c_s and ratio ρ, the dose of the periodic stent with a finite number of struts always lies above the dose for the asymptotic stent. Hence the same c_s can be used for an N-strut stent to achieve the same lower therapeutic bound. Then the

designer can adjust the number of struts and possibly reduce c_s in order to keep the dose under the upper therapeutic bound. This problem is two-dimensional and shares with the asymptotic stent the same geometrical simplicity.

As is often the case in design problems associated with complex systems, a simple operational model incorporating sensitive design parameters and retaining the essential features of the system can be extremely useful in quickly and economically identifying a good *suboptimal design*. Our analysis and design approach fall into that category and readily extend to more complex periodic stents (cf. section 5.3). As a better understanding of the biochemical and biophysical mechanisms in the wall becomes available, it will be incorporated into the model and the formulation of the design objectives in order to sharpen our results and improve the design process.

REFERENCES

[1] O. F. Bertand, R. Mongrain, J. F. Tanguay, and L. Bilodeau, *Radioactivity Local Delivery System*, PCT patent WO 97/38730, October, 1997.

[2] A. Chalifour and M. C. Delfour, *Optimal distribution of larvicide in running waters*, SIAM J. Optim., 2 (1992), pp. 264–303.

[3] D. Cioranescu and F. Murat, *Un terme étrange venu d'ailleurs* [A strange term brought from somewhere else], in Nonlinear Partial Differential Equations and Their Applications (Collège de France Seminar, Paris, 1979/1980), Pitman Res. Notes Math. 60, Vol. II, Pitman, Boston, London, 1982, pp. 98–138, 389–390 (in French with English summary).

[4] D. Cioranescu and F. Murat, *Un terme étrange venu d'ailleurs.* II [A strange term brought from somewhere else. II], in Nonlinear Partial Differential Equations and Their Applications (Collège de France Seminar, Paris, 1980/1981), Pitman Res. Notes Math. 70, Vol. III, Pitman, Boston, London, 1982, pp. 154–178, 425–426 (in French with English summary).

[5] A. Damlamian, *Le problème de la passoire de Neumann* [The Neumann sieve problem], Rend. Sem. Mat. Univ. Politec. Torino, 43 (1985), pp. 427–450 (in French).

[6] M. C. Delfour and J.-P. Zolésio, *Shapes and Geometries: Analysis, Differential Calculus and Optimization*, Adv. Des. Control 4, SIAM, Philadelphia, 2001.

[7] T. Del Vecchio, *The thick Neumann's sieve*, Ann. Mat. Pura Appl. (4), 147 (1987), pp. 363–402.

[8] J. Horváth, *Topological Vector Spaces and Distributions*, Vol. I, Addison-Wesley, Reading, MA, 1966.

[9] S. Julien, *Étude numérique de la dispersion et la convection du* ^{45}Ca-*DTPA émis par un stent pour le contrôle de la resténose*, Master of Applied Sciences thesis, Mémoire de maîtrise ès sciences appliquées, Génie Mécanique, École Polytechnique, Montréal, 2000.

[10] S. Julien, A. Garon, and J. Manseau, *Numerical simulation of local mass transfer from an endovascular device*, in Proceedings of the 35th ASME National Heat Transfer Conference, Anaheim, CA, 2002, pp. 1–7.

[11] J. Manseau, *Étude numérique d'un modèle de transport de macromolécules à travers la paroi artérielle*, Master of Applied Sciences thesis, Mémoire de maîtrise, École Polytechnique, Montréal, 2002.

[12] V. A. Marchenko, V. A. Marčenko, and E. A. Khruslov, *Kraevye zadachi v oblastyakh s melkozernistoĭgranitseĭ* [Boundary value problems in domains with a fine-grained boundary], Izdat. "Naukova Dumka," Kiev, 1974 (in Russian).

[13] E. Sánchez-Palencia, *Boundary value problems in domains containing perforated walls*, in Nonlinear Partial Differential Equations and Their Applications (Collège de France Seminar, Paris, 1980/1981), Pitman Res. Notes Math. 70, Vol. III, Pitman, Boston, London, 1982, pp. 309–325.

SIAM J. APPL. MATH.
Vol. 65, No. 3, pp. 882–940

A MODEL OF CONTINUOUS SEDIMENTATION OF FLOCCULATED SUSPENSIONS IN CLARIFIER-THICKENER UNITS*

RAIMUND BÜRGER†, KENNETH H. KARLSEN‡, AND JOHN D. TOWERS§

Abstract. The chief purpose of this paper is to formulate and partly analyze a new mathematical model for continuous sedimentation-consolidation processes of flocculated suspensions in clarifier-thickener units. This model appears in two variants for cylindrical and variable cross-sectional area units, respectively (Models 1 and 2). In both cases, the governing equation is a scalar, strongly degenerate parabolic equation in which both the convective and diffusion fluxes depend on parameters that are discontinuous functions of the depth variable. The initial value problem for this equation is analyzed for Model 1. We introduce a simple finite difference scheme and prove its convergence to a weak solution that satisfies an entropy condition. A limited analysis of steady states as desired stationary modes of operation is performed. Numerical examples illustrate that the model realistically describes the dynamics of flocculated suspensions in clarifier-thickeners.

Key words. clarifier-thickener units, discontinuous flux, degenerate diffusion, uniqueness, stationary solutions, finite difference scheme, numerical simulation

AMS subject classifications. 35K65, 35L65, 65M06, 76T20

DOI. 10.1137/04060620X

1. Introduction.

Continuously operated clarifier-thickener units for the solid-liquid separation of suspensions are widely used in chemical engineering, mineral processing, the pulp-and-paper and food industries, and wastewater treatment. For many purposes, spatially one-dimensional mathematical models of these units are sufficient. They are usually based on the kinematic sedimentation theory by Kynch [62], which describes the batch settling of a so-called ideal suspension of small, equal-sized rigid spheres in a viscous fluid by the conservation law $u_t + b(u)_x = 0$ for the solids volume fraction u as a function of depth x and time t. The material-specific properties of the suspension are described by the Kynch batch flux density function $b(u)$. If a global conservation of mass principle is taken into account, then the extension of this theory to clarifier-thickener units leads to a conservation law with a flux that depends discontinuously on x, since the suspension feed flow is split into upwards- and downwards-directed bulk flows into the clarification and thickening zones, respectively. The discontinuous flux makes the well-posedness analysis and numerical simulation of the clarifier-thickener model difficult.

As is well known, the solution of the conservation law arising from the kinematic theory propagates along characteristics, which are straight lines in cylindrical vessels. However, most suspensions are not ideal; rather, they consist of small flocs, or as we

*Received by the editors April 2, 2004; accepted for publication (in revised form) August 23, 2004; published electronically March 31, 2005. This work was supported by the Collaborative Research Center (Sonderforschungsbereich) 404 at the University of Stuttgart, the BeMatA program of the Research Council of Norway, and the European network HYKE, funded by the EC as contract HPRN-CT-2002-00282.

http://www.siam.org/journals/siap/65-3/60620.html

†Institute of Applied Analysis and Numerical Simulation, University of Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany (buerger@mathematik.uni-stuttgart.de).

‡Centre of Mathematics for Applications (CMA), University of Oslo, P.O. Box 1053, Blindern, N–0316 Oslo, Norway (kennethk@math.uio.no). The research of this author was supported in part by an Outstanding Young Investigators Award from the Research Council of Norway.

§MiraCosta College, 3333 Manchester Avenue, Cardiff-by-the-Sea, CA 92007-1516 (jtowers@cts.com).

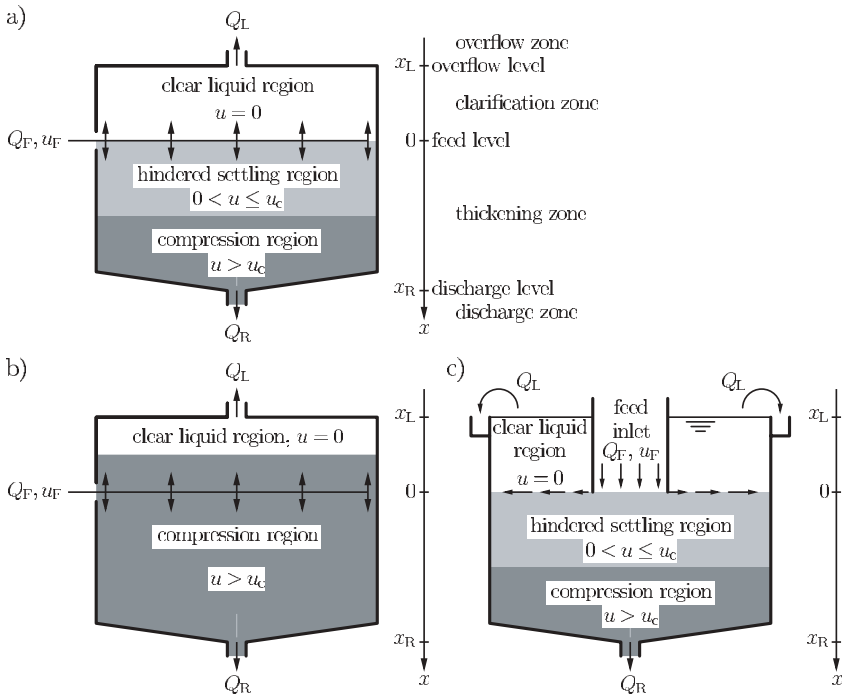


FIG. 1.1. *Clarifier-thickener units treating a flocculated suspension:* (a) *steady-state operation in conventional mode,* (b) *steady-state operation in high-rate mode,* (c) *a variant of the clarifier-thickener setup with a vertical feed inlet.*

say, they are *flocculated*. These mixtures include inorganic slurries such as tailings from mineral processing, which are flocculated artificially in order to enhance settling rates, as well as biological sludges in wastewater treatment. They form compressible sediment layers, which are characterized by curved isoconcentration lines in settling columns, and can therefore not be predicted by the kinematic theory. Instead, an extended dynamic model including pore pressure and effective solids stress has to be used. Such a model is provided by a theory of sedimentation-consolidation processes [10, 27], whose governing equation (if the model is reduced to one space dimension) is a quasi-linear degenerate parabolic equation, which degenerates into the equation of first-order hyperbolic type of the kinematic sedimentation model when $u \le u_c$, where u_c is a material-dependent *critical concentration* or *gel point* at which the solid flocs start to touch each other.

It is the purpose of this paper to present and analyze a clarifier-thickener model for flocculated suspensions as a combination of the first-order models for ideal suspensions with the sedimentation-consolidation theory, which contributes a strongly degenerate diffusion term. The proposed model consists of an initial value problem for a strongly degenerate parabolic PDE, in which both the convective flux and the diffusion flux depend discontinuously on the spatial variable x.

To be more precise, we consider a continuously operated axisymmetric clarifier-thickener vessel as drawn in two variants in Figures 1.1(a) and (b) and Figure 1.1(c), respectively. Throughout this paper, we assume that all flow variables depend on depth x and time t only. This means in particular that u is assumed to be constant across each horizontal cross section. We subdivide the vessel into four different zones:

the thickening zone ($0 < x < x_{\mathrm{R}}$), which is usually the unique zone considered in conventional analyses of continuous sedimentation, the clarification zone ($x_{\mathrm{L}} < x < 0$) located above, the underflow zone ($x > x_{\mathrm{R}}$), and the overflow zone ($x < x_{\mathrm{L}}$). The vessel is continuously fed at depth $x = 0$, the feed level, with fresh feed suspension at a volume feed rate $Q_{\mathrm{F}}(t) \geq 0$. The concentration of the feed suspension is $u_{\mathrm{F}}(t)$. The prescribed volume underflow rate, at which the thickened sediment is removed from the unit, is $Q_{\mathrm{R}}(t) \geq 0$. Consequently, the overflow rate is $Q_{\mathrm{L}}(t) = Q_{\mathrm{R}}(t) - Q_{\mathrm{F}}(t)$, where we assume that the two control functions $Q_{\mathrm{F}}(t)$ and $Q_{\mathrm{R}}(t)$ are chosen such that $Q_{\mathrm{L}}(t) \leq 0$. Of course, the solids concentrations in the underflow and overflow cannot be prescribed and are part of the solution. Furthermore, we distinguish between the four abovementioned *zones* in the clarifier-thickener, which are a property of the equipment modeled, and the clear liquid, hindered settling, and compression regions, in which a suspension at a given point of time has the concentrations zero, $0 < u \leq u_{\mathrm{c}}$, and $u > u_{\mathrm{c}}$, respectively. Thus, the time-dependent location of the regions is a property of a particular flow, that is, of the solution to the problem. Finally, let us mention that the hypothetical assumption $Q_{\mathrm{F}} < 0$ would mean that material is suctioned from rather than injected into the unit (as corresponding to our assumption $Q_{\mathrm{F}} \geq 0$). This case is not included in the present analysis.

The model includes two different stationary modes of operation that are usually distinguished in the applicative literature [34]: *conventional operation*, as shown in Figure 1.1(a), when the sediment level (where $u = u_{\mathrm{c}}$) is located below the feed level, and *high-rate* (also known as *high-capacity*) operation (Figure 1.1(b)), when one lets the sediment level (and thus the compression region) rise into the clarification zone. In the latter mode of operation, practitioners observe that the concentration above the compression region usually is zero. These distinctions are made in engineering applications, and we will show that both modes are captured by the model which we analyze in this paper. Figure 1.1(c) shows a variant of the clarifier-thickener setup of Figures 1.1(a) and (b), in which the feed flow enters the vessel from above through a feed inlet. Note that the feed inlet will usually occupy some of the cross-sectional area of the vessel. We assume that the vessel drawn in Figure 1.1(c) is controlled by regulating the feed flow Q_{F} and the discharge flow Q_{R}, such that no active control of the overflow rate Q_{L} is necessary. In any circumstance we consider a submerged feed source at a fixed vertical location. The notion "high rate" stems from the observation that this mode of operation usually permits higher solids throughput than the conventional mode, since the clarification zone can handle part of the solids feed flux. Capacity and design calculations based on the new model are, however, outside the scope of this paper. For the sake of simplicity, we also neglect the action of the rake provided in most industrial thickeners, which rotates above the gently sloped floor of the thickener to move the concentrated sediment towards the discharge opening.

Similar clarifier-thickener models were proposed by several authors including Barton, Li, and Spencer [6], Chancelier, Cohen de Lara, and Pacard [30], and Lev, Rubin, and Sheintuch [64]. All available treatments are, however, limited to the case of an ideal (nonflocculated) suspension, which is included as a special case in our analysis. In addition, we point out that in [30] the problem of flux discontinuities is circumvented by smoothing out the flux in small ε-neighborhoods of the flux around the levels zero and x_{R} (in our notation). However, uniqueness for $\varepsilon \to 0$ is proved in [30] for steady-state solutions only. Important contributions to the analysis and the determination of solutions to clarifier-thickener models for ideal suspensions have been made by Diehl [39, 40, 41, 42, 43], in which local-in-time existence and uniqueness

results for problems with piecewise constant initial data are obtained [39, 40, 41] and stationary solutions are completely classified [41, 43]. Numerical simulations using a Godunov-type scheme are presented in [40, 41, 42]. The paper [34] presents a limited discussion of a steady-state clarifier-thickener model for flocculated suspensions that has many features in common with the one presented here but is incomplete in that boundary conditions or flux transitions at the discharge level are lacking.

In a recent series of papers [19, 21, 23, 25] the authors with collaborators have initiated an activity aiming at providing a firm rigorous ground of mathematical (existence and uniqueness) and numerical analysis for the first-order clarifier-thickener models. Roughly speaking, the main ingredient in these clarifier-thickener models is a first-order scalar conservation law of the type

$$u_t + f\big(\boldsymbol{\gamma}(x), u\big)_x = 0, \tag{1.1}$$

where the (with respect to u, nonconvex) flux f and the discontinuous vector-valued coefficient $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ are given functions. As is well known, independently of the smoothness of $\boldsymbol{\gamma}(x)$, solutions to (1.1) are in general not smooth, and weak solutions must be sought. Moreover, discontinuous weak solutions are in general not uniquely determined by their initial data. Consequently, an entropy condition must be imposed to single out the physically correct solution. These "physically relevant" solutions are called entropy weak solutions. When $\boldsymbol{\gamma}$ is smooth, Kružkov's theory [61] ensures the existence of a unique and stable entropy weak solution to (1.1). Kružkov's theory does not apply when $\boldsymbol{\gamma}$ is discontinuous. In our previous work cited above, which culminated in [25], we suggested using a variant of Kružkov's notion of entropy weak solution for (1.1) that accounts for the discontinuities in $\boldsymbol{\gamma}$. Moreover, we proved existence and uniqueness (stability) of such entropy weak solutions in a certain functional class. The existence of such solutions was a consequence of convergence results for various numerical schemes such as front tracking [19], a relaxation scheme [21], and upwind difference schemes [23, 25].

The papers [19, 21, 23, 25] were inspired by previous work in the area of conservation laws with discontinuous fluxes. Due to their many applications, this is an area that has enjoyed a lot of interest in recent years [2, 5, 9, 12, 39, 40, 48, 49, 50, 51, 52, 57, 59, 60, 65, 66, 67, 70, 72, 73, 75, 76, 77, 78]. (This list is not complete.) Without entering into too many details, let us just mention that the usual way to cope with the discontinuous parameter $\boldsymbol{\gamma}(x)$ is to express it as an additional conservation law $\boldsymbol{\gamma}_t = 0$, which yields a system of conservations laws for the "unknowns" $(\boldsymbol{\gamma}, u)$. The equation $\boldsymbol{\gamma}_t = 0$ introduces linearly degenerate fields in this system with eigenvalues that are zero. Consequently, if f_u is zero at some points $(\boldsymbol{\gamma}, u)$, then the system is nonstrictly hyperbolic and it experiences so-called nonlinear resonant behavior, which means more complicated wave interactions than in strictly hyperbolic systems. Indeed, one cannot in general expect to bound the total variation of the conserved quantities directly but only when measured under a certain singular mapping, as was done first in [76] for a related system. An alternative "scalar" approach in which $\boldsymbol{\gamma}$ is not treated as a separate unknown is presented in [52, 54, 55, 57, 77, 78] and further developed in [21, 23, 25] in the context of the first-order clarifier-thickener models. If we took the model studied herein and discretized the discontinuity vector $\boldsymbol{\gamma}(x)$ as an additional conservation law $\boldsymbol{\gamma}_t = 0$, then we should expect similar nonlinear resonance phenomena as known for first-order systems, since our model degenerates to first-order type on a solution value interval (u-interval) of positive length.

The main ingredient in the models suggested herein, which accounts for compression effects, is not a first-order equation like (1.1) but rather a second-order strongly

degenerate parabolic (or mixed hyperbolic-parabolic) equation of the type

$$u_t + f\big(\boldsymbol{\gamma}(x), u\big)_x = (\gamma_1(x)A(u)_x)_x, \tag{1.2}$$

where $A(\cdot)$ is nondecreasing with $A(0) = 0$. Note that $A(\cdot)$ can have "flat" regions, and thus (1.2) is strongly degenerate. As a consequence, independently of the smoothness of $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$, solutions to (1.2) will in general be discontinuous, and it becomes necessary to work within a framework of entropy weak solutions also for (1.2). In the case of smooth coefficients, the general mathematical theory of hyperbolic conservation laws was developed more than 30 years ago. On the other hand, the mathematical theory for strongly degenerate parabolic equations (with smooth coefficients) has advanced significantly only in the last few years [7, 8, 28, 31, 32, 53, 68, 69, 82, 83, 84]. (This list is not complete either.) Although conservation laws with discontinuous fluxes are well studied by now, strongly degenerate parabolic equations with discontinuous fluxes are much less studied. In fact, the only papers that we are aware of are [54, 55, 56], among which the latter two are of importance for the present paper. In [55] equations like (1.2) are studied with a concave convective flux $u \mapsto f(\boldsymbol{\gamma}(x), u)$ and with $(\gamma_1(x)A(u)_x)_x$ replaced by $A(u)_{xx}$. Existence of an entropy weak solution is established by proving convergence of a difference scheme of the type discussed in this paper. Uniqueness and stability issues for entropy weak solutions are studied in [56] for a particular class of equations.

Herein we develop further the methods used in [25, 55, 56] in order to apply them to our new mathematical model for the dynamics of flocculated suspensions in clarifier-thickener units. The new results of this paper can be summarized as follows. First, we introduce a suitable definition of entropy weak solutions for the model variant with constant cross-sectional area (to which the mathematical and numerical analysis is limited). We argue that the x-discontinuity of the diffusion term $(\gamma_1(x)A(u)_x)_x$ requires an additional condition in this definition, which states that $A(u)$ is continuous across the jumps of γ_1 (in our model γ_1 is the characteristic function on an interval $(x_{\mathrm{L}}, x_{\mathrm{R}})$). Support for the necessity to state this condition comes from analyses of similar equations for two-phase flow in heterogeneous porous media, in which a similar condition is stated, and from the uniqueness analysis of our problem, which is the second novel point and in particular relies on this condition. Third, we formulate a simple finite difference scheme for the clarifier-thickener and prove its convergence by a compactness analysis. A feature of the compactness analysis is that the discontinuities in the fluxes make it hard to bound the total variation of the conserved variable. Instead, we introduce a particular nonlinear functional under which we are able to bound the total variation. We show that the limit element satisfies all parts of the definition of entropy weak solutions, except for the continuity of $A(u)$. This issue is left as an open problem. Fourth, we present an analysis of admissible stationary solutions based on the discussion of entropy weak solutions of the stationary ODE variant of the governing PDE of the transient model and, finally, a limited selection of numerical examples illustrating the clarifier-thickener model. Both the steady-state analysis and the numerical simulations support the view that it is reasonable to require $A(u)$ to be continuous.

Before outlining the remainder of this paper, let us briefly mention that we do not explicitly include the effect of hydrodynamic diffusion. This effect is also omitted in the majority of clarifier-thickener papers by other authors [6, 30, 34, 39, 40, 41, 42, 43] but is included in the analyses by Lev, Rubin, and Sheintuch [64] and Verdickt et al. [81]. A profound justification of the omission of hydrodynamic diffusion is

beyond the scope of this paper but is provided in section 7.4 of [10] on the basis of practical limitations, theoretical considerations, computational comparisons, and experimental results. If we had decided to include hydrodynamic diffusion by adding a term, say μu_{xx}, with $\mu > 0$ to the right-hand side of (1.2), then the resulting governing PDE would possess smooth solutions and allow for a simpler analytical and numerical treatment than the one advanced in this work. In essence, the discontinuities appearing in transient solutions would be blurred, and in Remark 5.4 we discuss how hydrodynamic diffusion affects steady states for the clarifier-thickener problem. Finally, let us mention that hydrodynamic diffusion is not explicitly modeled but is in a sense implicitly present in our model, since our concept of Kružkov entropy weak solution is equivalent to stating that our solution is obtained in the limit $\mu \to 0$ of smooth solutions of strictly parabolic equations with regularizing (hydrodynamic) diffusion term μu_{xx}. See section 4.3.

The remainder of this paper is organized as follows. In section 2, the clarifier-thickener model is derived. We consider two variants for units with constant and variable interior cross-sectional area, respectively (Models 1 and 2). In particular, we incorporate the governing equation of the sedimentation-consolidation theory developed in full detail in [10]. We describe in section 3 the finite difference scheme for the simulation of both models. The scheme appears in two variants, an explicit one which also is analyzed, and a semi-implicit one for which a less restrictive condition for the time step size is valid, and which therefore is suitable for large-time simulations. In section 4 we analyze the initial value problem for Model 1, relying on our previous efforts [25, 55, 56]. A definition of entropy weak solutions is given (and discussed extensively), jump and entropy jump conditions are derived, and uniqueness of entropy weak solutions is proved. We study the explicit difference scheme described in section 3 and prove compactness of a family of approximate solutions generated by this difference scheme. We prove that the limit function u is a weak solution of Model 1 that satisfies the entropy condition. The question of whether $A(u)$ is continuous for this limit function is left open. An important problem for practitioners are steady-state solutions, which correspond to the normal operation of a clarifier-thickener unit for constant feed and discharge control parameters. In section 5, we construct steady-state solutions to Model 1 as piecewise continuous solutions to a time-independent ODE version of the transient Model 1. These solutions are again based on the continuity of $A(u)$, but this time this property *follows* from the ODE formulation. Finally, section 6 presents a limited choice of numerical examples illustrating Models 1 and 2.

2. Mathematical model.

2.1. Balance equations. Consider a vessel with a variable cross-sectional area $S(x)$. Since we assume $u = u(x,t)$, the continuity equations for the solids and the fluid are given by

$$S(x)u_t + \big(S(x)uv_{\mathrm{s}}\big)_x = 0, \tag{2.1}$$

$$-S(x)u_t + \big(S(x)(1-u)v_{\mathrm{f}}\big)_x = 0, \tag{2.2}$$

where v_{s} and v_{f} are the solids and the fluid phase velocity, respectively. The mixture flux, that is, the volume average flow velocity weighted with the cross-sectional area at height x, is given by $Q(x,t) := S(x)(uv_{\mathrm{s}} + (1-u)v_{\mathrm{f}})$. The sum of (2.1) and (2.2) produces the continuity equation of the mixture, $Q_x(x,t) = 0$, which implies that $Q(\cdot,t)$ is constant as a function of x. When Q suffers no jumps with respect to x, we obtain $Q(x,t) = Q(x_{\mathrm{R}},t) = Q(t)$. This equation is equivalent to one of the mass

balance equations. We let it replace (2.2) and rewrite (2.1) in terms of the flow rate $Q(t)$ and the solid-fluid relative velocity or slip velocity $v_{\mathrm{r}} := v_{\mathrm{s}} - v_{\mathrm{f}}$, for which a constitutive equation will be formulated. Observing that

$$\begin{aligned} S(x)uv_{\mathrm{s}} &= S(x)\big[\big(uv_{\mathrm{s}} + (1-u)v_{\mathrm{f}}\big)u + u(1-u)(v_{\mathrm{s}} - v_{\mathrm{f}})\big] \\ &= Q(t)u + S(x)u(1-u)v_{\mathrm{r}}, \end{aligned} \tag{2.3}$$

we obtain from (2.1) the equation

$$S(x)u_t + \big(Q(t)u + S(x)u(1-u)v_{\mathrm{r}}\big)_x = 0. \tag{2.4}$$

The kinematic sedimentation theory [62] is based on the assumption that v_{r} is a function of u only, $v_{\mathrm{r}} = v_{\mathrm{r}}(u)$. However, the slip velocity is usually expressed in terms of the Kynch batch flux density function $b(u)$, such that $v_{\mathrm{r}}(u) = b(u)/(u(1-u))$ and (2.4) takes the form

$$S(x)u_t + \big(Q(t)u + S(x)b(u)\big)_x = 0. \tag{2.5}$$

The function b is usually assumed to be piecewise differentiable with $b(u) = 0$ for $u \le 0$ or $u \ge u_{\max}$, where $u_{\max}$ is the maximum solids concentration, $b(u) > 0$ for $0 < u < u_{\max}$, $b'(0) > 0$, and $b'(u_{\max}) \le 0$. A typical example is [74]

$$b(u) = v_\infty u(1-u)^C \quad \text{if } 0 < u < u_{\max}, \quad b(u) = 0 \quad \text{otherwise}, \tag{2.6}$$

where $C \ge 1$ and $v_\infty > 0$ is the settling velocity of a single floc in pure fluid. It should be pointed out that in the presence of diffusion terms modeling compression effects, to be introduced later, the maximum concentration attained in a settling system depends on the balance between convection and diffusion terms but not critically on the choice of $u_{\max}$. In order to facilitate the analysis, we assume in this paper that $u_{\max} = 1$ and that $b(u)$ is smooth on $[0,1]$.

We now apply the sedimentation-consolidation theory outlined in [10, 27] to include the sediment compressibility. By constitutive assumptions, a dimensional analysis, and considering one space dimension only, this theory leads to the following equation for the relative velocity v_{r}, which plays the role of one of the linear momentum balances:

$$v_{\mathrm{r}} = v_{\mathrm{r}}(u, u_x) = \frac{b(u)}{u(1-u)}\left(1 - \frac{\sigma_{\mathrm{e}}'(u)}{\Delta\varrho g u}u_x\right), \tag{2.7}$$

where $\Delta\varrho > 0$ denotes the solid-fluid density difference, g the acceleration of gravity, and $\sigma_{\mathrm{e}}(u)$ is the effective solid stress function, which is now the second constitutive function (besides b) characterizing the suspension. This function is assumed to satisfy $\sigma_{\mathrm{e}}(u) \ge 0$ for all u and

$$\sigma_{\mathrm{e}}'(u) := \frac{d\sigma_{\mathrm{e}}(u)}{du}\begin{cases} = 0 & \text{for } u \le u_{\mathrm{c}}, \\ > 0 & \text{for } u > u_{\mathrm{c}}. \end{cases} \tag{2.8}$$

A commonly used semiempirical effective stress formula is the power law

$$\sigma_{\mathrm{e}}(u) = 0 \quad \text{for } u \le u_{\mathrm{c}}; \quad \sigma_{\mathrm{e}}(u) = \sigma_0\big((u/u_{\mathrm{c}})^k - 1\big) \quad \text{for } u > u_{\mathrm{c}}, \tag{2.9}$$

with parameters $\sigma_0 > 0$ and $k > 1$. Note that the derivative $\sigma_e'(u)$ of the function defined in (2.9) is in general discontinuous at $u = u_c$. Inserting (2.7) into (2.4) and defining

$$a(u) := \frac{b(u)\sigma_e'(u)}{\Delta\varrho g u}, \quad A(u) := \int_0^u a(s)\,ds, \tag{2.10}$$

we obtain the governing equation

$$\big(S(x)u\big)_t + \big(Q(t)u + S(x)b(u)\big)_x = \big(S(x)A(u)_x\big)_x. \tag{2.11}$$

Since $a(u) = 0$ for $u \le u_c$ and $u = u_{\max}$ and $a(u) > 0$ otherwise, (2.11) is first-order hyperbolic for $u \le u_c$ and second-order parabolic for $u > u_c$. Since (2.11) degenerates into hyperbolic type on a solution value interval of positive length, (2.11) is called strongly degenerate parabolic. The location of the type-change interface $u = u_c$ (the sediment level) is in general unknown beforehand.

For the determination of appropriate functions b and σ_e for real materials, see [15, 16, 45]. Moreover, the sedimentation-consolidation model is equivalent to the suspension dewatering theory employed in [4, 38, 63, 79] and other works by the same group of authors.

2.2. The clarifier-thickener model. In the present model, the volume bulk flows are $Q(x,t) = Q_R(t)$ for $x > 0$ and $Q(x,t) = Q_L(t)$ for $x < 0$. This suggests employing (2.11) with $Q(t) = Q_R(t)$ for $0 < x < x_R$ and $Q(t) = Q_L(t)$ for $x_L < x < 0$. Furthermore, we assume that in the overflow and underflow zones, the solid material is transported with the same velocity as the liquid. This means that the solid-fluid relative velocity vanishes, $v_r = 0$. Moreover, the cross-sectional area $S(x)$ needs to be positive outside the interval $[x_L, x_R]$. We assume that $S(x) = S_0$ for $x < x_L$ and $x > x_R$, where $S_0 > 0$ is a small but positive pipe diameter. From (2.3) we now read off that

$$S(x)uv_s|_{x\notin[x_L,x_R]} = S_0 u v_s = \begin{cases} Q_L(t)u & \text{for } x < x_L, \\ Q_R(t)u & \text{for } x > x_R. \end{cases} \tag{2.12}$$

The feed mechanism is introduced by adding a singular source term to the right-hand side of the solids continuity equation (2.1). If we prescribe an initial concentration u_0 in the vessel, we can summarize the resulting dynamic model as

$$S(x)u_t + \tilde{G}(x,t,u)_x = \big(\gamma_1(x)A(u)_x\big)_x + Q_F(t)u_F(t)\delta(x), \quad x \in \mathbb{R}, \quad t > 0, \tag{2.13}$$

$$u(x,0) = u_0(x), \quad x \in \mathbb{R}, \quad u_0(x) \in [0, u_{\max}], \tag{2.14}$$

$$\tilde{G}(x,t,u) = S(x)uv_s = \begin{cases} Q_L(t)u & \text{for } x < x_L, \\ Q_L(t)u + S(x)b(u) & \text{for } x_L < x < 0, \\ Q_R(t)u + S(x)b(u) & \text{for } 0 < x < x_R, \\ Q_R(t)u & \text{for } x > x_R, \end{cases} \tag{2.15}$$

$$\gamma_1(x) := \begin{cases} S(x) & \text{if } x_L \le x \le x_R, \\ 0 & \text{otherwise.} \end{cases}$$

For the mathematical analysis we assume that the control functions Q_L, Q_R, and u_F are constant. Finally, we may express the singular source term in (2.13) in terms of the derivative of the Heaviside function. Adding the term $-H(x)Q_F u_F$ to $\tilde{G}(x,u)$

and subtracting the term $Q_{\mathrm{L}}u_{\mathrm{F}}$, which is constant with respect to x, we obtain the strongly degenerate convection-diffusion problem

$$S(x)u_t + G(x,u)_x = \big(\gamma_1(x)A(u)_x\big)_x, \quad x \in \mathbb{R}, \quad t > 0, \tag{2.16}$$

$$u(x,0) = u_0(x), \quad x \in \mathbb{R}, \quad u_0(x) \in [0, u_{\max}], \tag{2.17}$$

$$G(x,u) = \begin{cases} Q_{\mathrm{L}}(u-u_{\mathrm{F}}) & \text{for } x < x_{\mathrm{L}}, \\ Q_{\mathrm{L}}(u-u_{\mathrm{F}}) + S(x)b(u) & \text{for } x_{\mathrm{L}} < x < 0, \\ Q_{\mathrm{R}}(u-u_{\mathrm{F}}) + S(x)b(u) & \text{for } 0 < x < x_{\mathrm{R}}, \\ Q_{\mathrm{R}}(u-u_{\mathrm{F}}) & \text{for } x > x_{\mathrm{R}}. \end{cases} \tag{2.18}$$

2.3. Model 1 (constant interior cross-sectional area). A simple but important subcase is a vessel whose cross-sectional area is constant in the interior; i.e., we consider

$$S(x) = \begin{cases} S_0 & \text{for } x < x_{\mathrm{L}} \text{ and } x > x_{\mathrm{R}}, \\ S_{\mathrm{int}} & \text{for } x_{\mathrm{L}} \le x \le x_{\mathrm{R}}. \end{cases} \tag{2.19}$$

In this case, the solution of (2.16)–(2.18) does not depend on the value of S_0. To see this, we introduce the new space variable $w = w(x)$ defined by the bijective mapping $\mathbb{R} \ni x \mapsto w \in \mathbb{R}$,

$$w(x) := \begin{cases} (S_0/S_{\mathrm{int}})(x - x_{\mathrm{L}}) + x_{\mathrm{L}} & \text{for } x < x_{\mathrm{L}}, \\ x & \text{for } x_{\mathrm{L}} \le x \le x_{\mathrm{R}}, \\ (S_0/S_{\mathrm{int}})(x - x_{\mathrm{R}}) + x_{\mathrm{R}} & \text{for } x > x_{\mathrm{R}}, \end{cases} \tag{2.20}$$

and from (2.16) we infer that the function v defined by $v(w(x),t) = u(x,t)$ satisfies the following initial value problem, where we define the velocities $q_{\mathrm{R}} := Q_{\mathrm{R}}/S_{\mathrm{int}}$, $q_{\mathrm{L}} := Q_{\mathrm{L}}/S_{\mathrm{int}}$ and the diffusion functions $\tilde{a}(\cdot) := a(\cdot)/S_{\mathrm{int}}$, $\mathcal{A}(\cdot) := A(\cdot)/S_{\mathrm{int}}$:

$$v_t + g(w,v)_w = \big(\gamma_1(w)\mathcal{A}(v)_w\big)_w, \quad w \in \mathbb{R}, \quad t > 0, \tag{2.21}$$

$$v(w,0) = u_0\big(w(x)\big), \quad x \in \mathbb{R}, \tag{2.22}$$

$$g(w,v) := \begin{cases} q_{\mathrm{L}}(v-u_{\mathrm{F}}) & \text{for } w < x_{\mathrm{L}}, \\ q_{\mathrm{L}}(v-u_{\mathrm{F}}) + b(v) & \text{for } x_{\mathrm{L}} < w < 0, \\ q_{\mathrm{R}}(v-u_{\mathrm{F}}) + b(v) & \text{for } 0 < w < x_{\mathrm{R}}, \\ q_{\mathrm{R}}(v-u_{\mathrm{F}}) & \text{for } w > x_{\mathrm{R}}. \end{cases} \tag{2.23}$$

We refer to (2.21)–(2.23) as *Model* 1. Since the variation of $S(x)$ at $x = x_{\mathrm{L}}$ and $x = x_{\mathrm{R}}$ no longer appears in (2.21), Model 1 is formally attained by setting $S \equiv 1$ in (2.18) for all $x \in \mathbb{R}$. This significantly facilitates the analysis. Finally, we define the vector of discontinuity parameters

$$\boldsymbol{\gamma} := (\gamma_1, \gamma_2), \quad \gamma_1(w) := \begin{cases} 1 & \text{for } w \in (x_{\mathrm{L}}, x_{\mathrm{R}}), \\ 0 & \text{for } w \notin (x_{\mathrm{L}}, x_{\mathrm{R}}), \end{cases} \qquad \gamma_2(w) := \begin{cases} q_{\mathrm{L}} & \text{for } w < 0, \\ q_{\mathrm{R}} & \text{for } w > 0 \end{cases}$$

and the flux function

$$f\big(\boldsymbol{\gamma}(w), u\big) := g(x,u) = \gamma_1(x)b(u) + \gamma_2(x)(u - u_{\mathrm{F}}). \tag{2.24}$$

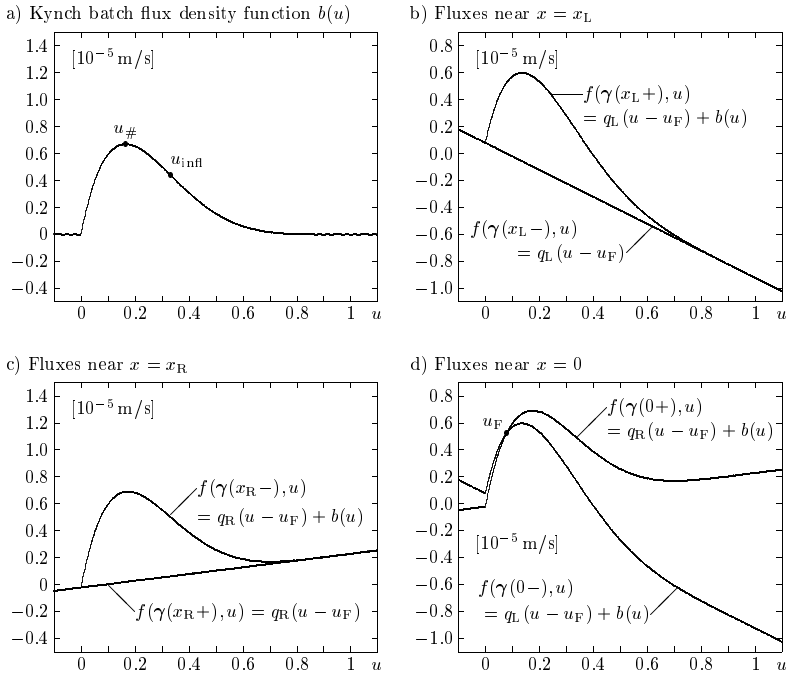


FIG. 2.1. (a) *The Kynch batch flux density function* $b(u)$ *and the fluxes adjacent to* (b) $x = x_{\mathrm{L}}$, (c) $x = x_{\mathrm{R}}$, *and* (d) $x = 0$.

REMARK 2.1. *Consider a nonflocculated ideal suspension for which* $A \equiv 0$. *Then Model* 1 *recovers the clarifier-thickener model with* $S \equiv 1$ *and* $x_{\mathrm{L}} = -x_{\mathrm{R}}$ *we analyzed previously* [19, 21, 22, 25]. *Our derivation now clearly shows that these analyses (including well-posedness and convergence of numerical schemes) are in fact not restricted to the assumption of transport pipes (leading away from the unit for* $x < x_{\mathrm{L}}$ *and* $x > x_{\mathrm{R}}$*) that have the same diameter as the thickening vessel. Rather, by application of the inverse of* (2.20)*, they are also valid for vessels with cylindrical interior and transport pipes of arbitrarily small (but positive) pipe diameter* S_0.

For the function $b(u)$ given by (2.6) with $v_\infty = 10^{-4}\,\mathrm{m/s}$, $C = 5$, the velocities $q_{\mathrm{L}} = -10^{-5}\,\mathrm{m/s}$ and $q_{\mathrm{R}} = 2.5 \times 10^{-6}\,\mathrm{m/s}$, and $u_{\mathrm{F}} = 0.08$, the flux functions $b(u)$ and the fluxes adjacent to the discontinuities of $\boldsymbol{\gamma}$ near $x = x_{\mathrm{L}}$, $x = 0$, and $x = x_{\mathrm{R}}$ are plotted in Figure 2.1. These parameters will also be utilized in sections 5 and 6.

2.4. Model 2 (variable interior cross-sectional area). In the case that $S(x)$ is variable on $(x_{\mathrm{L}}, x_{\mathrm{R}})$, we refer to (2.16)–(2.18) as *Model* 2. It is convenient to rewrite (2.16) as

$$S(x)u_t + F(\boldsymbol{\gamma}(x), u)_x = (\gamma_1(x)A(u)_x)_x$$

and rewrite the flux function $G(x, u)$ as

$$F(\boldsymbol{\gamma}(x), u) := G(x, u) = \gamma_1(x)b(u) + \gamma_2(x)(u - u_{\mathrm{F}}),$$

where

$$\gamma_1(w) := \begin{cases} S(x) & \text{for } x \in (x_{\mathrm{L}}, x_{\mathrm{R}}), \\ 0 & \text{for } x \notin (x_{\mathrm{L}}, x_{\mathrm{R}}), \end{cases} \qquad \gamma_2(w) := \begin{cases} Q_{\mathrm{L}} & \text{for } x < 0, \\ Q_{\mathrm{R}} & \text{for } x > 0. \end{cases}$$

3. Numerical scheme. The numerical scheme for the solution of (2.16)–(2.18) is a straightforward extension of that used in [23] for the first-order variant of (2.16). To define it, choose $\Delta x > 0$, set $x_j := j\Delta x$, and discretize the parameter vector $\boldsymbol{\gamma}$, the initial data, and the cross-sectional area by

$$\boldsymbol{\gamma}_{j+1/2} := \boldsymbol{\gamma}(x_{j+1/2}), \quad U_j^0 := u_0(x_j), \quad S_j := \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} S(x)\,dx.$$

Here $x_{j+1/2} := x_j + \Delta x/2$, i.e., the midpoint in the interval $[x_j, x_{j+1})$. In contrast to [25], we discretize u_0 and $\boldsymbol{\gamma}$ in a pointwise manner, rather than via cell averages. The discretization of u_0 circumvents some analytical difficulties that would otherwise turn up in the proof of Lemma 4.18 and is not unreasonable from a computational point of view. For $n > 0$ we define the approximations according to the explicit marching formula

$$U_j^{n+1} = U_j^n - \lambda_j \Delta_- h\big(\boldsymbol{\gamma}_{j+1/2}, U_{j+1}^n, U_j^n\big) + \frac{\lambda_j}{\Delta x}\Delta_-\big(\gamma_{1,j+1/2}\Delta_+ A(U_j^n)\big), \tag{3.1}$$

where $\lambda_j := \Delta t/(S_j\Delta x)$, $\Delta_- V_j := V_j - V_{j-1}$, $\Delta_+ V_j := V_{j+1} - V_j$, and

$$h(\boldsymbol{\gamma}, v, u) := \frac{1}{2}\left[f(\boldsymbol{\gamma}, u) + f(\boldsymbol{\gamma}, v) - \int_u^v \big|f_u(\boldsymbol{\gamma}, w)\big|dw\right] \tag{3.2}$$

is the Engquist–Osher flux [44]. Let $t_n := n\Delta t$, and let χ^n, χ_j, and $\chi_{j+1/2}$ denote the characteristic functions of the intervals $[t_n, t_{n+1})$, $[x_{j-1/2}, x_{j+1/2})$, and $[x_j, x_{j+1})$, respectively. We then define

$$u^\Delta(x,t) := \sum_{n\ge 0}\sum_{j\in\mathbb{Z}} U_j^n \chi_j(x)\chi^n(t), \quad \boldsymbol{\gamma}^\Delta(x) := \sum_{j\in\mathbb{Z}} \boldsymbol{\gamma}_{j+1/2}\chi_{j+1/2}(x). \tag{3.3}$$

Note that the discontinuity vector $\boldsymbol{\gamma}$ is discretized on a spatial mesh which is staggered (i.e., shifted by $\Delta x/2$) with respect to that of the conserved quantity u. This makes it possible to use the scalar Engquist–Osher function (3.2) for the convective part of the problem. A natural alternative would be to align the two discretizations. However, in that case one would have to solve (exactly or approximately) a Riemann problem for a system of two equations in two variables (namely, for u and the volume average velocity q) at each cell boundary, which makes the resulting numerical method rather complicated; see [48, 59, 60, 65, 66]. In particular, our treatment of the convective part is simpler than the complicated (but accurate) front tracking algorithm used in [19]. Staggering the discretizations also simplifies the analysis, making it possible to apply, with some allowances for the parabolic terms, some of the analytical techniques developed for monotone schemes for purely hyperbolic problems.

Let us recall that the scheme stated here comprises both Model 1 and Model 2 and is employed for numerical examples for both models in section 6. The analysis of the scheme is, however, limited to Model 1 (with $S \equiv 1$).

In section 6 we also use the following semi-implicit variant of (3.1) for large-time computations:

$$U_j^{n+1} = U_j^n - \lambda_j \Delta_- h\big(\boldsymbol{\gamma}_{j+1/2}, U_{j+1}^n, U_j^n\big) + \frac{\lambda_j}{\Delta x}\Delta_-\big(\gamma_{1,j+1/2}\Delta_+ A(U_j^{n+1})\big). \tag{3.4}$$

The scheme (3.4) requires the solution of a system of nonlinear equations in each time step by the Newton–Raphson method. This can be done efficiently by Thomas's algorithm, since the coefficient matrix is tridiagonal. The advantage of using (3.4) lies in the fact that we need only to satisfy a CFL condition requiring that $\Delta t/\Delta x$ is bounded; while (3.1) enforces that $\Delta t/(\Delta x)^2$ be bounded; see Lemmas 4.16 and 4.17.

4. Mathematical analysis. In several instances we will not repeat arguments that are only minor modifications of proofs that have already appeared in [25], [55], or [56]. In various bounds, C denotes a universal constant.

4.1. The initial value problem. For the sake of consistency with our previous papers, we will abuse the notation slightly by replacing the transformed spatial variable w by x, and the transformed functions v and $\mathcal{A}$ by u and A, respectively. The Cauchy problem of interest is then

$$\begin{aligned} &u_t + f(\boldsymbol{\gamma}(x), u)_x = (\gamma_1(x) A(u)_x)_x, \ (x,t) \in \Pi_T := \mathbb{R} \times (0,T); \\ &u(x,0) = u_0(x), \ x \in \mathbb{R}, \end{aligned} \tag{4.1}$$

where $f(\boldsymbol{\gamma}, u) = \gamma_1 b(u) + \gamma_2 (u - u_{\mathrm{F}})$. The parameter vector for this problem is $\boldsymbol{\gamma} := (\gamma_1, \gamma_2)$, where

$$\gamma_1(x) := \begin{cases} 1 & \text{for } x \in (x_{\mathrm{L}}, x_{\mathrm{R}}), \\ 0 & \text{for } x \notin (x_{\mathrm{L}}, x_{\mathrm{R}}), \end{cases} \qquad \gamma_2(x) := \begin{cases} q_{\mathrm{L}} & \text{for } x \le 0, \\ q_{\mathrm{R}} & \text{for } x > 0. \end{cases}$$

We assume that $q_{\mathrm{L}} < 0$ and $q_{\mathrm{R}} > 0$. This rules out the case of batch settling ($q_{\mathrm{L}} = 0$, $q_{\mathrm{R}} = 0$). However, once our analysis is complete, it will be clear that one can accommodate batch processing as a separate case where one restricts the analysis to the interval $[x_{\mathrm{L}}, x_{\mathrm{R}}]$. We assume that $b \in C^2([0,1])$, and $b(0) = b(1) = 0$. Furthermore, we assume that b' vanishes at exactly one location $u_{\#} \in (0,1)$, where the function has a maximum, and that b'' vanishes at no more than one inflection point in $u_{\mathrm{infl}} \in (0,1)$; if such a point is present, we assume that $u_{\mathrm{infl}} \in (u_{\#}, 1)$. For example, $b(u)$ may be given by (2.6). In accordance with (2.9), (2.10), we will assume that $A \in \mathrm{Lip}([0,1])$, $A'(u) = 0$ for $u < u_{\mathrm{c}}$ and that $A'(u) > 0$ for $u \in (u_{\mathrm{c}}, 1)$.

For the initial data, we assume that u_0 satisfies

$$\begin{cases} u_0 \in L^1(\mathbb{R}) \cap BV(\mathbb{R}); \ \ u_0(x) \in [0,1] \ \forall x \in \mathbb{R}; \\ A(u_0) \text{ is absolutely continuous on } [x_{\mathrm{L}}, x_{\mathrm{R}}]; \ \ \gamma_1 A(u_0)_x \in BV(\mathbb{R}). \end{cases} \tag{4.2}$$

In this paper, $\boldsymbol{\gamma}$ is allowed to take values in $\mathcal{G} := \{(q_{\mathrm{L}}, 0), (q_{\mathrm{L}}, 1), (q_{\mathrm{R}}, 0), (q_{\mathrm{R}}, 1)\}$ only. This simplifies matters somewhat compared to our previous paper [25], where the cell average discretization of $\boldsymbol{\gamma}$ required us to consider several lateral sides of the rectangle marked by the vertices in $\mathcal{G}$.

4.2. Definition of entropy weak solution. If (4.1) is allowed to degenerate at certain points, that is, $A'(s) = 0$ for some values of s, solutions are not necessarily smooth, and weak solutions must be sought. This property is independent of the smoothness of $\boldsymbol{\gamma}$. Moreover, weak solutions are not necessarily unique, requiring some additional condition, a so-called entropy condition, to single out the physically meaningful solution.

As in [25], the function space $BV_t(\Pi_T)$ plays an important role in our definition of entropy weak solution. We denote by $\mathcal{M}(\Pi_T)$ the locally finite Radon (signed) measures on Π_T. The space $BV_t(\Pi_T)$ consists of locally integrable functions $W : \Pi_T \to \mathbb{R}$ for which $\partial_t W \in \mathcal{M}(\Pi_T)$.

Let $\mathcal{J} := \{x_{\mathrm{L}}, 0, x_{\mathrm{R}}\}$ denote the set of points where $\boldsymbol{\gamma}$ is discontinuous. For a point $m \in \mathcal{J}$, we use the notation $\boldsymbol{\gamma}(m-)$ and $\boldsymbol{\gamma}(m+)$ for the one-sided limits at the point m:

$$\boldsymbol{\gamma}(m-) := \lim_{x \uparrow m} \boldsymbol{\gamma}(x), \quad \boldsymbol{\gamma}(m+) := \lim_{x \downarrow m} \boldsymbol{\gamma}(x).$$

The following definition is motivated by [25, 55, 56].

DEFINITION 4.1 (BV_t entropy weak solution). *A measurable function $u : \Pi_T \to R$ is a BV_t entropy weak solution of the initial value problem* (4.1) *if it satisfies the following conditions:*

(D.1) $u \in L^1(\Pi_T) \cap BV_t(\Pi_T)$, $u \in [0,1]$ *a.e. in* Π_T, *and* $A(u)_x \in L^\infty((x_{\mathrm{L}}, x_{\mathrm{R}}) \times (0,T))$.

(D.2) *For all test functions* $\phi \in \mathcal{D}(\Pi_T)$

$$\iint_{\Pi_T} \Big(u\phi_t + \big[f\big(\boldsymbol{\gamma}(x), u\big) - \gamma_1(x) A(u)_x \big] \phi_x \Big)\, dx\, dt = 0. \tag{4.3}$$

(D.3) *The initial condition is satisfied in the following strong* L^1 *sense:*

$$\operatorname*{ess\,lim}_{t \downarrow 0} \int_{\mathbb{R}} \big| u(x,t) - u_0(x) \big|\, dx = 0. \tag{4.4}$$

(D.4) *For a.e.* $t \in [0,T]$, $x \mapsto A(u(x,t))$ *is continuous at* $x = x_{\mathrm{L}}$ *and* $x = x_{\mathrm{R}}$.

(D.5) *The following Kružkov-type entropy inequality holds for all* $c \in \mathbb{R}$ *and all test functions* $0 \le \phi \in \mathcal{D}(\Pi_T)$*:*

$$\begin{aligned} \iint_{\Pi_T} \Big(& |u-c| \phi_t + \operatorname{sgn}(u-c) \big[f\big(\boldsymbol{\gamma}(x), u\big) - f\big(\boldsymbol{\gamma}(x), c\big) \big] \phi_x \\ & - \gamma_1(x) \big| A(u) - A(c) \big|_x \phi_x \Big)\, dx\, dt \\ & + \int_0^T \sum_{m \in \mathcal{J}} \big| f\big(\boldsymbol{\gamma}(m+), c\big) - f\big(\boldsymbol{\gamma}(m-), c\big) \big| \phi(m,t)\, dt \ge 0. \end{aligned} \tag{4.5}$$

A function $u : \Pi_T \to \mathbb{R}$ satisfying conditions (D.1), (D.2), *and* (D.3) *is called a BV_t weak solution of the initial value problem* (4.1).

4.3. Comments on the entropy weak solution concept.

4.3.1. Motivation of the entropy condition (4.5). It is possible to motivate the entropy condition (4.5) by the fact that limit functions constructed by the method of vanishing viscosity, in combination with smoothing of the coefficients, satisfy this condition. To this end, let $\boldsymbol{\gamma}_\varepsilon(x) := (\gamma_{1,\varepsilon}, \gamma_{2,\varepsilon})$ be a C^∞ approximation of $\boldsymbol{\gamma}$, such that $\boldsymbol{\gamma}_\varepsilon$ equals $\boldsymbol{\gamma}$ everywhere except on $(m-\varepsilon, m+\varepsilon)$ for $m \in \{x_{\mathrm{L}}, 0, x_{\mathrm{R}}\}$, and such that the sign of $\boldsymbol{\gamma}_\varepsilon'$ is constant on $(m-\varepsilon, m+\varepsilon)$ for $m \in \{x_{\mathrm{L}}, 0, x_{\mathrm{R}}\}$. Since $\boldsymbol{\gamma} \mapsto f(\boldsymbol{\gamma}, c)$ is linear, $f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}, c)$ does not depend on x. For each $\mu, \varepsilon > 0$, let $u^{\mu,\varepsilon}$ be a classical solution to the uniformly parabolic equation

$$u_t^{\mu,\varepsilon} + f\big(\boldsymbol{\gamma}_\varepsilon(x), u^{\mu,\varepsilon}\big)_x = \big(\gamma_{1,\varepsilon}(x) A(u^{\mu,\varepsilon})_x\big)_x + \mu u_{xx}^{\mu,\varepsilon}. \tag{4.6}$$

Let us suppose that $u^{\mu,\varepsilon} \to u$ in L^p_{loc} for any $1 \le p < \infty$. The rest of this remark is devoted to proving that the limit u satisfies the entropy condition (4.5).

Since $\boldsymbol{\gamma}_\varepsilon$ is smooth, the chain rule and the convexity of η imply that $u^{\mu,\varepsilon}$ satisfies the following entropy condition for all convex C^2 functions $\eta : \mathbb{R} \to \mathbb{R}$:

$$\begin{aligned} & \eta(u^{\mu,\varepsilon})_t + q\big(\boldsymbol{\gamma}_\varepsilon(x), u^{\mu,\varepsilon}\big)_x - \big(\gamma_{1,\varepsilon}(x) r(u^{\mu,\varepsilon})_x\big)_x - \mu \eta(u^{\mu,\varepsilon})_{xx} \\ & \qquad + \boldsymbol{\gamma}_\varepsilon'(x) \cdot \big(\eta'(u^{\mu,\varepsilon}) f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_\varepsilon(x), u^{\mu,\varepsilon}) - q_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_\varepsilon(x), u^{\mu,\varepsilon}) \big) \le 0 \quad \text{in } \mathcal{D}'(\Pi_T), \end{aligned} \tag{4.7}$$

where $q : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ and $r : \mathbb{R} \to \mathbb{R}$ are defined, respectively, by

$$q_u(\boldsymbol{\gamma}, u) = \eta'(u) f_u(\boldsymbol{\gamma}, u), \qquad r'(u) = \eta'(u) A'(u).$$

By a standard approximation argument, (4.7) implies that the following Kružkov-type entropy condition holds for any constant $c \in \mathbb{R}$:

$$
\begin{aligned}
&|u^{\mu,\varepsilon}-c|_t + \big[\operatorname{sgn}(u^{\mu,\varepsilon}-c)\big(f(\boldsymbol{\gamma}_\varepsilon(x),u^{\mu,\varepsilon})-f(\boldsymbol{\gamma}_\varepsilon(x),c)\big)\big]_x \\
&\qquad - \big(\gamma_{1,\varepsilon}\big|A(u^{\mu,\varepsilon})-A(c)\big|_x\big)_x \\
&\qquad - \mu|u^{\mu,\varepsilon}-c|_x + \operatorname{sgn}(u^{\mu,\varepsilon}-c) f\big(\boldsymbol{\gamma}_\varepsilon(x),c\big)_x \le 0 \quad \text{in } \mathcal{D}'(\Pi_T),
\end{aligned} \tag{4.8}
$$

where

$$
f(\boldsymbol{\gamma}_\varepsilon(x),c)_x = \boldsymbol{\gamma}_\varepsilon'(x)\cdot f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_\varepsilon(x),c) = \gamma_{1,\varepsilon}'(x) f_{\gamma_{1,\varepsilon}}(\boldsymbol{\gamma}_\varepsilon(x),c) + \gamma_{2,\varepsilon}'(x) f_{\gamma_{2,\varepsilon}}(\boldsymbol{\gamma}_\varepsilon(x),c).
$$

From (4.8) it is clear that the following "less precise" inequality holds:

$$
\begin{aligned}
&|u^{\mu,\varepsilon}-c|_t + \big[\operatorname{sgn}(u^{\mu,\varepsilon}-c)\big(f(\boldsymbol{\gamma}_\varepsilon(x),u^{\mu,\varepsilon})-f(\boldsymbol{\gamma}_\varepsilon(x),c)\big)\big]_x \\
&- \big(\gamma_{1,\varepsilon}\big|A(u^{\mu,\varepsilon})-A(c)\big|_x\big)_x - \mu|u^{\mu,\varepsilon}-c|_{xx} - \big|f\big(\boldsymbol{\gamma}_\varepsilon(x),c\big)_x\big| \le 0 \quad \text{in } \mathcal{D}'(\Pi_T);
\end{aligned} \tag{4.9}
$$

that is, for any $0 \le \phi \in \mathcal{D}'(\Pi_T)$,

$$
\begin{aligned}
&\iint_{\Pi_T} \Big(|u^{\mu,\varepsilon}-c|\phi_t + \operatorname{sgn}(u^{\mu,\varepsilon}-c)\big(f(\boldsymbol{\gamma}_\varepsilon(x),u^{\mu,\varepsilon})-f(\boldsymbol{\gamma}_\varepsilon(x),c)\big)\phi_x \\
&\qquad - \gamma_{1,\varepsilon}\big|A(u^{\mu,\varepsilon})-A(c)\big|_x\phi_x + \mu|u^{\mu,\varepsilon}-c|\phi_{xx}\Big)\,dx\,dt \\
&\quad + \iint_{\Pi_T} \big|f(\boldsymbol{\gamma}_\varepsilon(x),c)_x\big|\phi\,dx\,dt \ge 0.
\end{aligned} \tag{4.10}
$$

Using the properties of $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_\varepsilon$, we can write (notice the signs)

$$
\begin{aligned}
\iint_{\Pi_T} \big|f(\boldsymbol{\gamma}_\varepsilon(x),c)_x\big|\phi\,dx\,dt &= \int_0^T \int_{x_{\mathrm{L}}-\varepsilon}^{x_{\mathrm{L}}+\varepsilon} \gamma_{1,\varepsilon}'(x)|b(c)|\phi(x,t)\,dx\,dt \\
&\quad - \int_0^T \int_{x_{\mathrm{R}}-\varepsilon}^{x_{\mathrm{R}}+\varepsilon} \gamma_{1,\varepsilon}'(x)|b(c)|\phi(x,t)\,dx\,dt + \int_0^T \int_{-\varepsilon}^{\varepsilon} \gamma_{2,\varepsilon}'(x)|c-u_{\mathrm{F}}|\phi(x,t)\,dx\,dt.
\end{aligned}
$$

After three integrations by parts and subsequently sending ε to zero, we obtain

$$
\begin{aligned}
\lim_{\varepsilon\to 0}\iint_{\Pi_T} \big|f(\boldsymbol{\gamma}_\varepsilon(x),c)_x\big|\phi\,dx\,dt &= \int_0^T \big(\gamma_1(x_{\mathrm{L}}+)-\gamma_1(x_{\mathrm{L}}-)\big)\,|b(c)|\phi(x_{\mathrm{L}},t)\,dt \\
&\quad - \int_0^T \big(\gamma_1(x_{\mathrm{R}}+)-\gamma_1(x_{\mathrm{R}}-)\big)\,|b(c)|\phi(x_{\mathrm{R}},t)\,dt \\
&\quad + \int_0^T \big(\gamma_2(0+)-\gamma_2(0-)\big)\,|c-u_{\mathrm{F}}|\phi(0,t)\,dt. \\
&= \int_0^T \sum_{m\in\mathcal{J}} \big|f\big(\boldsymbol{\gamma}(m+),c\big)-f\big(\boldsymbol{\gamma}(m-),c\big)\big|\phi(m,t)\,dt.
\end{aligned}
$$

In view of the previous calculations, we can send $\mu,\varepsilon \to 0$ in (4.10) to conclude that the limit function u satisfies the entropy condition (4.5).

4.3.2. Status of the weak formulation (4.3). We point out that the entropy inequality (4.5) does not imply the weak formulation (4.3). In fact, the usual procedure to derive the weak formulation from the Kružkov entropy inequality consists

in taking first $c > \|u\|_\infty$ and then $c < -\|u\|_\infty$, and then combining the resulting inequalities. This does not work here, since the term $|f(\boldsymbol{\gamma}(m+), c) - f(\boldsymbol{\gamma}(m-), c)|$ does not have compact support with respect to c, and therefore the sum over $m \in \mathcal{J}$ in (4.5) will not disappear for these values of c. Consequently, we state the weak formulation (4.3) and the entropy inequality (4.5) as independent ingredients of the solution concept.

4.3.3. Alternative entropy weak solution concepts. Section 4.3.1 shows that the (Kružkov) entropy formulation (4.5) for the clarifier-thickener model can be derived from the parabolic regularization (4.6) when the regularization parameter μ and the parameter ε that smoothes the flux discontinuities tend to zero. In section 1 we mentioned that the term μu_{xx} corresponds to hydrodynamic diffusion. Consequently, the entropy formulation (4.5), which eventually leads to uniqueness, encodes that our "physically relevant" solutions are those obtained in the limit of vanishing hydrodynamic diffusion.

Alternative entropy concepts for conservation laws with discontinuous flux, which equally lead to unique solutions (that are, however, different from the ones constructed here), are possible. For example, a mathematical model for two-phase flow in a one-dimensional porous medium that changes its type at $x = 0$ can be written as the conservation law with discontinuous flux

$$u_t + \big(H(x)f(u) + (1 - H(x))g(u)\big)_x = 0, \tag{4.11}$$

where u is the saturation of one of the phases, the functions $f(u)$ and $g(u)$ are the rock-type dependent Darcy velocities for $x > 0$ and $x < 0$, respectively, and $H(x)$ is the Heaviside function. These models are usually based on the assumption that the capillary pressure is continuous across heterogeneities of the porous medium. Consequently, the appropriate viscous regularization term of (4.11) for this model is not given by μu^ε_{xx} but by $\mu(\lambda_c(u^\varepsilon)p_c(u^\varepsilon)_x)_x$, where λ_c and p_c are the mobility and capillary pressure functions for the phase considered and $x < 0$ ($c = \mathrm{L}$) and $x > 0$ ($c = \mathrm{R}$), respectively [51, 71]. Analyzing the limit $\varepsilon \to 0$ for this regularization term, Kaasschieter [51] shows that the corresponding viscosity limit produces an entropy condition for the limiting equation (4.11) that excludes that characteristics leave the discontinuity at $x = 0$ from both sides. In other words, the capillary pressure characterization does not allow undercompressive shocks emerging from $x = 0$. This contrasts with the role of the flux discontinuity at $x = 0$ in the clarifier-thickener model, in which material is injected at $x = 0$ and transported from both sides of that source term location, which allows characteristics to emerge to both sides of $x = 0$.

Based on the properties of this two-phase flow in heterogeneous porous media model, Adimurthi, Jaffré, and Veerappa Gowda propose in [1] an entropy formulation alternative to ours, which is essentially based on the usual Kružkov entropy characterization away from flux discontinuities, supplemented by an additional entropy jump condition that rules out undercompressive shocks. In [1] it is shown that regular solutions satisfying these conditions are unique, and convergence of a Godunov-type scheme to weak solutions of this type is proved. The entropy concepts of [1] and [25] (to which the present treatment reduces for $A \equiv 0$) coincide if the "left" and "right" fluxes $f(u)$ and $g(u)$ do *not* intersect and in general yield different results if these fluxes intersect in an "undercompressive" way. This means that if we suppose, for simplicity, that the functions $f(u)$ and $g(u)$ intersect at only one point, called u_χ here, then in the vicinity of u_χ we have $g'(u) < 0$ for $u \le u_\chi$ and $f'(u) > 0$ for $u \ge u_\chi$. Such a situation frequently occurs in the two-phase flow in porous media model and

is also possible in the clarifier-thickener model. For example, consider the situation near the feed level $x = 0$, and assume that we take the parameters q_{L} and q_{R} as used in Figure 2.1 but set $u_{\mathrm{F}} = 0.8$. Graphically, this situation corresponds to moving the two curves in Figure 2.1(d) further apart until they intersect at $u_\chi = u_{\mathrm{F}} = 0.8$. From the shape of these curves it becomes clear that this is indeed a case of an "undercompressive" intersection, where $f(u) = q_{\mathrm{R}}(u - u_{\mathrm{F}}) + b(u)$ and $g(u) = q_{\mathrm{L}}(u - u_{\mathrm{F}}) + b(u)$. Furthermore, if we assume that the clarifier-thickener is initially filled with suspension of that same concentration $u_0 = u_{\mathrm{F}}$, we expect that filling up the vessel with (strongly concentrated) suspension of concentration $u = u_{\mathrm{F}}$ produces a solution that is constant near $x = 0$ (for small times at least, until waves coming from x_{L} or x_{R} start to interfere). In this scenario, however, the theory of [1], which excludes undercompressive shocks, leads to a different solution, which includes one stationary oscillation between solution values θ_f and θ_g such that $f'(\theta_f) \le 0$ or $g'(\theta_g) \ge 0$; i.e., the resulting stationary discontinuity is not undercompressive. In [20] we also present a numerical example for a similar case showing that both entropy theories, along with the convergence of the respective associated schemes, lead to different admissible solutions. Of course, these alternative theories mirror the different physics involved, as the different viscous regularizations of the clarifier-thickener model and of the two-phase flow in porous media model show. Finally, let us clarify that the problem of whether the fluxes $f(u)$ and $g(u)$ intersect in an "undercompressive way" is *not* equivalent to the satisfaction of the so-called crossing condition, to which we appeal in section 4.5 for the uniqueness proof. See the comment following Lemma 4.14 for a discussion of this point in the context of the clarifier-thickener model.

4.3.4. Motivation of condition (D.4). One consequence of this definition is that for a.e. $t \in [0, T]$, $A(u(x,t))$ is absolutely continuous as a function of x on $[x_{\mathrm{L}}, x_{\mathrm{R}}]$ and that $A(u)_x$ exists a.e. in $[x_{\mathrm{L}}, x_{\mathrm{R}}]$. The point that deserves to be discussed extensively is, however, our explicit requirement that $A(u)$ be continuous across $x = x_{\mathrm{L}}$ and $x = x_{\mathrm{R}}$. To justify this assumption, we follow the analysis of van Duijn, de Neef, and Molenaar [80] and Molenaar [71]. We consider the situation near x_{R}. (The jump across x_{L} is handled in the same way.) We approximate the discontinuous coefficient $\gamma_1(x)$ by a smooth function $\gamma_1^\varepsilon(x)$, where $\varepsilon > 0$ is a small smoothing parameter and $\gamma_1^\varepsilon(x) = \gamma_1(x)$ outside $(x_{\mathrm{R}} - \varepsilon, x_{\mathrm{R}} + \varepsilon)$. The differential equation that should be satisfied by the limit function in this interval for $\varepsilon \to 0$ provides an additional interface condition. We assume that v^ε is the solution of Model 1 with $\gamma_1(x)$ replaced by γ_1^ε and introduce the rescaled space variable $y := (x - x_{\mathrm{R}})/\varepsilon$. Inside $(x_{\mathrm{R}} - \varepsilon, x_{\mathrm{R}} + \varepsilon)$, our equation can be written as $\varepsilon v_t^\varepsilon + (f^\varepsilon(y, v^\varepsilon))_y = 0$ with

$$f^\varepsilon(y, v^\varepsilon) = q_{\mathrm{R}}(v^\varepsilon - u_{\mathrm{F}}) + \gamma_1^\varepsilon b(v^\varepsilon) - \gamma_1^\varepsilon \frac{1}{\varepsilon} A(v^\varepsilon)_y.$$

We assume that v^ε with $0 \le v^\varepsilon \le 1$ converges to a function v as $\varepsilon \to 0$ and that the total flux f^ε remains bounded uniformly in ε as $\varepsilon \to 0$. Then the limit function v should satisfy $A(v)_y = 0$ for $-1 < y < 1$. Integrating this equation yields the following relation between $v(-1,t)$ and $v(1,t)$, which are identified with the limits of u to the left and to right of $x = x_{\mathrm{R}}$, $u(x_{\mathrm{R}}-,t)$, and $u(x_{\mathrm{R}}+,t)$:

$$A\big(u(x_{\mathrm{R}}+,t)\big) - A\big(u(x_{\mathrm{R}}-,t)\big) = A\big(v(1,t)\big) - A\big(v(-1,t)\big) = \int_{-1}^{1} A(v)_y \, dy = 0,$$

which motivates our condition (D.4).

To highlight the status of condition (D.4), we first mention that (D.4) is the analogue of the "extended pressure condition" postulated in problems of multiphase flow in heterogeneous porous media [71, 80, 85]. These problems lead to equations with discontinuous flux and discontinuous (with respect to the space variable) diffusion, which require an additional jump condition across jumps of the diffusion coefficient (apart from the appropriate Rankine–Hugoniot condition) to ensure uniqueness. This analogy and the observation that in our case the diffusion terms are not present at all for $x < x_{\mathrm{L}}$ and $x > x_{\mathrm{R}}$ and it is therefore unlikely to obtain control on the limits of $A(u)_x$ for $x \downarrow x_{\mathrm{L}}$ and $x \uparrow x_{\mathrm{R}}$ strongly support the necessity to postulate (D.4) as a separate ingredient of the definition of an entropy weak solution. This view is also supported by the fact that our uniqueness proof relies on (D.4). In other words, (D.4) is necessary to ensure uniqueness.

It should be mentioned, however, that it is currently unclear how to prove that the numerical scheme converges to a solution that satisfies (D.4). In fact, we will prove that the scheme converges to a limit u that satisfies all components of Definition 4.1 except (D.4). However, our numerical results support that $A(u)$ is continuous across $x = x_{\mathrm{L}}$ and $x = x_{\mathrm{R}}$. In particular, transient numerical simulations converge (for large times) to steady-state solutions. These are x-dependent piecewise continuous functions that are defined by a time-independent version of Definition 4.1 which does *not* include a postulate of continuity of $A(u)$, since satisfaction of this condition can be derived in the ODE context.

4.4. Existence of traces. In what follows, we often use the notation

$$F(\boldsymbol{\gamma}, u, c) := \operatorname{sgn}(u - c)\big(f(\boldsymbol{\gamma}, u) - f(\boldsymbol{\gamma}, c)\big) \tag{4.12}$$

for the Kružkov entropy flux appearing in (4.5). Here, the sign function is defined by $\operatorname{sgn}(w) := w/|w|$ for $w \neq 0$ and $\operatorname{sgn}(0) := 0$. It is sometimes convenient to work with the function $\hat{f}(\boldsymbol{\gamma}, u) := f(\boldsymbol{\gamma}, u) + \gamma_2 u_{\mathrm{F}}$ and to use the identity $F(\boldsymbol{\gamma}, u, c) = \operatorname{sgn}(u - c)(\hat{f}(\boldsymbol{\gamma}, u) - \hat{f}(\boldsymbol{\gamma}, c))$. We will also find the following notation useful: $a \vee b := \max\{a, b\}$, $a \wedge b := \min\{a, b\}$, $a_+ := a \vee 0$, and $a_- := a \wedge 0$.

Our analysis makes use of certain jump conditions that relate limits from the right and left of the entropy weak solution u at jumps in the spatially varying coefficient $\boldsymbol{\gamma}(x)$. Thus, we need a notion of one-sided limits (from both the right and left) at the points $x \in \mathcal{J}$.

Definition 4.2 (traces). *Let $W \in L^\infty(\Pi_T)$ be a real function. By the right and left traces of $W(\cdot, t)$ at a point $x = x_0 \in \mathbb{R}$ we understand functions $t \mapsto W(x_0\pm, t) \in L^\infty(0, T)$ that satisfy*

$$\operatorname*{ess\,lim}_{x \downarrow x_0} \big|W(x, t) - W(x_0+, t)\big| = 0 \quad \textit{for a.e. } t \in (0, T),$$
$$\operatorname*{ess\,lim}_{x \uparrow x_0} \big|W(x, t) - W(x_0-, t)\big| = 0 \quad \textit{for a.e. } t \in (0, T).$$

As in [25, 56], we now derive the existence of the required traces from our definition of BV_t entropy weak solution. To this end, we introduce the so-called singular mapping. Let $\mathcal{S}(w)$ denote the characteristic function of the interval $[0, u_{\mathrm{c}}]$, where $A'(u) = 0$, and recall that

$$f\big(\boldsymbol{\gamma}(w), u\big) = \gamma_1(x) b(u) + \gamma_2(x)(u - u_{\mathrm{F}}).$$

The singular mapping is defined by

(4.13)
$$\Psi(\boldsymbol{\gamma},u) := \mathcal{F}(\boldsymbol{\gamma},u) + \gamma_1 A(u), \quad \mathcal{F}(\boldsymbol{\gamma},u) := \int_0^u \big(\gamma_1 \mathcal{S}(w) + 1 - \gamma_1\big)\big|f_u(\boldsymbol{\gamma},w)\big|\,dw.$$

Following [55], we have broken the singular mapping into two parts: $\mathcal{F}$ for the hyperbolic spatial operator and $\gamma_1 A$ for the parabolic operator. Note that if $\gamma_1 = 0$, which means that $x \notin [x_{\mathrm{L}}, x_{\mathrm{R}}]$, the parabolic term will not be present, and the singular mapping simplifies to $\Psi(\boldsymbol{\gamma},u) = \mathcal{F}(\boldsymbol{\gamma},u) = \int_0^u |f_u(\boldsymbol{\gamma},w)|\,dw$. If $x \in (x_{\mathrm{L}}, x_{\mathrm{R}})$, then $\gamma_1 = 1$, and

$$\Psi(\boldsymbol{\gamma},u) = \mathcal{F}(\boldsymbol{\gamma},u) + A(u), \quad \mathcal{F}(\boldsymbol{\gamma},u) = \int_0^u \mathcal{S}(w)|f_u(\boldsymbol{\gamma},w)|\,dw.$$

Thus, for $x \in (x_{\mathrm{L}}, x_{\mathrm{R}})$,

$$\partial_u \Psi(\boldsymbol{\gamma},u) = \begin{cases} |f_u(\boldsymbol{\gamma},u)| & \text{for } u \le u_{\mathrm{c}}, \\ A'(u) & \text{for } u > u_{\mathrm{c}}. \end{cases}$$

Thus, we see that for any fixed value of x (and hence $\boldsymbol{\gamma}$) and u, exactly one of the mappings $u \mapsto \mathcal{F}(\boldsymbol{\gamma},u)$, $u \mapsto \gamma_1 A(u)$ is increasing, and the other is constant. This allows us to separate the hyperbolic and parabolic terms somewhat in our analysis and is the motivation behind the particular form of the singular mapping given by (4.13).

The following lemma records some easily verified properties of Ψ. We omit the elementary proofs.

LEMMA 4.3. *The mapping $u \mapsto \Psi(\boldsymbol{\gamma},u)$ is Lipschitz continuous on $[0,1]$, uniformly for $\gamma \in \mathcal{G}$. In addition, $u \mapsto \Psi(\boldsymbol{\gamma},u)$ is strictly increasing on $[0,1]$ for each fixed vector $\boldsymbol{\gamma} \in \mathcal{G}$.*

The proof of the following lemma is easily adapted to the present situation from that of Lemma 3.1 of [56]. The key ingredients are (4.3), (4.5), and the fact that $u_t \in \mathcal{M}(\Pi_T)$.

LEMMA 4.4. *Suppose u is a BV_t entropy weak solution. Then, for any $c \in \mathbb{R}$,*

$$\Big(f\big(\boldsymbol{\gamma}(x),u\big) - f\big(\boldsymbol{\gamma}(x),c\big) - \gamma_1(x)\big(A(u) - A(c)\big)_x\Big)_x \in \mathcal{M}(\Pi_T), \tag{4.14}$$

$$\Big(\operatorname{sgn}(u-c)\big[f\big(\boldsymbol{\gamma}(x),u\big) - f\big(\boldsymbol{\gamma}(x),c\big)\big] - \gamma_1(x)\big|A(u) - A(c)\big|_x\Big)_x \in \mathcal{M}(\Pi_T). \tag{4.15}$$

LEMMA 4.5. *Let u be a BV_t entropy weak solution of (4.1), and consider the transformed function $z(x,t) := \Psi(\boldsymbol{\gamma}(x), u(x,t))$. Then $\int_0^T \mathrm{TV}(z(\cdot,t))\,dt < C$ for some finite constant $C > 0$. In other words, $z_x \in \mathcal{M}(\Pi_T)$.*

Proof. For $A \equiv 0$, the proof of Lemma 2.2 of [25] applies unchanged up to minor differences in notation. Here we modify that proof to account for the presence of $A(u)$. Let $\mathrm{TV}(z(\cdot,t)|_{\mathcal{I}})$ denote the spatial variation of $z(\cdot,t)$ measured over the interval $\mathcal{I}$. Then it suffices to show that $\int_0^T \mathrm{TV}(z(\cdot,t)|_{\mathcal{I}})\,dt$ is bounded for each of the open intervals $\mathcal{I} = (-\infty, x_{\mathrm{L}})$, $(x_{\mathrm{L}}, 0)$, $(0, x_{\mathrm{R}})$, and (x_{R}, ∞). Due to the factor $\gamma_1(x)$, the singular mapping Ψ simplifies to $\Psi(\boldsymbol{\gamma},u) = \int_0^u |f_u(\boldsymbol{\gamma},w)|\,dw$ when $\mathcal{I} = (-\infty, x_{\mathrm{L}})$ or $\mathcal{I} = (x_{\mathrm{R}}, \infty)$, and so the proof of Lemma 2.2 of [25] applies to those intervals

without any modifications. We will focus on the interval $\mathcal{I} = (0, x_{\mathrm{R}})$ and omit the proof for $(x_{\mathrm{L}}, 0)$ since it is similar. Thus, we now set out to show that

$$\int_0^T \mathrm{TV}\big(z(\cdot,t)|_{\{x|\,0<x<x_{\mathrm{R}}\}}\big)\,dt < \infty. \tag{4.16}$$

To this end, recall that for $x \in (0, x_{\mathrm{R}})$, we have $\boldsymbol{\gamma} = (1, q_{\mathrm{R}})$, and

$$\begin{aligned}
f(\boldsymbol{\gamma}, u) &= q_{\mathrm{R}} u + b(u) - q_{\mathrm{R}} u_{\mathrm{F}} = \hat{f}(\boldsymbol{\gamma}, u) - q_{\mathrm{R}} u_{\mathrm{F}},\\
\Psi(\boldsymbol{\gamma}, u) &= \int_0^u \mathcal{S}(w)\big|\hat{f}_u(\boldsymbol{\gamma}, w)\big|\,dw + A(u) = \int_0^u \mathcal{S}(w)\big|q_{\mathrm{R}} + b'(w)\big|\,dw + A(u),\\
F(\boldsymbol{\gamma}, u, c) &= \operatorname{sgn}(u-c)\big(f(\boldsymbol{\gamma}, u) - f(\boldsymbol{\gamma}, c)\big) = \operatorname{sgn}(u-c)\big(\hat{f}(\boldsymbol{\gamma}, u) - \hat{f}(\boldsymbol{\gamma}, c)\big).
\end{aligned}$$

Let $\Pi_T^{\mathrm{R}} := (0, x_{\mathrm{R}}) \times (0, T) \subset \Pi_T$. Since $A(u)_x \in L^\infty((x_{\mathrm{L}}, x_{\mathrm{R}}) \times (0,T))$, we have

$$A(u)_x \in \mathcal{M}(\Pi_T^{\mathrm{R}}), \quad \big(A(u) - A(c)\big)_x \in \mathcal{M}(\Pi_T^{\mathrm{R}}), \quad \big|A(u) - A(c)\big|_x \in \mathcal{M}(\Pi_T^{\mathrm{R}}).$$

Thus, it suffices to show that $\mathcal{F}(\boldsymbol{\gamma}, u)_x \in \mathcal{M}(\Pi_T^{\mathrm{R}})$. Note that

$$\mathcal{F}\big(\boldsymbol{\gamma}(x), u(x,t)\big) = \int_0^{u(x,t)} \mathcal{S}(w)\big|\hat{f}_u(\boldsymbol{\gamma}, w)\big|\,dw \quad \text{for } (x,t) \in \Pi_T^{\mathrm{R}}.$$

Due to the assumptions on q_{R} and $b(u)$, the function $\hat{f}$ has at most two extrema for $u \in (0,1)$. We assume that q_{R} is chosen such that there are exactly two extrema $u_1^* < u_2^*$. The cases with one or no extremum will be omitted; they are handled in a similar manner. It is clear that $u \mapsto \hat{f}(\boldsymbol{\gamma}, u)$ is strictly monotone on intervals not containing extrema. We need the following fact, which follows by subtracting (4.14) from (4.15) and then dividing by 2:

$$\Big(\chi_l(w;c)\big(f(\boldsymbol{\gamma}_{\mathrm{R}}, u) - f(\boldsymbol{\gamma}_{\mathrm{R}}, c)\big) - \gamma_1(x)\big((A(u) - A(c))_-\big)_x\Big)_x \in \mathcal{M}(\Pi_T^{\mathrm{R}}) \quad \forall c \in \mathbb{R}. \tag{4.17}$$

Here $\chi_l(w;c)$ is the characteristic function for $[0,c]$, and we have restricted our attention to the smaller domain Π_T^{R}. Finally, we have used the fact that $\boldsymbol{\gamma}(x) \equiv \boldsymbol{\gamma}_{\mathrm{R}}$ for $x \in (0, x_{\mathrm{R}})$.

Next, note that for $c \le u_{\mathrm{c}}$ in (4.17), the term $(A(u) - A(c))_-$ vanishes. This is easy to see by considering the two possible cases $u > c$ and $u \le c$. In the first case, $A(u) - A(c) \ge 0$, since $A(\cdot)$ is nondecreasing, and in the second case $A(u) - A(c) = 0$, since $A(\cdot)$ is constant on $[0, u_{\mathrm{c}}]$. Also, $\hat{f}(\boldsymbol{\gamma}, u) - \hat{f}(\boldsymbol{\gamma}, c) = f(\boldsymbol{\gamma}, u) - f(\boldsymbol{\gamma}, c)$. Thus, we conclude from (4.17) that

$$\Big(\chi_l(w;c)\big(\hat{f}(\boldsymbol{\gamma}_{\mathrm{R}}, u) - \hat{f}(\boldsymbol{\gamma}_{\mathrm{R}}, c)\big)\Big)_x \in \mathcal{M}(\Pi_T^{\mathrm{R}}) \quad \text{for } c \le u_{\mathrm{c}}. \tag{4.18}$$

In (4.18), we now take $c_1 := u_1^* \wedge u_{\mathrm{c}}$, $c_2 := u_2^* \wedge u_{\mathrm{c}}$, and $c_3 := 1 \wedge u_{\mathrm{c}}$, and letting

$$P_i(\boldsymbol{\gamma}_{\mathrm{R}}, u) := \chi_l(w; c_i)\big(\hat{f}(\boldsymbol{\gamma}_{\mathrm{R}}, u) - \hat{f}(\boldsymbol{\gamma}_{\mathrm{R}}, c_i)\big), \quad i = 1,2,3,$$

we have $\partial_x P_i \in \mathcal{M}(\Pi_T^{\mathrm{R}})$. It is a straightforward exercise to verify that

$$\begin{aligned}
\mathcal{F}(\boldsymbol{\gamma}, u) = {} & P_3(\boldsymbol{\gamma}_{\mathrm{R}}, u) - 2P_2(\boldsymbol{\gamma}_{\mathrm{R}}, u) + 2P_1(\boldsymbol{\gamma}_{\mathrm{R}}, u)\\
& + \hat{f}(\boldsymbol{\gamma}_{\mathrm{R}}, c_3) - 2\hat{f}(\boldsymbol{\gamma}_{\mathrm{R}}, c_2) + 2\hat{f}(\boldsymbol{\gamma}_{\mathrm{R}}, c_1),
\end{aligned} \tag{4.19}$$

from which it follows immediately that $\mathcal{F}(\boldsymbol{\gamma}, u)_x \in \mathcal{M}(\Pi_T^{\mathrm{R}})$. ∎

The proof of the following lemma is a direct application of Lemma 4.5. Its proof follows from the proofs of Lemmas 3.3 and 3.4 of [56].

LEMMA 4.6. *A BV_t entropy solution u and the quantities $\gamma_1 A(u)$, $\gamma_1 A(u)_x$, and $\gamma_1|A(u)-A(c)|_x$ admit right and left traces at each jump in $\boldsymbol{\gamma}$.*

4.5. Entropy jump conditions and uniqueness of entropy solutions. Our objective in this section is to prove the L^1 stability of entropy weak solutions, which is stated in Theorem 4.15. If we had in our problem (4.1) the parabolic term $A(u)_{xx}$ instead of $(\gamma_1(x)A(u)_x)_x$, i.e., if $\gamma_1 \equiv 1$, section 2 of [56] would apply verbatim, and we could simply appeal to Theorem 2.1 of that paper. Thus we will follow closely section 2 of [56]. Since the spatially varying parameter γ_1 plays a key role here, we remind the reader that γ_1 is simply the characteristic function for the interval $(x_{\mathrm{L}}, x_{\mathrm{R}})$.

As in [56], it is convenient, and sufficient, to work with limits in the sense of Lebesgue. Specifically, let $W = W(x)$ be any function on $\mathbb{R}$, and fix a point $x_0 \in \mathbb{R}$. We define Lebesgue-type one-sided limits as follows:

$$\operatorname*{L\,lim}_{x\downarrow x_0} W(x) := \lim_{\epsilon\downarrow 0} \frac{1}{\epsilon}\int_{x_0}^{x_0+\epsilon} W(x)\,dx, \quad \operatorname*{L\,lim}_{x\uparrow x_0} W(x) := \lim_{\epsilon\downarrow 0} \frac{1}{\epsilon}\int_{x_0-\epsilon}^{x_0} W(x)\,dx.$$

The key fact here (see Lemma 2.1 of [56]) is the following.

LEMMA 4.7. *Let $W \in L^\infty(\Pi_T)$, and fix a point $x_0 \in \mathbb{R}$. If the right and left traces $t \mapsto W(x_0\pm, t)$ exist in the sense of Definition 4.2, then for a.e. $t \in (0,T)$ they also exist as right and left traces in the sense of Lebesgue points in L^1:*

$$\operatorname*{L\,lim}_{x\downarrow x_0} W(x,t) = W(x_0+,t), \qquad \operatorname*{L\,lim}_{x\uparrow x_0} W(x,t) = W(x_0-,t).$$

Next, we record the versions of Lemmas 2.2 and 2.3 of [56] that account for the coefficient $\gamma_1(x)$ multiplying $A(u)_x$. The proofs in [56] are easily modified to deal with γ_1 and are omitted here.

LEMMA 4.8. *Let u and v be a pair of BV_t entropy weak solutions. Let F be the Kružkov entropy flux defined in* (4.12). *Fix one of the jumps in $\boldsymbol{\gamma}$ located at $m \in \mathcal{J}$. Then for a.e. $t \in (0,T)$*

$$\begin{aligned} &\operatorname*{L\,lim}_{x\downarrow m} F\big(\boldsymbol{\gamma}(x), u(x,t), v(x,t)\big) = F\big(\boldsymbol{\gamma}(m+), u(m+,t), v(m+,t)\big),\\ &\operatorname*{L\,lim}_{x\uparrow m} F\big(\boldsymbol{\gamma}(x), u(x,t), v(x,t)\big) = F\big(\boldsymbol{\gamma}(m-), u(m-,t), v(m-,t)\big), \end{aligned} \tag{4.20}$$

$$\operatorname*{L\,lim}_{x\downarrow m}\big(\gamma_1(x)\big|A(u)-A(v)\big|_x\big)(x,t)$$

$$= \begin{cases} \gamma_1(m+)\sigma(m+,t)\big(A(u)_x(m+,t) - A(v)_x(m+,t)\big) & \text{if } A(u(m+,t)) \\ & \neq A(v(m+,t)), \\ \gamma_1(m+)\big|A(u)_x(m+,t) - A(v)_x(m+,t)\big| & \text{otherwise}, \end{cases} \tag{4.21}$$

$$\operatorname*{L\,lim}_{x\uparrow m}\big(\gamma_1(x)\big|A(u)-A(v)\big|_x\big)(x,t)$$

$$= \begin{cases} \gamma_1(m-)\sigma(m-,t)\big(A(u)_x(m-,t) - A(v)_x(m-,t)\big) & \text{if } A(u(m-,t)) \\ & \neq A(v(m-,t)), \\ -\gamma_1(m-)\big|A(u)_x(m-,t) - A(v)_x(m-,t)\big| & \text{otherwise}, \end{cases} \tag{4.22}$$

where $\sigma(m+,t) := \operatorname{sgn}(u(m+,t)-v(m+,t))$ and $\sigma(m-,t) := \operatorname{sgn}(u(m-,t)-v(m-,t))$.

LEMMA 4.9. *Let u be a BV_t entropy weak solution. Let F be the Kružkov entropy flux defined in* (4.12)*. Fix one of the jumps in $\boldsymbol{\gamma}$ located at $m \in \mathcal{J}$. For any $c \in \mathbb{R}$, we have for a.e. $t \in (0,T)$*

$$\begin{aligned} \operatorname*{L\,lim}_{x \downarrow m} F\big(\boldsymbol{\gamma}(x), u(x,t), c\big) &= F\big(\boldsymbol{\gamma}(m+), u(m+,t), c\big), \\ \operatorname*{L\,lim}_{x \uparrow m} F\big(\boldsymbol{\gamma}(x), u(x,t), c\big) &= F\big(\boldsymbol{\gamma}(m-), u(m-,t), c\big), \end{aligned} \tag{4.23}$$

$$\begin{aligned} &\operatorname*{L\,lim}_{x \downarrow m} \big(\gamma_1(x) \big| A(u) - A(c) \big|_x\big)(x,t) \\ &= \begin{cases} \gamma_1(m+)\sigma(m+,t) A(u)_x(m+,t) & \text{if } u(m+,t) \neq c, \\ \gamma_1(m+) \big| A(u)_x(m+,t) \big| & \text{if } u(m+,t) = c, \end{cases} \end{aligned} \tag{4.24}$$

$$\begin{aligned} &\operatorname*{L\,lim}_{x \uparrow m} \big(\gamma_1(x) \big| A(u) - A(c) \big|_x\big)(x,t) \\ &= \begin{cases} \gamma_1(m-)\sigma(m-,t) A(u)_x(m-,t) & \text{if } u(m-,t) \neq c, \\ -\gamma_1(m-) \big| A(u)_x(m-,t) \big| & \text{if } u(m-,t) = c, \end{cases} \end{aligned} \tag{4.25}$$

where $\sigma(m-,t) := \operatorname{sgn}(u(m-,t) - c)$ and $\sigma(m+,t) := \operatorname{sgn}(u(m+,t) - c)$.

Before continuing, we introduce a notational convention that we hope will simplify the appearance of the formulas that follow. Whenever we are discussing a fixed element $m \in \mathcal{J}$, and the time is fixed at $t \in [0,T]$ where all of the relevant right and left limits exist, we use the notation $u_\pm = u_\pm(t) = u(m\pm,t)$, $\boldsymbol{\gamma}_\pm = \boldsymbol{\gamma}(m\pm)$, and $(\gamma_1 A(u_x))_\pm = (\gamma_1 A(u)_x)(m\pm,t)$.

We collect in the following lemma several properties of a BV_t entropy weak solution near a jump in $\boldsymbol{\gamma}$. The relationship (4.26) is the Rankine–Hugoniot condition for this problem, while (4.28) is an entropy condition. The relationship (4.27) restricts the sign of $A(u)_x$ at a jump in $\boldsymbol{\gamma}$.

REMARK 4.1. *To highlight once again the significance of assumption* (D.4)*, we mention that the proof of* (4.27) *requires the hypothesis that $A(u)$ is continuous across the jumps in $\boldsymbol{\gamma}$ at x_{L} and x_{R}. Thus,* (D.4) *is crucial for the uniqueness of entropy weak solutions of Model* 1*.*

LEMMA 4.10. *Let u be a BV_t entropy weak solution. Fix one of the jumps in $m \in \mathcal{J}$. Then the following relationships hold across the jump for a.e. $t \in (0,T)$:*

$$f\big(\boldsymbol{\gamma}_+, u_+\big) - (\gamma_1 A(u)_x)_+ = f\big(\boldsymbol{\gamma}_-, u_-\big) - (\gamma_1 A(u)_x)_-, \tag{4.26}$$

$$\operatorname{sgn}(u_+ - u_-) \operatorname{sgn}\big((A(u)_x)_+\big) \geq 0, \quad \operatorname{sgn}(u_+ - u_-) \operatorname{sgn}\big((A(u)_x)_-\big) \geq 0. \tag{4.27}$$

And for $u_-(t) \neq u_+(t)$,

$$\begin{aligned} &\big[F\big(\boldsymbol{\gamma}_+, u_+, c\big) - \operatorname{sgn}(u_+ - c)\,(\gamma_1 A(u)_x)_+\big] \\ &- \big[F\big(\boldsymbol{\gamma}_-, u_-, c\big) - \operatorname{sgn}(u_- - c)\,(\gamma_1 A(u)_x)_-\big] \\ &\qquad \leq \big|f(\boldsymbol{\gamma}_+, c) - f(\boldsymbol{\gamma}_-, c)\big| \qquad \forall c \in \mathbb{R}, \end{aligned} \tag{4.28}$$

where F is the Kružkov entropy flux function defined in (4.12)*. In addition, the appropriate inequality in Table* 4.1 *holds for all c between u_- and u_+.*

Proof. The proofs of these assertions are similar to the proofs of Lemmas 2.4, 2.5, 2.6, and 2.7 of [56]. Since the proof of (4.27) relies on the assumption (D.4), we will review its proof but will not repeat the proofs of the other assertions. We start by fixing a time $t \in (0,T)$ where all of the relevant right and left spatial (essential)

TABLE 4.1
Entropy jump conditions.

	$f(\boldsymbol{\gamma}_-, c) \le f(\boldsymbol{\gamma}_+, c)$	$f(\boldsymbol{\gamma}_-, c) \ge f(\boldsymbol{\gamma}_+, c)$
$u_- \le c \le u_+$	$f(\boldsymbol{\gamma}_+, u_+) - (\gamma_1 A(u)_x)_+$ $\le f(\boldsymbol{\gamma}_+, c)$	$f(\boldsymbol{\gamma}_-, u_-) - (\gamma_1 A(u)_x)_-$ $\le f(\boldsymbol{\gamma}_-, c)$
$u_+ \le c \le u_-$	$f(\boldsymbol{\gamma}_-, u_-) - (\gamma_1 A(u)_x)_-$ $\ge f(\boldsymbol{\gamma}_-, c)$	$f(\boldsymbol{\gamma}_+, u_+) - (\gamma_1 A(u)_x)_+$ $\ge f(\boldsymbol{\gamma}_+, c)$

limits exist at $x = m$. We prove only the first inequality in (4.27); the proof of the other inequality is similar. Let us suppress the dependence on t for the remainder of the proof. If $u_+ = u_-$, the inequality is obvious, so assume that $u_+ > u_-$. We must show that $(A(u)_x)_+ \ge 0$. From

$$\operatorname*{ess\,lim}_{\varepsilon \downarrow 0} u(m+\varepsilon) =: u_+ > u_-,$$

it is clear that $u(m+\varepsilon) > u_-$ for a.e. sufficiently small $\varepsilon > 0$. Next, we apply A to both sides of $u(m+\varepsilon) > u_-$. Since A is nondecreasing, we obtain $A(u(m+\varepsilon)) \ge A(u_-)$ for a.e. sufficiently small $\varepsilon > 0$. If the jump point is at the origin, i.e., $m = 0$, then continuity of $A(u)$ follows from assumption (D.1). If $m = x_{\mathrm{L}}$ or $m = x_{\mathrm{R}}$, then assumption (D.4) gives us continuity of $A(u)$. In either case, we have $A(u_-) = A(u_+)$. Thus,

$$\frac{1}{\varepsilon} \int_m^{m+\varepsilon} A(u)_x \, dx = \frac{1}{\varepsilon} \left(A(u(m+\varepsilon)) - A(u_+) \right) \ge 0 \tag{4.29}$$

for a.e. sufficiently small $\varepsilon > 0$. Letting $\varepsilon \downarrow 0$ (along a subsequence for which (4.29) holds) yields $(A(u)_x)_+ \ge 0$. The proof is completed by showing via a similar argument that if $u_+ < u_-$, then $(A(u)_x)_+ \le 0$. ∎

REMARK 4.2. *Since A is nondecreasing, we can write the entropy condition* (4.28) *in the alternative form*

$$\Phi(\boldsymbol{\gamma}, u, c)_+ - \Phi(\boldsymbol{\gamma}, u, c)_- \le \left| f(\boldsymbol{\gamma}_+, c) - f(\boldsymbol{\gamma}_-, c) \right| \quad \forall c \in \mathbb{R}, \tag{4.30}$$

where $\Phi(\boldsymbol{\gamma}, u, c) := F(\boldsymbol{\gamma}, u, c) - \gamma_1 |A(u) - A(c)|_x$. Note that the entropy jump condition (4.28) *is the same as the one stated in Lemma* 2.6 *of* [56]*, with the exception that $\gamma_1 = \gamma_1(x)$ is not present in* [56]*. Similarly, this is the only difference between Table* 1 *of* [56] *and our Table* 4.1.

The next lemma is basically Lemma 2.8 of [56], adapted to the present setup.

LEMMA 4.11. *Let u be a BV_t entropy weak solution. Fix the jump in $\boldsymbol{\gamma}$ located at $m = 0$ and a time $t \in [0, T]$ where all of the relevant right and left limits exist. If $u_- \ne u_+$, then $A'(w) = 0$ for w between u_- and u_+, and thus $A(\cdot)$ is constant on the interval connecting u_- to u_+; that is,*

$$A(w) = A(u_-) = A(u_+) \quad \text{for } w \text{ between } u_- \text{ and } u_+. \tag{4.31}$$

Taken together, Lemma 4.11 and assumption (D.4) guarantee continuity of $x \mapsto A(u(x,t))$ for a.e. $t \in [0, T]$ at each of the jumps $m \in \{x_{\mathrm{L}}, 0, x_{\mathrm{R}}\}$. Using this fact, along with the relationships in Table 4.1, it is possible to prove the following lemma, whose proof we omit, since it is essentially the same as the proof of Lemma 2.9 of [56].

LEMMA 4.12. *Let u and v be a pair of BV_t entropy weak solutions. Fix a jump in $\boldsymbol{\gamma}$ located at $m \in \mathcal{J}$ and a time $t \in (0,T)$ where all of the relevant right and left traces exist. Assume that $u_- > v_-$, $u_+ < v_+$. If $u_+ \le u_-$, then*

$$\begin{aligned} &v_- \in [u_+, u_-), \;\; f(\boldsymbol{\gamma}_+, v_-) \ge f(\boldsymbol{\gamma}_-, v_-) \\ &\qquad \Longrightarrow f(\boldsymbol{\gamma}_-, v_-) - \big(\gamma_1 A(v)_x\big)_- \le f(\boldsymbol{\gamma}_-, u_-) - \big(\gamma_1 A(u)_x\big)_-, \\ &v_+ \in (u_+, u_-], \;\; f(\boldsymbol{\gamma}_-, v_+) \ge f(\boldsymbol{\gamma}_+, v_+) \\ &\qquad \Longrightarrow f(\boldsymbol{\gamma}_+, v_+) - \big(\gamma_1 A(v)_x\big)_+ \le f(\boldsymbol{\gamma}_+, u_+) - \big(\gamma_1 A(u)_x\big)_+. \end{aligned} \tag{4.32}$$

If $v_- \le v_+$, then

$$\begin{aligned} &u_- \in (v_-, v_+], \;\; f(\boldsymbol{\gamma}_+, u_-) \le f(\boldsymbol{\gamma}_-, u_-) \\ &\qquad \Longrightarrow f(\boldsymbol{\gamma}_-, v_-) - \big(\gamma_1 A(v)_x\big)_- \le f(\boldsymbol{\gamma}_-, u_-) - \big(\gamma_1 A(u)_x\big)_-, \\ &u_+ \in [v_-, v_+), \;\; f(\boldsymbol{\gamma}_+, u_+) \ge f(\boldsymbol{\gamma}_-, u_+) \\ &\qquad \Longrightarrow f(\boldsymbol{\gamma}_+, v_+) - \big(\gamma_1 A(v)_x\big)_+ \le f(\boldsymbol{\gamma}_+, u_+) - \big(\gamma_1 A(u)_x\big)_+. \end{aligned} \tag{4.33}$$

Before proceeding to our main uniqueness theorem, let us recall the so-called crossing condition that we introduced in [56].

DEFINITION 4.13 (crossing condition). *For any jump in $\boldsymbol{\gamma}$ with associated left and right limits $(\boldsymbol{\gamma}_-, \boldsymbol{\gamma}_+)$, we say that the crossing condition holds if, for any states u and v,*

$$f(\boldsymbol{\gamma}_+, u) - f(\boldsymbol{\gamma}_-, u) < 0 < f(\boldsymbol{\gamma}_+, v) - f(\boldsymbol{\gamma}_-, v) \text{ implies } u < v. \tag{4.34}$$

The geometric interpretation of this condition is that if the graphs of $u \mapsto f(\boldsymbol{\gamma}_-, u)$ and $u \mapsto f(\boldsymbol{\gamma}_+, u)$ cross, then there can be at most one crossing, say at $u = u_\chi$, and in this case the graph of $f(\boldsymbol{\gamma}_-, u)$ lies above (below) the graph of $f(\boldsymbol{\gamma}_+, u)$ for $u < u_\chi$ ($u > u_\chi$). The crossing condition is satisfied automatically if there is no crossing. Figure 2.1(d) shows an example of a flux crossing that satisfies the crossing condition, with crossing point $u_\chi = u_{\mathrm{F}}$. A motivation for the crossing condition in the present context is given by the following lemma, whose elementary proof is provided in [25]. See also Figure 2.1.

LEMMA 4.14. *With our assumptions on b, q_{L}, and q_{R}, the crossing condition is satisfied at each jump $m \in \mathcal{J}$. Specifically, there are no flux crossings associated with the jumps $x = x_{\mathrm{L}}$, $x = x_{\mathrm{R}}$. There may be a single crossing at the jump $x = 0$, but it satisfies the crossing condition.*

Satisfaction of the crossing condition at the nontrivial crossing at $x = 0$ may be traced to the fact that a source is located there, and thus the flow diverges from the origin. This is most easily understood by ignoring for a moment the batch flux $b(u)$ and the parabolic term. It is easy to check that these terms do not affect the crossing relationship at $x = 0$. Then, if we use $\delta(x)$ to denote the delta function, our model simplifies to

$$u_t + q(x) u_x = (q_{\mathrm{R}} - q_{\mathrm{L}}) u_{\mathrm{F}}\, \delta(x), \quad q(x) = \begin{cases} q_{\mathrm{L}} & \text{for } x < 0, \\ q_{\mathrm{R}} & \text{for } x > 0. \end{cases}$$

Since $q_{\mathrm{L}} < 0 < q_{\mathrm{R}}$, we have diverging flows, balanced by a source term on the right-hand side. Notice that in this case the flux curves are the straight lines $u \mapsto q_{\mathrm{L}}(u - u_{\mathrm{F}})$ and $u \mapsto q_{\mathrm{R}}(u - u_{\mathrm{F}})$, and the crossing condition is satisfied. On the other hand, if we

had $q_{\mathrm{L}} > 0 > q_{\mathrm{R}}$, then our simplified model would result in converging flows, balanced by a sink term, and the crossing condition would be violated. Thus, from a physical point of view, our assumption that the crossing condition is satisfied is a natural one and follows directly from the fact that the clarifier-thickener model has a source term (as opposed to a sink term).

If any of the jumps in $\boldsymbol{\gamma}$ violated the crossing condition, our definition of entropy solution would not be strong enough to rule out so-called expansion shocks (see [56] for a detailed explanation), and our uniqueness theory would break down. It turns out that additional entropy conditions are required when the crossing condition is not satisfied; we defer further discussion of this issue to a future paper since the crossing condition is satisfied in the present context.

We are finally able to prove our main uniqueness theorem.

THEOREM 4.15 (L^1 stability and uniqueness). *Let v and u be two BV_t entropy weak solutions to the initial value problem* (4.1). *For a.e. $t \in (0,T)$,*

$$\big\|v(\cdot,t) - u(\cdot,t)\big\|_{L^1(\mathbb{R})} \le \|v_0 - u_0\|_{L^1(\mathbb{R})}. \tag{4.35}$$

Proof. For BV_t entropy weak solutions u and v, a "doubling of variables" argument appearing in Appendix A of [56] yields

$$-\iint_{\Pi_T} \Big(|v-u|\varphi_t + F(\boldsymbol{\gamma}(x),v,u)\varphi_x + \big|\gamma_1 A(v) - \gamma_1 A(u)\big|\varphi_{xx} \Big)\, dt\, dx \le 0 \tag{4.36}$$

for any $0 \le \varphi \in \mathcal{D}(\Pi_T \setminus \mathcal{J})$. Next, via a limiting argument (see the proof of Theorem 2.1 of [56]) we extend this inequality to the larger class of test functions which do not vanish near $x \in \mathcal{J}$. Specifically, we obtain for any $0 \le \phi \in \mathcal{D}(\Pi_T)$

$$\begin{aligned} &-\iint_{\Pi_T} \Big(|v-u|\phi_t + F(\boldsymbol{\gamma}(x),v,u)\phi_x - \big|\gamma_1 A(v) - \gamma_1 A(u)\big|\phi_{xx} \Big)\, dt\, dx \\ &\quad \le \sum_{m \in \mathcal{J}} \int_0^T \Big[F(\gamma(x),v,u) - \big|\gamma_1 A(v) - \gamma_1 A(u)\big|_x \Big]_{x=m-}^{x=m+} \phi(m,t)\, dt, \end{aligned} \tag{4.37}$$

where the notation indicates limits from the right and left at $x = m$.

We wish to show that each term in the sum on the right-hand side of (4.37) is nonpositive. If we fix a jump point $m \in \mathcal{J}$, then the contribution to this sum from the jump point m is given by

$$R := \Phi\big(\boldsymbol{\gamma}(m+), v(m+,t), u(m+,t)\big) - \Phi\big(\boldsymbol{\gamma}(m-), v(m-,t), u(m-,t)\big). \tag{4.38}$$

Here Φ is defined in Remark 4.2 and appears in the entropy condition (4.30). Our goal is now to show that R is nonpositive. Let us fix a time $t \in (0,T)$ where all of the relevant essential right and left limits exist. If $m = 0$, then since $\gamma_1 = 1$ on the interval $(x_{\mathrm{L}}, x_{\mathrm{R}})$ containing $x = 0$, $R \le 0$ is immediate by repeating the relevant portion (the seven cases) of the proof of Theorem 2.1 of [56]. (Note that in [56] we used the symbol S for the quantity known here as R.) We will not reproduce that proof here, but we emphasize that this ($m = 0$) is the only case where a nontrivial flux crossing occurs, and thus we rely on the fact that the crossing condition is satisfied.

If $x = x_{\mathrm{L}}$ or $x = x_{\mathrm{R}}$, then because of the jump in γ_1, we cannot appeal directly to [56], which did not address the case of a spatially varying parabolic term. We will focus on the case $x = x_{\mathrm{R}}$ and omit the similar case $x = x_{\mathrm{L}}$. The approach is to verify that $R \le 0$ in each of the seven cases identified in [56]. The assumptions on $b(u)$

ensure that at $x = x_{\mathrm{R}}$, we will always have $f(\boldsymbol{\gamma}_-, u) \geq f(\boldsymbol{\gamma}_+, u)$. In particular, there are no flux crossings, which simplifies the proofs of Cases 6 and 7.

Case 1 ($v_- = u_-$, $v_+ = u_+$). Then $F(\boldsymbol{\gamma}_+, v_+, u_+) = 0$ and $F(\boldsymbol{\gamma}_-, v_-, u_-) = 0$, and by Lemma 4.8 and the fact that $\gamma_1(x_{\mathrm{R}}+) = 0$, R reduces to

$$R = -\big|\big(\gamma_1 A(u)_x\big)_- - \big(\gamma_1 A(v)_x\big)_-\big| \leq 0.$$

Case 2 ($v_- = u_-$, $u_+ \neq v_+$). Assume that $v_+ > u_+$. In this case

$$\begin{aligned} R = {} & f(\boldsymbol{\gamma}_+, v_+) - f(\boldsymbol{\gamma}_+, u_+) - \big(\gamma_1 A(v)_x\big)_+ + \big(\gamma_1 A(u)_x\big)_+ \\ & - \big|\big(\gamma_1 A(v)_x\big)_- - \big(\gamma_1 A(u)_x\big)_-\big|, \end{aligned} \tag{4.39}$$

where we have used the equality $f(\boldsymbol{\gamma}_-, v_-) = f(\boldsymbol{\gamma}_-, u_-)$. Via the Rankine–Hugoniot condition and another application of $f(\boldsymbol{\gamma}_-, u_-) = f(\boldsymbol{\gamma}_-, v_-)$, we get

$$f(\boldsymbol{\gamma}_+, v_+) - f(\boldsymbol{\gamma}_+, u_+) = -\big(\gamma_1 A(v)_x\big)_- + \big(\gamma_1 A(u)_x\big)_-.$$

We have again used the fact that $\gamma_1(x_{\mathrm{R}}+) = 0$. Substituting this into (4.39) gives

$$R = -\big|\big(\gamma_1 A(v)_x\big)_- - \big(\gamma_1 A(u)_x\big)_-\big| \leq 0.$$

The situation where $v_+ < u_+$ is handled similarly.

Case 3 ($v_+ = u_+$, $u_- \neq v_-$). The proof of this case is similar to that of Case 2 and is therefore omitted.

Case 4 ($u_- < v_-$, $u_+ < v_+$). In this case, using $\gamma_1(x_{\mathrm{R}}+) = 0$, we obtain from (4.38)

$$\begin{aligned} R = {} & \big(f(\boldsymbol{\gamma}_+, v_+) - f(\boldsymbol{\gamma}_+, u_+)\big) \\ & - \big[f(\boldsymbol{\gamma}_-, v_-) - f(\boldsymbol{\gamma}_+, u_+) - \big(\gamma_1 A(v)_x\big)_- + \big(\gamma_1 A(u)_x\big)_-\big], \end{aligned} \tag{4.40}$$

which equals zero, by the Rankine–Hugoniot condition (4.26).

Case 5 ($u_- > v_-$, $u_+ > v_+$). As in the preceding case, $R = 0$ due to a similar calculation.

Case 6 ($u_- > v_-$, $u_+ < v_+$). In this case, (4.38) becomes

$$R = f(\boldsymbol{\gamma}_-, v_-) + f(\boldsymbol{\gamma}_+, v_+) - f(\boldsymbol{\gamma}_-, u_-) - f(\boldsymbol{\gamma}_+, u_+) - \big(\gamma_1 A(v)_x\big)_- + \big(\gamma_1 A(u)_x\big)_- \tag{4.41}$$

$$= 2f(\boldsymbol{\gamma}_+, v_+) - 2f(\boldsymbol{\gamma}_+, u_+) \tag{4.42}$$

$$= 2\big[f(\boldsymbol{\gamma}_-, v_-) - \big(\gamma_1 A(v)_x\big)_-\big] - 2\big[f(\boldsymbol{\gamma}_-, u_-) - \big(\gamma_1 A(u)_x\big)_-\big],$$

where (4.41) and (4.42) follow from the Rankine–Hugoniot condition (4.26). In (4.41) we have used again the fact that $\gamma_1(x_{\mathrm{R}}+) = 0$. It follows from the assumption $u_- > v_-$, $u_+ < v_+$ that $u_+ < v_+ \leq u_-$ or $v_- < u_- \leq v_+$ must hold. Take the case where $u_+ < v_+ \leq u_-$. Recalling that at $m = x_{\mathrm{R}}$ we always have $f(\boldsymbol{\gamma}_-, \cdot) \geq f(\boldsymbol{\gamma}_+, \cdot)$, we can apply (4.32) of Lemma 4.12, giving us $f(\boldsymbol{\gamma}_+, v_+) \leq f(\boldsymbol{\gamma}_+, u_+)$. With (4.41) in mind, we see that $R \leq 0$. In the case where $v_- < u_- \leq v_+$, (4.33) of Lemma 4.12 yields $f(\boldsymbol{\gamma}_-, v_-) - (\gamma_1 A(v)_x)_- \leq f(\boldsymbol{\gamma}_-, u_-) - (\gamma_1 A(u)_x)_-$, again implying that $R \leq 0$, this time using (4.42).

Case 7 ($u_- < v_-$, $u_+ > v_+$). The proof is identical to that of Case 6; we switch the roles of u and v and use the version of Lemma 4.12 that results by also switching the roles of u and v.

We have established that for any $0 \le \phi \in \mathcal{D}(\Pi_T)$

$$-\iint_{\Pi_T} \Big(|v-u|\phi_t + F(\boldsymbol{\gamma}(x), v, u)\phi_x - \big|\gamma_1 A(v) - \gamma_1 A(u)\big|\phi_{xx} \Big)\, dt\, dx \le 0. \tag{4.43}$$

The proof is concluded via a standard test function argument which is outlined in the proof of Theorem 2.1 of [56]. ∎

4.6. Convergence of the numerical approximations. In what follows, let us denote by Δ the pair $\Delta := (\Delta t, \Delta x)$. Our purpose in this section is to prove convergence (along a subsequence) of the numerical approximations as $\Delta \downarrow 0$, i.e., as $\Delta t, \Delta x \to 0$ with $\Delta t, \Delta x > 0$. For the sake of simplicity, we will concentrate on the explicit version of the algorithm.

Let $(\boldsymbol{\gamma}^\Delta, u^\Delta)$ be the finite difference approximation defined in (3.3). A significant part of the convergence analysis consists of establishing a spatial total variation estimate for the approximate solution u^Δ, measured with respect to a particular transformed variable. More precisely, we prove that u^Δ converges (along a subsequence) to a weak solution by introducing a singular mapping $\Psi : (\boldsymbol{\gamma}, u) \mapsto (\boldsymbol{\gamma}, z)$ such that strong compactness of $z^\Delta = \Psi(\boldsymbol{\gamma}^\Delta, u^\Delta)$ can be obtained. As always in problems involving resonance phenomena, one should measure the space translates with respect to a nonlinear transformation; as already mentioned in the introduction, there is generally no spatial total variation bound for the conserved variable u itself. The singular mapping approach has been used for many years in the purely hyperbolic setting, starting with the paper [76].

On the other hand, the construction of a suitable singular mapping Ψ for second-order equations is more recent and was done first in [55]. The idea is to construct a singular mapping that includes a contribution from both the convective flux and the diffusion function. We first prove compactness for the two parts of the singular mapping separately. We then combine the two portions to recover the original singular mapping and conclude that since the mapping is strictly increasing as a function of the conserved variable u, convergence of the transformed variable implies convergence of u.

Since we are applying the scheme described in section 3 to Model 1 (constant cross section), we can simplify the analysis by taking $S_j \equiv 1$, and then $\lambda_j =: \lambda = \Delta t/\Delta x$. To simplify the notation a little further, let $\mu = \lambda/\Delta x$, $h^n_{j+1/2} = h(\gamma_{j+1/2}, U^n_{j+1}, U^n_j)$, ${\gamma_1}_{j+1/2} = s_{j+1/2}$, and $A^n_j = A(U^n_j)$. The marching formula (3.1) then takes the form

$$U^{n+1}_j = U^n_j - \lambda \Delta_+ h^n_{j-1/2} + \mu \Delta_+ \left(s_{j-1/2} \Delta_- A^n_j \right). \tag{4.44}$$

The Engquist–Osher numerical flux is consistent with the actual flux, i.e., $h(\boldsymbol{\gamma}, u, u) = f(\boldsymbol{\gamma}, u)$. In addition, for fixed $\boldsymbol{\gamma}$, $h(\boldsymbol{\gamma}, v, u)$ is a two-point monotone flux, meaning that it is nonincreasing with respect to v and nondecreasing with respect to u. Due to the regularity assumptions on f, the numerical flux h is Lipschitz continuous with respect to each of its arguments and in fact satisfies

$$f^-_u(\boldsymbol{\gamma}, v) = h_v(\boldsymbol{\gamma}, v, u) \le 0 \le h_u(\boldsymbol{\gamma}, v, u) = f^+_u(\boldsymbol{\gamma}, u), \tag{4.45}$$

where $f^-_u(\boldsymbol{\gamma}, u) := \min\{0, f_u(\boldsymbol{\gamma}, u)\}$ and $f^+_u(\boldsymbol{\gamma}, u) := \max\{0, f_u(\boldsymbol{\gamma}, u)\}$ denote the negative and positive parts of f_u. Thus, whenever the flux $u \mapsto f(\boldsymbol{\gamma}, u)$ is C^1, the

numerical flux is also C^1 as a function of the conserved variables u and v. The following bound is easily checked:

$$\|h\| := \max\Big\{ \big|h(\boldsymbol{\gamma}, v, u)\big| \;\Big|\; \boldsymbol{\gamma} \in \mathcal{G},\ v, u \in [0,1] \Big\} \le \|f\| + \frac{1}{2}\|f_u\|.$$

From formula (4.45) it is clear that $\|f_u\|$ is a Lipschitz constant for the numerical flux h with respect to the conserved variables u and v.

We assume that the discretization parameters $(\Delta x, \Delta t)$ are chosen so that the following Courant–Friedrichs–Lewy (CFL) [35] condition holds:

$$\lambda\Big(\max\{-q_{\mathrm{L}}, q_{\mathrm{R}}\} + \max_{u\in[0,1]} |b'(u)|\Big) + \mu \max_{u\in[0,1]} A'(u) \le 1/2. \tag{4.46}$$

In our case, the stability analysis relies very much on the monotonicity property of the scheme, where we recall that a finite difference scheme such as (4.44) is monotone [36] if

$$U_j^n \le V_j^n \quad \forall j \quad \Longrightarrow \quad U_j^{n+1} \le V_j^{n+1} \quad \forall j. \tag{4.47}$$

The following lemma and its proof illustrate how the CFL condition (4.46) is derived from the requirement that our scheme (4.44) be monotone.

LEMMA 4.16. *If the initial data $\{U_j^0\}_{j\in\mathbb{Z}}$ lies in the interval $[0,1]$ and the CFL condition* (4.46) *is satisfied, then the solution $\{U_j^n\}_{j\in\mathbb{Z}}$ computed by the explicit scheme* (4.44) *also belongs to the interval $[0,1]$ for each $n \ge 0$. Moreover, the difference scheme* (4.44) *remains monotone at each time level $n \ge 0$.*

Proof. Let us first rewrite (4.44) as $U_j^{n+1} = G_j(U_{j+1}^n, U_j^n, U_{j-1}^n, \gamma_{j+1/2}, \gamma_{j-1/2})$ for $j \in \mathbb{Z}$. Then the scheme is monotone, i.e., satisfies (4.47), if

$$\partial U_j^{n+1}/\partial U_{j+1}^n \ge 0, \quad \partial U_j^{n+1}/\partial U_j^n \ge 0, \quad \partial U_j^{n+1}/\partial U_{j-1}^n \ge 0, \quad j \in \mathbb{Z}. \tag{4.48}$$

However, in our case we have for $j \in \mathbb{Z}$

$$\partial U_j^{n+1}/\partial U_{j+1}^n = -\lambda f_u^-\big(\boldsymbol{\gamma}_{j+1/2}, U_{j+1}^n\big) + \mu\gamma_{1\,j+1/2} A'\big(U_{j+1}^n\big), \tag{4.49}$$

$$\partial U_j^{n+1}/\partial U_{j-1}^n = \lambda f_u^+\big(\boldsymbol{\gamma}_{j-1/2}, U_{j-1}^n\big) + \mu\gamma_{1\,j-1/2} A'\big(U_{j-1}^n\big), \tag{4.50}$$

$$\begin{aligned} \partial U_j^{n+1}/\partial U_j^n = {} & 1 + \lambda f_u^-\big(\boldsymbol{\gamma}_{j+1/2}, U_j^n\big) - \lambda f_u^+\big(\boldsymbol{\gamma}_{j-1/2}, U_j^n\big) \\ & - \mu\big(\gamma_{1\,j-1/2} + \gamma_{1\,j+1/2}\big) A'\big(U_j^n\big). \end{aligned} \tag{4.51}$$

Since $f_u^- \le 0$, $f_u^+ \ge 0$, and $A'(u) \ge 0$ by definition, we see that the right-hand sides of (4.49) and (4.50) are always nonnegative. If $U_j^n \in [0,1]$, then it is easy to deduce from (4.51) and from $\gamma_1 \in [0,1]$ that also $\partial U_j^{n+1}/\partial U_j^n \ge 0$ if the CFL condition (4.46) is satisfied. Precisely speaking, $U_j^n \in [0,1]$ and the CFL condition (4.46) ensure that the scheme is monotone *at time* t_n.

It remains to prove that if we have $U_j^n \in [0,1]$, then the quantities U_j^{n+1} calculated by the scheme also satisfy $U_j^{n+1} \in [0,1]$ for $j \in \mathbb{Z}$. To this end, we now apply the scheme (4.44) to the initial data $V_j^0 \equiv 0$. The parabolic terms vanish, since the data is constant, and at time level 1 we get $V_j^1 = V_j^0 - \lambda\Delta_- h(\gamma_{j+1/2}, V_{j+1}^0, V_j^0)$. Since $b(u) = 0$ for $u = 0$ and $u = 1$, it is easy to check that $V_0^1 = \lambda(q_{\mathrm{R}} - q_{\mathrm{L}})u_{\mathrm{F}}$ and $V_j^1 = 0$ for $j \ne 0$. The CFL condition implies that $0 \le \lambda(-q_{\mathrm{L}}) \le 1/2$ and $0 \le \lambda q_{\mathrm{R}} \le 1/2$, which yields $0 \le \lambda(q_{\mathrm{R}} - q_{\mathrm{L}}) \le 1$, and thus $V_j^1 \in [0,1]$.

Next, we apply the scheme (4.44) to the initial data $W_j^0 \equiv 1$, yielding at time level 1 $W_j^1 = W_j^0 - \lambda\Delta_- h(\gamma_{j+1/2}, W_{j+1}^0, W_j^0)$. This time we find that $W_0^1 = 1 - \lambda(q_{\mathrm{R}} - q_{\mathrm{L}})(1 - u_{\mathrm{F}})$ and $W_j^1 = 1$ for $j \neq 0$. The CFL condition again guarantees that $W_j^1 \in [0,1]$. Thus, $0 \leq V_j^1, W_j^1 \leq 1$, the CFL condition remains valid for $\{V_j^1\}_{j\in\mathbb{Z}}$ and $\{W_j^1\}_{j\in\mathbb{Z}}$, and monotonicity implies that $V_j^1 \leq U_j^1 \leq W_j^1$. Continuing inductively, we see that $0 \leq V_j^n \leq U_j^n \leq W_j^n \leq 1$, the CFL condition remains satisfied at each successive time step, and we continue to have monotonicity for each $n \geq 0$. ∎

When sending $\Delta \downarrow 0$, as we will do in the analysis of the explicit scheme (3.1), the ratio $\mu = \lambda/\Delta x = \Delta t/\Delta x^2$ will be kept constant, which means that $\lambda = \mu\Delta x$ is variable with $\lambda \to 0$ as $\Delta \downarrow 0$.

The CFL condition for the semi-implicit scheme (3.4), which we do not analyze here but use for some of the numerical examples, is

$$\lambda\Big(\max\{-q_{\mathrm{L}}, q_{\mathrm{R}}\} + \max_{u\in[0,1]} |b'(u)|\Big) \leq 1/2. \tag{4.52}$$

Consequently, the semi-implicit scheme behaves stably for $\Delta \downarrow 0$ if we fix $\lambda = \Delta t/\Delta x$ such that (4.52) is satisfied. The semi-implicit scheme (3.4) allows a faster computation than the explicit scheme (3.1), since Δt needs to be chosen proportional to Δx, not Δx^2 (as for the explicit scheme). Again, the CFL condition (4.52) accrues from the requirement that the scheme be monotone, as we shall see in the following version of Lemma 4.16 for the semi-implicit scheme (3.4). This lemma can be considered as a motivation for the CFL condition (4.52).

LEMMA 4.17. *If the initial data $\{U_j^0\}_{j\in\mathbb{Z}}$ lies in the interval $[0,1]$ and the CFL condition* (4.52) *is satisfied, then the solution $\{U_j^n\}_{j\in\mathbb{Z}}$ computed by the semi-implicit scheme* (3.4) *also belongs to the interval $[0,1]$ for each $n \geq 0$. Moreover, the difference scheme* (3.4) *remains monotone at each time level $n \geq 0$.*

Proof. Let $V^n := \{V_j^n\}_{j\in\mathbb{Z}}$ and $W^n := \{W_j^n\}_{j\in\mathbb{Z}}$ satisfy $V_j^n, W_j^n \in [0,1]$ for all $j \in \mathbb{Z}$. If we compute V^{n+1} and W^{n+1} using the implicit scheme, then with the help of (4.45) we can write their difference as

$$\begin{aligned} W_j^{n+1} - V_j^{n+1} = {} & W_j^n - V_j^n + \alpha_{j+1/2}\left(W_{j+1}^n - V_{j+1}^n\right) - \beta_{j+1/2}\left(W_j^n - V_j^n\right) \\ & - \alpha_{j-1/2}\left(W_j^n - V_j^n\right) + \beta_{j-1/2}\left(W_{j-1}^n - V_{j-1}^n\right) \\ & + s_{j+1/2}\theta_{j+1}\left(W_{j+1}^{n+1} - V_{j+1}^{n+1}\right) \\ & - (s_{j+1/2} + s_{j-1/2})\theta_j\left(W_j^{n+1} - V_j^{n+1}\right) \\ & + s_{j-1/2}\theta_{j-1}\left(W_{j-1}^{n+1} - V_{j-1}^{n+1}\right), \end{aligned} \tag{4.53}$$

where we define for $j \in \mathbb{Z}$

$$\begin{aligned} \alpha_{j+1/2} &:= -\lambda \int_0^1 f_u^-\left(\boldsymbol{\gamma}_{j+1/2}, V_{j+1}^n + \phi(W_{j+1}^n - V_{j+1}^n)\right) d\phi \geq 0, \\ \beta_{j+1/2} &:= \lambda \int_0^1 f_u^+\left(\boldsymbol{\gamma}_{j+1/2}, V_j^n + \phi(W_j^n - V_j^n)\right) d\phi \geq 0, \end{aligned} \tag{4.54}$$

and

$$\theta_j := \mu \frac{A(W_j^{n+1}) - A(V_j^{n+1})}{W_j^{n+1} - V_j^{n+1}} \geq 0. \tag{4.55}$$

Let us abbreviate $D_j^n := W_j^n - V_j^n$ and rearrange (4.53) into the form

$$\begin{aligned}\big(1+(s_{j+1/2}+s_{j-1/2})\theta_j\big)D_j^{n+1} &= (1-\beta_{j+1/2}-\alpha_{j-1/2})D_j^n + \alpha_{j+1/2}D_{j+1}^n \\ &\quad + \beta_{j-1/2}D_{j-1}^n + s_{j+1/2}\theta_{j+1}D_{j+1}^{n+1} \\ &\quad + s_{j-1/2}\theta_{j-1}D_{j-1}^{n+1}.\end{aligned} \tag{4.56}$$

Thanks to the CFL condition (4.52), we have $1-\beta_{j+1/2}-\alpha_{j-1/2}\geq 0$. Thus all of the coefficients appearing in (4.56) are nonnegative, and taking absolute values results in

$$\begin{aligned}\big(1+(s_{j+1/2}+s_{j-1/2})\theta_j\big)\big|D_j^{n+1}\big| &\leq (1-\beta_{j+1/2}-\alpha_{j-1/2})\big|D_j^n\big| \\ &\quad + \alpha_{j+1/2}\big|D_{j+1}^n\big| + \beta_{j-1/2}\big|D_{j-1}^n\big| \\ &\quad + s_{j+1/2}\theta_{j+1}\big|D_{j+1}^{n+1}\big| + s_{j-1/2}\theta_{j-1}\big|D_{j-1}^{n+1}\big|.\end{aligned} \tag{4.57}$$

Summing (4.57) over $j\in\mathbb{Z}$, canceling wherever possible, and recalling the definition of D_j^n, we find that

$$\sum_{j\in\mathbb{Z}}\big|W_j^{n+1}-V_j^{n+1}\big| \leq \sum_{j\in\mathbb{Z}}\big|W_j^n-V_j^n\big|, \tag{4.58}$$

indicating that the semi-implicit scheme is L^1-contractive on data that is constrained to the interval $[0,1]$. It now follows from the Crandall–Tartar lemma [37] that the scheme is also monotone (on data that is constrained to the interval $[0,1]$).

We still must show that the solution remains in $[0,1]$. First, observe that the L^1-contraction property (4.58) implies that the solution to the implicit scheme is unique. Referring back to the portion of the proof of Lemma 4.16 where we used the specific data $V_j^0\equiv 0$, and $W_j^0\equiv 1$, we see that the solutions V_j^1 and W_j^1 are also (the unique) solutions to the implicit scheme at time level 1. With these observations in mind, it is clear that the relevant portion of the proof of Lemma 4.16 also shows invariance of the interval $[0,1]$ for the semi-implicit scheme. ∎

We now continue our analysis of the explicit scheme (3.1).

LEMMA 4.18. *Our numerical approximation satisfies the following discrete time continuity estimate (which is uniform in n and Δ):*

$$\Delta x\sum_{j\in\mathbb{Z}}\big|U_j^{n+1}-U_j^n\big| \leq \Delta x\sum_{j\in\mathbb{Z}}\big|U_j^1-U_j^0\big| \leq C\Delta t. \tag{4.59}$$

It also satisfies a uniform (in n and Δ) L^1 bound:

$$\|u^\Delta(\cdot,t^n)\|_{L^1(\mathbb{R})} \leq CT. \tag{4.60}$$

Proof. The proof of Lemma 3.3 of [55] is almost entirely applicable to (4.59), the only possible complication arising when we have to bound the quantity

$$\sum_{j\in\mathbb{Z}}\left|\Delta_-\frac{1}{\Delta x}s_{j+1/2}\Delta_+A(U_j^0)\right|.$$

One finds that the proof of the analogous bound in [55] can be modified to accommodate the present situation. The key ingredients are the assumption that $\gamma_1 A(u_0)_x\in BV(\mathbb{R})$, along with the pointwise discretization of u_0. For the proof of (4.60), see Lemma 3.4 of [55]. ∎

In what follows, it will be convenient to have available the notation $\mathcal{O}(1)$ to denote a quantity that is bounded uniformly in n and Δ.

LEMMA 4.19. *The following bound holds independently of Δ and the time level n:*

$$\sum_{j\in\mathbb{Z}} s_{j-1/2}\big|\Delta_- A_j^n\big| \le C. \tag{4.61}$$

Proof. Let $\rho_{j-1/2}^n := h_{j-1/2}^n - s_{j-1/2}\Delta_- A_j^n/\Delta x$. By substituting $U_j^{n+1} - U_j^n = -\lambda\Delta_+\rho_{j-1/2}^n$ into (4.59) we find that

$$\sum_{j\in\mathbb{Z}} \big|\Delta_+\rho_{j-1/2}^n\big| = \mathcal{O}(1). \tag{4.62}$$

At the same time, if j is so large that $x_j > x_{\mathrm{R}} + 2\Delta x$, then $\rho_{j-1/2}^n = q_{\mathrm{R}}U_j^n$. From Lemma 4.16, we get that $|\rho_{j-1/2}^n| \le q_{\mathrm{R}}$. This bound, together with the bound (4.62), implies a uniform bound of the form $|\rho_{j-1/2}^n| = \mathcal{O}(1)$. Since the convective numerical flux $h(\gamma, v, u)$ is continuous, the quantity $h_{j-1/2}^n$ is also uniformly bounded, and so we have the bound $s_{j-1/2}|\Delta_- A_j^n|/\Delta x = \mathcal{O}(1)$. The proof is completed by multiplying both sides of this relationship by Δx, summing over j, and recalling that $s_{j-1/2}$ vanishes for x_j outside of the interval $[x_{\mathrm{L}} - \Delta x, x_{\mathrm{R}}]$. ∎

Let $z^\Delta(x,t) := \Psi(\boldsymbol{\gamma}(x), u^\Delta(x,t))$. Defining (see (4.13) for the definition of $\mathcal{F}$)

$$\mathcal{F}^\Delta(x,t) := \mathcal{F}(\boldsymbol{\gamma}(x), u^\Delta(x,t)), \quad A^\Delta(x,t) := A(u^\Delta(x,t)),$$

we can separate z^Δ into its hyperbolic and parabolic contributions:

$$z^\Delta(x,t) = \mathcal{F}^\Delta(x,t) + \gamma_1(x)A^\Delta(x,t). \tag{4.63}$$

To prove that the difference scheme converges, we follow [25] and first prove compactness for the transformed quantity z^Δ. We establish spatial variation bounds separately for each of the intervals $(-\infty, x_{\mathrm{L}})$, $(x_{\mathrm{L}}, 0)$, $(0, x_{\mathrm{R}})$, (x_{R}, ∞). The jumps in z^Δ where these intervals meet are bounded, and so we can ignore them. Indeed consider the jump in $z^\Delta(x, t^n)$ that occurs at $x = m \in \{x_{\mathrm{L}}, 0, x_{\mathrm{R}}\}$, which is given by

$$z^\Delta(m+, t^n) - z^\Delta(m-, t^n) = \Psi\big(\boldsymbol{\gamma}+, u^\Delta(m+, t^n)\big) - \Psi\big(\boldsymbol{\gamma}-, u^\Delta(m-, t^n)\big).$$

Since u^Δ is bounded uniformly (by Lemma 4.16), $\boldsymbol{\gamma}$ is bounded by assumption, and the transformation Ψ is Lipschitz continuous with respect to all variables, it is clear that the magnitude of this jump is uniformly bounded also.

In the intervals $(-\infty, x_{\mathrm{L}})$ and (x_{R}, ∞), the parabolic term is not present, and (4.63) simplifies to $z^\Delta(x,t) = \gamma_2 u^\Delta(x,t)$. This makes it clear that the proof of the variation bound for these intervals is the same as the proof of Lemma 3.5 of [25]. We record this fact in the following lemma.

LEMMA 4.20. *We have the following bounds, which are independent of Δ and n:*

$$\mathrm{TV}\big(z^\Delta(\cdot, t^n)|_{\{x|x<x_{\mathrm{L}}\}}\big) \le C, \qquad \mathrm{TV}\big(z^\Delta(\cdot, t^n)|_{\{x|x>x_{\mathrm{R}}\}}\big) \le C. \tag{4.64}$$

We now address the variation bound for the remaining intervals, $(x_{\mathrm{L}}, 0)$ and $(0, x_{\mathrm{R}})$. As in [25], we will focus on $(0, x_{\mathrm{R}})$, omitting the proof for the other interval, since it is similar. Let $\boldsymbol{\gamma}_{\mathrm{R}} := (q_{\mathrm{R}}, 1)$; i.e., let $\boldsymbol{\gamma}_{\mathrm{R}}$ denote the value that $\boldsymbol{\gamma}$ takes on

$(0, x_{\mathrm{R}})$. Recalling the definition (4.13), we see that $\gamma_1 = 1$ for $x \in (0, x_{\mathrm{R}})$, and so Ψ simplifies to

$$\Psi(\boldsymbol{\gamma}, u) = \mathcal{F}(\boldsymbol{\gamma}_{\mathrm{R}}, u) + A(u), \quad \mathcal{F}(\boldsymbol{\gamma}_{\mathrm{R}}, u) = \int_0^u \mathcal{S}(w) |f_u(\boldsymbol{\gamma}_{\mathrm{R}}, w)| \, dw. \tag{4.65}$$

Let J^- be the largest index j such that $x_j - \Delta x/2 \le 0$, and let J^+ be the smallest index j such that $x_j + \Delta x/2 \ge x_{\mathrm{R}}$. Thus $0 \in I_{J^-}$, $x_{\mathrm{R}} \in I_{J^+}$, and $[0, x_{\mathrm{R}}] \subseteq [x_{J^-} - \Delta x/2, x_{J^+} + \Delta x/2]$.

The following lemma records a discrete entropy inequality. It can be proved via a slight modification (to account for $s_{j-1/2}$) of the proof of Lemma 4.1 of [56].

LEMMA 4.21. *For any $c \in \mathbb{R}$, the following cell entropy inequality is satisfied by approximate solutions U_j^n generated by the scheme* (4.44)*:*

$$\begin{aligned} \left|U_j^{n+1} - c\right| \le \left|U_j^n - c\right| &- \lambda \Delta_- H_{j+1/2}^n + \mu \Delta_+ \left(s_{j-1/2} \Delta_- \left| A\left(U_j^n\right) - A(c) \right| \right) \\ &- \lambda \operatorname{sgn}\left(U_j^{n+1} - c\right) \Delta_+ f\left(\gamma_{j-1/2}, c\right), \end{aligned} \tag{4.66}$$

where the numerical entropy flux $H_{j+1/2}$ is defined by

$$H_{j+1/2}^n := h\left(\gamma_{j+1/2}, U_{j+1}^n \vee c, U_j^n \vee c\right) - h\left(\gamma_{j+1/2}, U_{j+1}^n \wedge c, U_j^n \wedge c\right). \tag{4.67}$$

Formally, the cell entropy inequality (4.66) can be motivated by assuming that the function u in the integrand of (4.5) is piecewise constant on the rectangle $R_j^n := (x_{j-1/2}, x_{j+1/2}) \times (t_n, t_{n+1})$ and by choosing a sequence of test functions ϕ with support on R_j^n that approximate the characteristic function χ_j^n of R_j^n. Moreover, the exact entropy flux defined in (4.12) is replaced by the numerical entropy flux (4.67). In this sense, the discrete entropy inequality (4.66) is consistent with the entropy inequality (4.5) for the exact solution, but observe that the term in the second line of (4.66), which mirrors the sum over $m \in \mathcal{J}$ in (4.5), is evaluated at time level t_{n+1}.

Let $\chi_l(w; c) := H(c - w)$, where $H(\cdot)$ is the Heaviside function, and $\chi_r(w; c) := 1 - \chi_l(w; c)$. The following lemma is easily established using the calculations used in Lemma 3.9 of [25], adapted to the cell entropy inequality (4.66) appearing in Lemma 4.21.

LEMMA 4.22. *Fix $c \in \mathbb{R}$ and $\boldsymbol{\gamma} \in \mathcal{G}$. The following inequalities are valid for $J^- \le j \le J^+$:*

$$\begin{aligned} &- \int_{U_j^n}^{U_{j+1}^n} \chi_l(w; c) f_u^-(\boldsymbol{\gamma}, w) \, dw - \int_{U_{j-1}^n}^{U_j^n} \chi_l(w; c) f_u^+(\boldsymbol{\gamma}, w) \, dw \\ &\le \frac{-1}{\lambda} \left(U_j^n - U_j^{n+1}\right)_- - \frac{1}{\Delta x} \Delta_+ \left(s_{j-1/2} \Delta_- (A(U_j^{n+1}) - A(c))_- \right) + \alpha_j^n, \end{aligned} \tag{4.68}$$

$$\begin{aligned} &\int_{U_j^n}^{U_{j+1}^n} \chi_r(w; c) f_u^-(\boldsymbol{\gamma}, w) \, dw + \int_{U_{j-1}^n}^{U_j^n} \chi_r(w; c) f_u^+(\boldsymbol{\gamma}, w) \, dw \\ &\le \frac{1}{\lambda} \left(U_j^n - U_j^{n+1}\right)_+ - \frac{1}{\Delta x} \Delta_+ \left(s_{j-1/2} \Delta_- (A(U_j^{n+1}) - A(c))_+ \right) + \beta_j^n. \end{aligned} \tag{4.69}$$

The quantities α_j^n and β_j^n are bounded independently of n and Δ. In fact, $\alpha_j^n = \beta_j^n = 0$ for $J^- + 2 \le j \le J^+ - 2$.

With the help of these entropy inequalities, we can prove the following lemma.

LEMMA 4.23. *The following spatial variation bounds are satisfied, independent of* Δ *and* n*:*

$$\mathrm{TV}\big(\mathcal{F}^{\Delta}(\cdot,t^n)|_{\{x|x_{\mathrm{L}}<x<0\}}\big)\le C,\qquad \mathrm{TV}\big(\mathcal{F}^{\Delta}(\cdot,t^n)|_{\{x|0<x<x_{\mathrm{R}}\}}\big)\le C.$$

Proof. We prove only the second assertion; the proof of the first is similar. We follow closely the proof of Lemma 3.10 of [25]. First, observe that if the term

$$\frac{1}{\Delta x}\Delta_-\big(\Delta_+ s_{j-1/2}(A(U_j^n)-A(c))_-\big)$$

was not present in (4.68), the proof of Lemma 3.10 of [25] would apply verbatim. Next, recall from the proof of Lemma 4.5 that if $c\le u_{\mathrm{c}}$, then $(A(U_j^n)-A(c))_-=0$. Thus, when $c\le u_{\mathrm{c}}$, the parabolic term in (4.68) disappears, giving us

(4.70)
$$-\int_{U_j^n}^{U_{j+1}^n}\chi_l(w;c)f_u^-(\boldsymbol{\gamma},w)\,dw-\int_{U_{j-1}^n}^{U_j^n}\chi_l(w;c)f_u^+(\boldsymbol{\gamma},w)\,dw\le\frac{-1}{\lambda}\big(U_j^n-U_j^{n+1}\big)_-+\alpha_j^n.$$

When $x\in(0,x_{\mathrm{R}})$, $(\gamma_1(x),\gamma_2(x))=(q_{\mathrm{R}},1)\equiv\boldsymbol{\gamma}_{\mathrm{R}}$, and by the assumptions on b and q_{R}, $u\mapsto f(\boldsymbol{\gamma}_{\mathrm{R}},u)$ has at most two extrema for $u\in(0,1)$. For the sake of argument, we assume that there are exactly two extrema. It will become clear that a simplified version of the following proof will suffice if there are fewer than two. So assume that there is one maximum located at $u_1^*\in(0,1)$ and one minimum located at $u_2^*\in(0,1)$, with $u_1^*<u_2^*$. The flux $u\mapsto f(\boldsymbol{\gamma}_{\mathrm{R}},u)$ is strictly monotone away from these critical points. Let $u_0^*:=0$ and $u_3^*:=1$, and for $\nu=0,1,2$, let $\chi^\nu(u)$ be the characteristic function of the interval $[\min\{u_\nu^*,u_{\mathrm{c}}\},\min\{u_{\nu+1}^*,u_{\mathrm{c}}\})$. Each of the intervals $[\min\{u_\nu^*,u_{\mathrm{c}}\},\min\{u_{\nu+1}^*,u_{\mathrm{c}}\})$ either is empty (if the left endpoint happens to equal u_{c}) or $f(\gamma_{\mathrm{R}},u)$ is strictly monotone in its interior. Define

$$\phi^\nu(\boldsymbol{\gamma}_{\mathrm{R}},u):=\int_0^u\chi^\nu(w)\big|f_u(\boldsymbol{\gamma}_{\mathrm{R}},w)\big|\,dw,\qquad\nu=0,1,2.$$

Clearly, $\mathcal{S}(u)=\chi^0(u)+\chi^1(u)+\chi^2(u)$, so that $\mathcal{F}(\boldsymbol{\gamma}_{\mathrm{R}},\cdot)$ has the decomposition

$$\mathcal{F}(\boldsymbol{\gamma}_{\mathrm{R}},u)=\phi^0(\boldsymbol{\gamma}_{\mathrm{R}},u)+\phi^1(\boldsymbol{\gamma}_{\mathrm{R}},u)+\phi^2(\boldsymbol{\gamma}_{\mathrm{R}},u).\tag{4.71}$$

We now use the entropy inequality (4.70) three times, just as in the proof of Lemma 3.10 of [25], except that now instead of $c=u_\nu$, $\nu=1,2,3$, we take $c=\min\{u_\nu^*,u_{\mathrm{c}}\}$, $\nu=1,2,3$. In order to keep the analysis somewhat self-contained, let us review the calculation appearing in [25] when $c=u_1^*$. We start by setting $c=u_1^*$ in inequality (4.70) and observe that $u\mapsto f(\boldsymbol{\gamma}_{\mathrm{R}},u)$ is strictly increasing on $(0,u_1^*)$. Then (4.70) simplifies to

$$-\int_{U_{j-1}^n}^{U_j^n}\chi_l(w;u_1^*)f_u^+(\boldsymbol{\gamma}_{\mathrm{R}},w)\,dw\le\frac{-1}{\lambda}\big(U_j^n-U_j^{n+1}\big)_-+\alpha_j^n.\tag{4.72}$$

Since $f_u^+(\boldsymbol{\gamma}_{\mathrm{R}},u)=|f_u(\boldsymbol{\gamma}_{\mathrm{R}},u)|$ for $u\in(0,u_1^*)$, we find that

$$\begin{aligned}\int_{U_{j-1}^n}^{U_j^n}\chi_l(w;u_1^*)f_u^+(\boldsymbol{\gamma}_{\mathrm{R}},w)\,dw&=\int_{U_{j-1}^n}^{U_j^n}\chi^0(w)\big|f_u(\boldsymbol{\gamma}_{\mathrm{R}},w)\big|dw\\&=\phi^0\big(\boldsymbol{\gamma}_{\mathrm{R}},U_j^n\big)-\phi^0\big(\boldsymbol{\gamma}_{\mathrm{R}},U_{j-1}^n\big).\end{aligned}$$

Combining this last relationship with (4.72), we have the inequality

$$\phi^0\big(\boldsymbol{\gamma}_{\mathrm{R}}, U_{j-1}^n\big) - \phi^0\big(\boldsymbol{\gamma}_{\mathrm{R}}, U_j^n\big) \le \frac{1}{\lambda}\big|U_j^{n+1} - U_j^n\big| + \alpha_j^n,$$

and, since the right-hand side of this inequality is nonnegative, we also have

$$-\Big(\phi^0\big(\boldsymbol{\gamma}_{\mathrm{R}}, U_j^n\big) - \phi^0\big(\boldsymbol{\gamma}_{\mathrm{R}}, U_{j-1}^n\big)\Big)_- \le \frac{1}{\lambda}\big|U_j^{n+1} - U_j^n\big| + \alpha_j^n. \tag{4.73}$$

Next, we sum (4.73) over j and invoke Lemmas 4.22 and 4.18 to obtain

$$\begin{aligned}
&-\sum_{j=J^-}^{J^+}\Big(\phi^0\big(\boldsymbol{\gamma}_{\mathrm{R}}, U_j^n\big) - \phi^0\big(\boldsymbol{\gamma}_{\mathrm{R}}, U_{j-1}^n\big)\Big)_- \le \sum_{j=J^-}^{J^+}\left(\frac{1}{\lambda}\big|U_j^{n+1} - U_j^n\big| + \big|\alpha_j^n\big|\right)\\
&\le \sum_{j\in\mathbb{Z}}\frac{1}{\lambda}\big|U_j^{n+1} - U_j^n\big| + \big|\alpha_{J^+}^n\big| + \big|\alpha_{J^+-1}^n\big| + \big|\alpha_{J^-}^n\big| + \big|\alpha_{J^-+1}^n\big| = \mathcal{O}(1).
\end{aligned}$$

Finally, we observe that since ϕ^0 is bounded uniformly in Δ and n, it follows from this bound on the negative variation that ϕ^0 also has uniformly bounded *total* variation. Similar calculations (see [25]) result in uniform bounds on the total variation of ϕ^1 and ϕ^2; i.e., we have

$$\sum_{j=J^-}^{J^+}\Big|\phi^\nu\big(\boldsymbol{\gamma}_{\mathrm{R}}, U_j^n\big) - \phi^\nu\big(\boldsymbol{\gamma}_{\mathrm{R}}, U_{j-1}^n\big)\Big| = \mathcal{O}(1), \quad \nu = 0, 1, 2.$$

In view of (4.71), the proof is completed by combining these three bounds. ∎

With this spatial variation bound established, we can prove the following lemma. We omit the proof, which is not essentially different from the proof of Lemma 3.8 of [55].

LEMMA 4.24. *There exists a subsequence of $\{\mathcal{F}^\Delta\}$, also denoted by $\{\mathcal{F}^\Delta\}$, and a function $\overline{\mathcal{F}} \in L^1(\Pi_T) \cap L^\infty(\Pi_T)$ such that $\mathcal{F}^\Delta \to \overline{\mathcal{F}}$ in $L^1_{\mathrm{loc}}(\Pi_T)$ and a.e. in Π_T. Furthermore, $\overline{\mathcal{F}}(\cdot, t) \in L^1(\mathbb{R})$ for all $t \in [0, T]$.*

The following lemma establishes convergence (along a subsequence) of the discrete diffusion term A^Δ. The proof is similar to the proofs of Lemmas 3.9, 3.10, 3.11, and 3.12 of [55].

LEMMA 4.25. *The following bounds are satisfied, independent of n and Δ:*

$$\left\{\Delta t\,\Delta x\sum_{n\ge 0}\sum_{j\in\mathbb{Z}} s_{j+1/2}\big(\Delta_+ A(U_j^n)\big)^2\right\}^{1/2} \le C\Delta x, \tag{4.74}$$

$$\big\|A\big(u^\Delta(\cdot + y, \cdot)\big) - A\big(u^\Delta(\cdot, \cdot)\big)\big\|_{L^2(\Omega_y)} \le C\sqrt{|y|(|y| + \Delta x)} \qquad \forall y \in (x_{\mathrm{L}}, x_{\mathrm{R}}),$$

$$\big\|A\big(u^\Delta(\cdot, \cdot + \tau)\big) - A\big(u^\Delta(\cdot, \cdot)\big)\big\|_{L^2(\Omega_\tau)} \le C\sqrt{\tau + \Delta t} \qquad \forall \tau \in (0, T), \tag{4.75}$$

where Ω_y consists of all $(x, t) \in \Pi_T$ such that x and $x + y$ belong to $(x_{\mathrm{L}}, x_{\mathrm{R}}) \times (0, T)$ and $\Omega_\tau := (x_{\mathrm{L}}, x_{\mathrm{R}}) \times (0, T - \tau))$. Finally, we have that there exists a subsequence of $\{A^\Delta\}$, also denoted by $\{A^\Delta\}$, and a function $\overline{A} \in L^2(0, T; H^1(x_{\mathrm{L}}, x_{\mathrm{R}}))$ such that $A^\Delta \to \overline{A}$ in $L^2((x_{\mathrm{L}}, x_{\mathrm{R}}) \times (0, T)$ and boundedly a.e. in $(x_{\mathrm{L}}, x_{\mathrm{R}}) \times (0, T)$. Furthermore, $\overline{A} = A(u)$ a.e. in $(x_{\mathrm{L}}, x_{\mathrm{R}}) \times (0, T)$, where u denotes the L^∞ weak-$*$ limit of u^Δ.

It is possible to establish more regularity of the diffusion function than displayed in Lemma 4.25. This additional regularity, which is stated in the lemma below, will be used later in the proof of Theorem 4.27.

LEMMA 4.26. *There exists a constant C, independent of Δ, such that*

$$\big|A(U_j^n) - A(U_i^n)\big| \le C|j-i|\Delta x, \quad \big|A(U_j^n) - A(U_j^m)\big| \le C\sqrt{|n-m|\Delta t}$$

for all i, j, n, m such that (x_i, t_n), (x_j, t_n), and (x_j, t_m) belong to $(x_{\mathrm{L}}, x_{\mathrm{R}}) \times (0, T)$.

Define $\tilde{A}^n(x)$ as

$$\tilde{A}^n(x) = \frac{1}{\Delta x}\Big((x - x_{j-1})\, A(U_j^n) + (x_j - x)\, A(U_{j-1}^n)\Big), \qquad x \in [x_{j-1}, x_j].$$

Then define

$$\tilde{A}^\Delta(x,t) = \frac{1}{\Delta t}\Big((t - t_n)\, \tilde{A}^{n+1}(x) + (t_{n+1} - t)\, \tilde{A}^n(x)\Big), \qquad t \in [t_n, t_{n+1}].$$

Then there exists a subsequence of $\tilde{A}^\Delta$, also denoted by $\tilde{A}^\Delta$, and a function

$$\tilde{A} \in C^{1,1/2}\big((x_{\mathrm{L}}, x_{\mathrm{R}}) \times (0,T)\big)$$

such that $\tilde{A}^\Delta \to \bar{A}$ in $L^\infty((x_{\mathrm{L}}, x_{\mathrm{R}}) \times (0,T))$. Moreover, there holds $(\tilde{A}^\Delta)_x \overset{\star}{\rightharpoonup} \tilde{A}_x$ in $L^\infty((x_{\mathrm{L}}, x_{\mathrm{R}}) \times (0,T))$.

This lemma can be proved by a straightforward adaptation of the proofs of Lemmas 4.1 and 4.2 and Theorem 4.1 in [55].

We can now prove our main convergence theorem.

THEOREM 4.27. *Assume that the hypotheses concerning the data stated in section* 4.1 *are satisfied. Then there exists a BV_t weak solution of the initial value problem* (4.1) *that satisfies the entropy condition* (D.5). *Let u^Δ be defined by* (3.3) *and the scheme* (4.44), *with the parameters Δx and Δt chosen so that the CFL condition* (4.46) *holds. Then, along a subsequence, $u^\Delta \to u$ in $L^1_{\mathrm{loc}}(\Pi_T)$ and a.e. in Π_T, where u is a BV_t weak solution.*

Proof. The proof of convergence (along a subsequence) to a function $u : \Pi_T \to \mathbb{R}$ is essentially the same as the proof of Theorem 3.1 of [55]. The main idea is to observe that $z^\Delta = \Psi(\boldsymbol{\gamma}^\Delta, u^\Delta) = \mathcal{F}^\Delta + \gamma_1 A^\Delta$. Convergence (along a subsequence) of $\{z^\Delta\}$ then follows from compactness for the sequences $\mathcal{F}^\Delta$ and $\gamma_1 A^\Delta$ (Lemmas 4.24 and 4.25). Letting $z(x,t)$ denote $\lim_{\Delta \to 0} z^\Delta(x,t)$, one then recovers the conserved quantity u via $u(x,t) = \Psi^{-1}(\boldsymbol{\gamma}(x), z(x,t))$. The arguments in [55] also (with some slight modifications to account for γ_1 multiplying $A(u)_x$) show that $u \in L^1(\Pi_T) \cap L^\infty(\Pi_T) \cap C(0,T; L^1(\mathbb{R}))$ and $A(u) \in L^2(0,T; H^1(x_{\mathrm{L}}, x_{\mathrm{R}}))$. It follows readily from the discrete time continuity estimate (4.59) that $u \in BV_t(\Pi_T)$ (see the proof of Theorem 3.1 of [25]) and that the initial data is assumed in the strong L^1 sense; i.e, (4.4) is satisfied.

To show that the limit u is a BV_t weak solution, it remains to verify that the weak formulation (4.3) is satisfied, for which a a Lax–Wendroff-type calculation is required. The proof of Theorem 3.1 of [55] applies in the present situation, with the exception that the spatially varying coefficient $s_{j-1/2}$ multiplying the parabolic term causes some new complications. We can lay this matter to rest if we can show that for $\phi \in \mathcal{D}(\Pi_T)$, and with $\phi_j^n := \phi(x_j, t_n)$,

$$\Delta x \Delta t \sum_{n \ge 0} \sum_{j \in \mathbb{Z}} \frac{1}{\Delta x^2} \Delta_+ \left(s_{j-1/2} \Delta_- A_j^n\right) \phi_j^n \to - \iint_{\Pi_T} \gamma_1(x) A(u)_x \phi_x \, dx \, dt. \tag{4.76}$$

Summing by parts, we get the following expression for the left-hand side of (4.76):

$$-\Delta x\Delta t\sum_{n\geq 0}\sum_{j\in\mathbb{Z}}\frac{1}{\Delta x}\left(s_{j-1/2}\Delta_- A_j^n\right)\left(\Delta_-\phi_j^n/\Delta x\right). \tag{4.77}$$

Let $\tilde{A}^\Delta$ be the interpolant defined in Lemma 4.26. Observe that $(\tilde{A}^\Delta)_x = \Delta_+ A(U_j^n)$ on the parallelogram P_j^n with vertices (x_j, t^{n-1}), (x_j, t^n), (x_{j+1}, t^n), and (x_{j+1}, t^{n+1}).

We now have

$$\begin{aligned}
&-\Delta x\Delta t\sum_{n\geq 0}\sum_{j\in\mathbb{Z}}\frac{1}{\Delta x}\left(s_{j-1/2}\Delta_- A_j^n\right)\left(\Delta_-\phi_j^n/\Delta x\right)\\
&=\sum_{n\geq 0}\sum_{j\in\mathbb{Z}}\iint_{P_j^{n+1}}\gamma_1(x)(\tilde{A}^\Delta)_x\phi_x\,dx\,dt+\mathcal{O}(\Delta x+\Delta t)\\
&=\iint_{\Pi_T}\gamma_1(x)(\tilde{A}^\Delta)_x\phi_x\,dx\,dt+\mathcal{O}(\Delta x+\Delta t).
\end{aligned} \tag{4.78}$$

According to Lemma 4.26 we can assume that $\tilde{A}^\Delta \to \bar{A}$ in $L^\infty((x_{\mathrm{L}}, x_{\mathrm{R}})\times(0,T))$. Since $u^\Delta \to u$ a.e. in Π_T, we can repeat the proof of Theorem 4.1 in [55] to show that $\tilde{A} = A(u)$ a.e. in $(x_{\mathrm{L}}, x_{\mathrm{R}})\times(0,T)$. Recall that the parameter $\gamma_1(x)$ takes the value 1 for $x \in (x_{\mathrm{L}}, x_{\mathrm{R}})$ and is zero elsewhere. Using this and the convergence $(\tilde{A}^\Delta)_x \overset{\star}{\rightharpoonup} A(u)_x$ in $L^\infty((x_{\mathrm{L}}, x_{\mathrm{R}})\times(0,T))$ when sending $\Delta \to 0$ in (4.78), we get (4.76).

The proof that u satisfies the entropy inequality (4.5) requires another Lax–Wendroff-type calculation, this time based on the cell entropy inequality (4.66). The proof of Theorem 5.1 of [55], or Lemma 4.1 of [56], suffices for the situation at hand, again with the exception of the parabolic terms, due to the presence of $s_{j-1/2}$. It is possible to resolve this matter by an argument (which we omit) similar to the one above. ∎

REMARK 4.3. *In this proof, we have verified all but condition* (D.4) *of Definition* 4.1. *Thus, if we were able to prove that* (D.4) *is satisfied, then the limit* u *of Theorem* 4.27 *would be the* BV_t *entropy solution whose uniqueness is guaranteed by Theorem* 4.15. *Although our numerical results suggest that this condition is satisfied by the limit of the sequence of approximate solutions, a rigorous proof of this property is still left as an open problem.*

5. Steady-state solutions. The construction of steady states is based on the stationary version of (2.16) or (2.21). We do not present here a thorough analysis of all steady states but identify some stationary solutions in order to motivate the choices of the control parameters for the transient simulations. Our construction of steady states will follow a procedure similar to that of the simpler continuous thickening models treated in [16, 17]. Specifically, we fix the material model and the vessel geometry and assume that the clarifier-thickener is to be operated at given values of Q_{L}, Q_{F}, and u_{F} and is supposed to produce a thickened sediment of a discharge concentration $u_{\mathrm{D}} > u_{\mathrm{c}}$. Although the construction given below can be extended in a straightforward manner to vessels with varying cross-sectional area, there are some subtle details that require us to restrict the rigorous discussion to vessels with constant cross-sectional interior area, so that we limit the discussion to Model 1.

Our notation is consistent with section 4; i.e., we refer to the space variable by x (instead of w), to the solution by u (instead of v), and to the integrated diffusion coefficient by A (instead of $\mathcal{A}$).

DEFINITION 5.1. *A piecewise twice differentiable function* $u : \mathbb{R} \to [0, u_{\max}]$ *is a* steady-state entropy weak solution *of Model* 1 *if the following conditions are satisfied:* (a) *the function* $\gamma_1(x)A(u)'$ *is bounded, where* $' = d/dx$*;* (b) *the function* u *is a weak solution to the following ODE that arises from* (2.21) *and where* $g(x,u)$ *is given by* (2.23)*:*

$$g(x,u)' = \big(\gamma_1(x)A(u)'\big)'; \tag{5.1}$$

i.e., for every test function $\phi \in C_0^2(\mathbb{R})$ *with compact support we have*

$$\int_{\mathbb{R}} \Big(f\big(\boldsymbol{\gamma}(x), u(x)\big) - \gamma_1(x) A\big(u(x)\big)' \Big) \phi'(x)\, dx = 0; \tag{5.2}$$

and (c) *the following entropy inequality holds for all test functions* $\phi \in C_0^2(\mathbb{R})$, $\phi \geq 0$, *and* $k \in \mathbb{R}$*:*

$$\begin{aligned} &\int_{\mathbb{R}} \Big(\operatorname{sgn}\big(u(\xi)-k\big)\big(f(\boldsymbol{\gamma}(\xi),u(\xi)) - f(\boldsymbol{\gamma}(\xi),k)\big) - \gamma_1(\xi)A(u)' \Big)\phi'(\xi) d\xi \\ &+ \sum_{m\in\mathcal{J}} \big| f\big(\boldsymbol{\gamma}(m^+),k\big) - f\big(\boldsymbol{\gamma}(m^-),k\big)\big| \phi(m) \geq 0. \end{aligned} \tag{5.3}$$

It is standard to conclude from (5.2) that the following jump condition has to be satisfied across any discontinuity of the steady-state solution, where $u(x^+)$ and $u(x^-)$ refer to limits of $u(\xi)$ taken for $\xi \to x$ with $\xi > x$ and $\xi < x$, respectively:

$$f\big(\boldsymbol{\gamma}(x^-),u(x^-)\big) - \gamma_1(x^-)A'(u)|_{x=x^-} = f\big(\boldsymbol{\gamma}(x^+),u(x^+)\big) - \gamma_1(x^+)A'(u)|_{x=x^+}. \tag{5.4}$$

It is easy to see that this condition implies that steady-state solutions are constant for $x < x_{\mathrm{L}}$ and $x > x_{\mathrm{R}}$.

LEMMA 5.2. *Inequality* (5.3) *implies the following entropy jump condition:*

$$\begin{aligned} &\operatorname{sgn}\big(u(x^+)-k\big)\big[f\big(\boldsymbol{\gamma}(x^+),u(x^+)\big) - f\big(\boldsymbol{\gamma}(x^+),k\big) - \gamma_1(x^+)A'(u)|_{x=x^+}\big] \\ &- \operatorname{sgn}\big(u(x^-)-k\big)\big[f\big(\boldsymbol{\gamma}(x^-),u(x^-)\big) - f\big(\boldsymbol{\gamma}(x^-),k\big) - \gamma_1(x^-)A'(u)|_{x=x^-}\big] \\ &\leq \big| f\big(\boldsymbol{\gamma}(x^+),k\big) - f\big(\boldsymbol{\gamma}(x^-),k\big)\big| \quad \forall k \in \mathbb{R}. \end{aligned} \tag{5.5}$$

(Note that the right-hand side of (5.5) *is zero for* $x \notin \mathcal{J}$*.)*

Proof. The proof is a simpler variant of the proof of Lemma 2.6 of [56]. To outline it, let us fix $m \in \mathcal{J} = \{x_{\mathrm{L}}, 0, x_{\mathrm{R}}\}$, and define the function

$$\theta_\varepsilon(x) := \begin{cases} (\varepsilon + x)/\varepsilon & \text{if } x \in [-\varepsilon, 0], \\ (\varepsilon - x)/\varepsilon & \text{if } x \in [0, \varepsilon], \\ 0 & \text{otherwise} \end{cases}$$

with a parameter $\varepsilon > 0$. A density argument will reveal that we may choose the compactly supported Lipschitz continuous function $\theta_\varepsilon(x-m)$ as a test function in (5.3). This yields

$$\begin{aligned} &\frac{1}{\varepsilon}\int_{m-\varepsilon}^{m} \Big(F\big(\boldsymbol{\gamma}(\xi),u(\xi),k\big) - \gamma_1(x)\big|A(u)-A(k)\big|' \Big)\, dx \\ &- \frac{1}{\varepsilon}\int_{m}^{m+\varepsilon} \Big(F\big(\boldsymbol{\gamma}(\xi),u(\xi),k\big) - \gamma_1(x)\big|A(u)-A(k)\big|' \Big)\, dx \\ &\qquad + \big| f\big(\boldsymbol{\gamma}(m^+),k\big) - f\big(\boldsymbol{\gamma}(m^-),k\big)\big| \geq 0, \end{aligned} \tag{5.6}$$

where we recall the notation (4.12). Since the solution $u(x)$ is piecewise smooth, we may apply a time-independent version of Lemma 4.9 to conclude that for $\varepsilon \to 0$, we obtain from (5.6)

$$\begin{aligned} &F\big(\boldsymbol{\gamma}(m^+), u(m^+), k\big) - \mathrm{L}\lim_{\xi \downarrow m}\big(\gamma_1(\xi)\big|A(u) - A(k)\big|'\big) \\ &- F\big(\boldsymbol{\gamma}(m^-), u(m^-), k\big) + \mathrm{L}\lim_{\xi \uparrow m}\big(\gamma_1(\xi)\big|A(u) - A(k)\big|'\big) \\ &\le \big|f\big(\boldsymbol{\gamma}(m^+), k\big) - f\big(\boldsymbol{\gamma}(m^-), k\big)\big|. \end{aligned} \tag{5.7}$$

Assume for the moment that $u(m^+) \neq k$ and $u(m^-) \neq k$. Then Lemma 4.9 implies that

$$\begin{aligned} \mathrm{L}\lim_{\xi \downarrow m}\big(\gamma_1(\xi)\big|A(u) - A(k)\big|'\big) &= \operatorname{sgn}\big(u(m^+) - k\big)\gamma_1(m^+)A'(u)|_{x=m^+}, \\ \mathrm{L}\lim_{\xi \uparrow m}\big(\gamma_1(\xi)\big|A(u) - A(k)\big|'\big) &= \operatorname{sgn}\big(u(m^-) - k\big)\gamma_1(m^-)A'(u)|_{x=m^-}, \end{aligned}$$

such that (5.7) already implies (5.5). To remove this restriction, assume that $k = u(m^-)$. (The other case is similar.) Then the left-hand side of (5.5) is just

$$\begin{aligned} L := \operatorname{sgn}\big(u(m^+) - u(m^-)\big)\big[&f\big(\boldsymbol{\gamma}(m^+), u(m^+)\big) \\ &- f\big(\boldsymbol{\gamma}(m^+), u(m^-)\big) - \gamma_1(m^+)A'(u)\big|_{x=m^+}\big]. \end{aligned}$$

Using the jump condition (5.4), we obtain

$$\begin{aligned} L &= \operatorname{sgn}\big(u(m^+) - u(m^-)\big)\big[f\big(\boldsymbol{\gamma}(m^-), u(m^-)\big) \\ &\qquad - f\big(\boldsymbol{\gamma}(m^+), u(m^-)\big) - \gamma_1(m^-)A'(u)\big|_{x=m^-}\big] \\ &\le \big|f\big(\boldsymbol{\gamma}(m^-), u(m^-)\big) - f\big(\boldsymbol{\gamma}(m^+), u(m^-)\big)\big| \\ &\qquad - \operatorname{sgn}\big(u(m^+) - u(m^-)\big)\gamma_1(m^-)A'(u)\big|_{x=m^-}, \end{aligned}$$

and applying a steady-state variant of the right inequality of (4.27) in Lemma 4.10, we finally get

$$L \le \big|f\big(\boldsymbol{\gamma}(m^-), u(m^-)\big) - f\big(\boldsymbol{\gamma}(m^+), u(m^-)\big)\big|,$$

which is the inequality (5.5). Finally, since we are dealing with time-independent solutions, inequality (5.6) and the remaining discussion remain valid if we replace $m \in \mathcal{J}$ by $x \in \mathbb{R}$. □

The following lemma states a useful continuity result.

LEMMA 5.3. *Let $u(x)$ be a piecewise differentiable steady-state entropy weak solution of Model 1. Then $A(u(x^+)) = A(u(x^-))$ for all $x \in \mathbb{R}$.*

Proof. We consider first a point $x \in (x_\mathrm{L}, x_\mathrm{R})$, at which γ_1 is continuous. Then the boundedness of $\gamma_1(x)A(u)'$ implies that

$$0 = \lim_{\varepsilon \to 0}\int_{x-\varepsilon}^{x+\varepsilon} \gamma_1(\xi)A\big(u(\xi)\big)'d\xi = \lim_{\varepsilon \to 0}\int_{x-\varepsilon}^{x+\varepsilon} A\big(u(\xi)\big)'d\xi = A\big(u(x^+)\big) - A\big(u(x^-)\big).$$

Furthermore, consider that boundedness of $\gamma_1(x)A(u)'$ implies that $A(u)'$ is uniformly bounded on $[x_\mathrm{L}, x_\mathrm{R}]$. On the other hand, for $x < x_\mathrm{L}$ and $x > x_\mathrm{R}$, the jump condition (5.4) reduces to $q_\mathrm{L}u(x^-) = q_\mathrm{L}(x^+)$ and $q_\mathrm{R}u(x^+) = q_\mathrm{R}u(x^+)$, respectively, which

implies that $u(x^-) = u(x^+)$ and therefore no jumps are possible for $x < x_L$ and $x > x_R$. We conclude that piecewise smooth steady-state solutions are constant on $(-\infty, x_L)$ and (x_R, ∞), and therefore we have $A(u)' = 0$ on $(-\infty, x_L) \cup (x_R, \infty)$. Thus $A(u)'$ is uniformly bounded, and therefore has a continuous primitive $A(u)$, which implies that $A(u(x^+)) = A(u(x^-))$ for all $x \in \mathbb{R}$. ∎

REMARK 5.1. *We point out that the continuity property established by Lemma* 5.3 *includes the steady-state analogue of condition* (D.4) *stated for the time-dependent Model* 1. *However, we see that for our class of steady-state solutions, the continuity of* $A(u)$ *across* $x = x_L$ *and* $x = x_R$ *is a result of a more general regularity property and need not be postulated separately.*

Constructively, we select the discharge concentration $u_D = u(x)|_{x>x_R}$ and then integrate the ODE arising from the steady-state version of Model 1 upwards by obeying jump conditions wherever necessary. In doing so, we shall establish the limitations the entropy condition imposes on the choice of control parameters. Thus, the one-sided boundary condition is

$$u(x_R^-) = u_D > u_c. \tag{5.8}$$

The discussion will be limited to those cases where the compression zone does not reach the overflow level. In addition, to further simplify the discussion, we assume that the functions $g_L(u) := q_L u + b(u)$ and $g_R(u) := q_R u + b(u)$ are monotone on the interval $[0, u_c]$; i.e.,

$$q_L + b'(u) > 0, \quad q_R + b'(u) > 0 \quad \text{for } u \in [0, u_c]. \tag{5.9}$$

Moreover, we limit ourselves to steady-state solutions for which the overflow or effluent concentration $u_E := u(x)|_{x<x_L}$ is zero; that is, we choose the parameters u_D and u_F such that

$$Q_F u_F = (Q_R - Q_L)u_F = Q_R u_D - Q_L u_E \tag{5.10}$$

is satisfied with $u_E = 0$, or, equivalently, and since we consider Model 1 only,

$$u_F(q_R - q_L)/q_R = u_D. \tag{5.11}$$

These steady states represent either the conventional or the high-rate mode of continuous operation shown in Figure 1.1(a) and Figure 1.1(b), respectively.

At this point it should be emphasized that our steady-state problem is in general overdetermined. In fact, fixing u_D and integrating (5.1) upwards and obeying entropy and jump conditions, we will in general not achieve a solution with $u|_{x<x_L} = u_E = 0$. All profiles with $u|_{w<x_L} \neq u_E = 0$ have to be rejected as candidates for steady-state entropy solutions, since the global mass balance (5.10) is a consequence of the weak formulation (5.2). To make the analysis transparent, we will in some instances write out the symbol u_E in manipulations before setting it to zero. One result of this procedure is that under our model assumptions, no steady states with the compression region completely contained in the thickening zone but with a nonzero effluent concentration exist.

To determine a steady-state entropy weak solution that satisfies the global mass balance, it is in general necessary, say, to fix u_F, to choose u_D, to solve (5.1), to verify whether (5.10) is satisfied with u_E replaced by $u(x_L^-)$, and to iterate this solution procedure (for example, by varying u_D) until the global mass balance (5.10) is attained. However, under the simplifying assumption (5.9), it turns out that solutions

with $u_{\mathrm{E}} = 0$ can easily be characterized: these are those steady-state entropy weak solutions for which the compression region is strictly contained in the container, i.e., for which $\inf\{x \in \mathbb{R} : u(x) > u_{\mathrm{c}}\} > x_{\mathrm{L}}$. This is the most important subclass of steady states, since they are the most desired mode of operation (see Figure 1.1). Moreover, it turns out that these steady-state entropy weak solutions are strictly increasing.

5.1. Steady-state solution in the discharge zone. Before proceeding to integrate the ODE (5.1) upwards from $x = x_{\mathrm{R}}$, we consider the discharge zone $x > x_{\mathrm{R}}$. Since we are seeking solutions for which $A(v)$ is continuous, we conclude that $A(u(x_{\mathrm{R}}^+)) = A(u(x_{\mathrm{R}}^-)) = A(u_{\mathrm{D}})$, and therefore $u(x_{\mathrm{R}}^+) = u_{\mathrm{D}}$. On the other hand, from (5.1) we infer that the steady-state solution must be constant for $x > x_{\mathrm{R}}$. We conclude that $u(x) = u_{\mathrm{D}}$ for $x > x_{\mathrm{R}}$.

5.2. Steady-state solution in the thickening zone. Now that the steady-state solution has been determined in the interval (x_{R}, ∞), we determine the solution in the interval $(0, x_{\mathrm{R}})$. To this end, note first that as a consequence of the jump condition (5.4), the steady-state solution must satisfy the condition $q_{\mathrm{R}} u_{\mathrm{D}} + b(u_{\mathrm{D}}) - A(u)'|_{x=x_{\mathrm{R}}^+} = q_{\mathrm{R}} u_{\mathrm{D}}$, which means

$$b(u_{\mathrm{D}}) - A(u)'|_{x=x_{\mathrm{R}}^+} = 0. \tag{5.12}$$

Assume now that $v(x)$ is a continuously differentiable solution of the following one-sided boundary value problem, which is the subcase of (5.1) occurring for the interval $(0, x_{\mathrm{R}}]$:

$$q_{\mathrm{R}}(u - u_{\mathrm{D}}) + b(u) - A(u)' = 0 \quad \text{for } x < x_{\mathrm{R}}, \quad u(x_{\mathrm{R}}) = u_{\mathrm{D}}. \tag{5.13}$$

Note that we have used (5.12) to reduce the second-order ODE (5.1) to the first-order ODE (5.13). We consider the solution of (5.13) on the interval $[x_{\mathrm{c}}, x_{\mathrm{R}}]$, where

$$x_{\mathrm{c}} := \inf\big\{x \in (0, x_{\mathrm{R}}] \;\big|\; u(x) \text{ is the solution of (5.13) and } u(x) > u_{\mathrm{c}}\big\}. \tag{5.14}$$

However, not every solution of (5.13) is an acceptable steady-state solution. Rather, the following lemma shows that the entropy condition (5.3) imposes an additional admissibility condition. This condition imposes a restriction on the choice of q_{R} and u_{D} for a given flux density function $b(u)$.

LEMMA 5.4. *Any steady-state entropy solution $u(x)$ of the one-sided boundary value problem* (5.13) *on the interval $(x_{\mathrm{c}}, x_{\mathrm{R}})$ is monotonically increasing; i.e., $u'(x) \geq 0$ for $x \in [x_{\mathrm{c}}, x_{\mathrm{R}}]$. This statement is equivalent to the requirement*

$$q_{\mathrm{R}} u_{\mathrm{D}} \leq q_{\mathrm{R}} k + b(k) \quad \forall k \text{ between } u(x) \text{ and } u_{\mathrm{D}} \text{ for } x \in [x_{\mathrm{c}}, x_{\mathrm{R}}]. \tag{5.15}$$

Proof. In view of (5.12), we obtain from (5.5) the inequality

$$\begin{aligned} \forall k \in \mathbb{R} :\ \forall x \in (x_{\mathrm{c}}, x_{\mathrm{R}}) :\quad & \operatorname{sgn}(u_{\mathrm{D}} - k)\big(q_{\mathrm{R}}(u_{\mathrm{D}} - k) - b(k)\big) \\ & - \operatorname{sgn}\big(u(x) - k\big)\big[q_{\mathrm{R}}\big(u(x) - k\big) + b\big(u(x)\big) - b(k) - A(u)'\big] \leq 0. \end{aligned} \tag{5.16}$$

We now fix $x \in (x_{\mathrm{c}}, x_{\mathrm{R}})$ and evaluate (5.16) for different values of k. Setting $k < \min\{u(x), u_{\mathrm{D}}\}$ and $k > \max\{u(x), u_{\mathrm{D}}\}$, we obtain $\pm[q_{\mathrm{D}} u_{\mathrm{D}} - q_{\mathrm{R}} u - b(u) + A(u)'] \leq 0$, which in view of (5.13) is no new information. The choices $k = u(x)$ and $k = u_{\mathrm{D}}$ are covered as limiting cases in the subsequent discussion of the two alternatives in which k is located strictly between $u(x)$ and u_{D}.

Assume first that $u_{\mathrm{D}} < k < u(x)$. Then (5.16) leads to the inequality

$$2\big(q_{\mathrm{R}}k + b(k)\big) - q_{\mathrm{R}}u_{\mathrm{D}} - q_{\mathrm{R}}u(x) - b\big(u(x)\big) + A(u)' \le 0 \quad \forall k \in (u_{\mathrm{D}}, u(x)). \tag{5.17}$$

Using that $-q_{\mathrm{R}}u(x) - b(u(x)) + A(u)' = -q_{\mathrm{R}}u_{\mathrm{D}}$, we obtain from (5.17)

$$q_{\mathrm{R}}(k - u_{\mathrm{D}}) + b(k) \le 0 \quad \forall k \in (u_{\mathrm{D}}, u(x)). \tag{5.18}$$

However, (5.18) can never be satisfied, since $q_{\mathrm{R}} > 0$, $k > u_{\mathrm{D}}$, and we assume $b \ge 0$.

The remaining case is the assumption $u(x) < k < u_{\mathrm{D}}$, which leads to the inequality

$$-2\big(q_{\mathrm{R}}k + b(k)\big) + q_{\mathrm{R}}u_{\mathrm{D}} + q_{\mathrm{R}}u(x) + b\big(u(x)\big) - A(u)' \le 0 \quad \forall k \in (u(x), u_{\mathrm{D}}). \tag{5.19}$$

Using that $q_{\mathrm{R}}u(x) + b(u(x)) - A(u)' = q_{\mathrm{R}}u_{\mathrm{D}}$ and that $b(u_{\mathrm{D}}) > 0$, we obtain from (5.19)

$$q_{\mathrm{R}}u_{\mathrm{D}} \le q_{\mathrm{R}}k + b(k) \quad \forall k \in [u(x), u_{\mathrm{D}}]. \tag{5.20}$$

Since (5.13) can be rearranged to give

$$u'(x) = \frac{q_{\mathrm{R}}(u(x) - u_{\mathrm{D}}) + b(u(x))}{a(u(x))}, \tag{5.21}$$

we see that (5.20) implies that $u'(x) \ge 0$ for $x \in [x_{\mathrm{c}}, x_{\mathrm{R}}]$. □

REMARK 5.2. *Note that* (5.20) *has a useful graphical interpretation: namely, the graph of* $g_{\mathrm{R}}(u) = q_{\mathrm{R}}u + b(u)$ *must lie above the horizontal line* $f = q_{\mathrm{R}}u_{\mathrm{D}}$ *fixed by the desired operation data. This condition implies a limitation of the attainable solids throughput for the given material and vessel.*

To proceed with the discussion, we distinguish between three cases: $x_{\mathrm{c}} > 0$ (Case 1), $x_{\mathrm{c}} = 0$ and $u(0^+) > u_{\mathrm{c}}$ (Case 2), and $x_{\mathrm{c}} = 0$ and $u(0^+) = u_{\mathrm{c}}$ (Case 3).

Case 1 ($x_{\mathrm{c}} > 0$). The Rankine–Hugoniot and entropy jump conditions across $x = x_{\mathrm{c}}$ are

$$q_{\mathrm{R}}u(x_{\mathrm{c}}^-) + b\big(u(x_{\mathrm{c}}^-)\big) - A(u)'|_{x=x_{\mathrm{c}}^-} = q_{\mathrm{R}}u(x_{\mathrm{c}}^+) + b\big(u(x_{\mathrm{c}}^+)\big) - A(u)'|_{x=x_{\mathrm{c}}^+}, \tag{5.22}$$

$$\begin{aligned}\forall k \in \mathbb{R}: \quad & \operatorname{sgn}\big(u(x_{\mathrm{c}}^+) - k\big)\big[q_{\mathrm{R}}\big(u(x_{\mathrm{c}}^+) - k\big) + b\big(u(x_{\mathrm{c}}^+)\big) - b(k) - A(u)'|_{x=x_{\mathrm{c}}^+}\big] \\ & - \operatorname{sgn}\big(u(x_{\mathrm{c}}^-) - k\big)\big[q_{\mathrm{R}}\big(u(x_{\mathrm{c}}^-) - k\big) + b\big(u(x_{\mathrm{c}}^-)\big) - b(k) - A(u)'|_{x=x_{\mathrm{c}}^-}\big] \le 0,\end{aligned} \tag{5.23}$$

respectively. Moreover, from Lemma 5.3 it follows that $A(u(x_{\mathrm{c}}^-)) = A(u(x_{\mathrm{c}}^+)) = A(u_{\mathrm{c}}) = 0$, so that $0 \le u(x_{\mathrm{c}}^-) \le u_{\mathrm{c}}$. From (5.13) and the definition of x_{c} it follows that $q_{\mathrm{R}}u(x_{\mathrm{c}}^+) + b(u(x_{\mathrm{c}}^+)) - A(u)'|_{x=x_{\mathrm{c}}^+} = q_{\mathrm{R}}u_{\mathrm{D}}$. Inserting this into (5.22), we get

$$q_{\mathrm{R}}u(x_{\mathrm{c}}^-) + b\big(u(x_{\mathrm{c}}^-)\big) - A(u)'|_{x=x_{\mathrm{c}}^-} = q_{\mathrm{R}}u_{\mathrm{D}}. \tag{5.24}$$

Inserting (5.13), (5.24), and $u(x_{\mathrm{c}}^+) = u_{\mathrm{c}}$ into (5.23) yields

$$\begin{aligned}\forall k \in \mathbb{R}: \quad & \operatorname{sgn}(u_{\mathrm{c}} - k)\big(q_{\mathrm{R}}u_{\mathrm{D}} - q_{\mathrm{R}}k - b(k)\big) \\ & - \operatorname{sgn}\big(u(x_{\mathrm{c}}^-) - k\big)\big(q_{\mathrm{R}}u_{\mathrm{D}} - q_{\mathrm{R}}k - b(k)\big) \le 0.\end{aligned}$$

Obviously, the unique nontrivial case that needs to be discussed here is $u(x_{\mathrm{c}}^-) < k < u_{\mathrm{c}}$. Then we have $\operatorname{sgn}(u_{\mathrm{c}} - k) = 1$, $\operatorname{sgn}(u(x_{\mathrm{c}}^-) - k) = -1$, and the inequality is reduced to

$$q_{\mathrm{R}}u_{\mathrm{D}} \le g_{\mathrm{R}}(k) = q_{\mathrm{R}}k + b(k) \quad \forall k \in (u(x_{\mathrm{c}}^-), u_{\mathrm{c}}). \tag{5.25}$$

On the other hand, from (5.15) we infer that $q_{\mathrm{R}}u_{\mathrm{D}} \le q_{\mathrm{R}}u_{\mathrm{c}} + b(u_{\mathrm{c}})$. This means that at $u = u_{\mathrm{c}}$, the graph of $g_{\mathrm{R}}(u)$ lies above or intersects the horizontal line $f = q_{\mathrm{R}}u_{\mathrm{D}}$. Consequently, $u(x_{\mathrm{c}}^-)$ is the largest intersection of $g_{\mathrm{R}}(u)$ with the horizontal line $f = q_{\mathrm{R}}u_{\mathrm{D}}$ that is smaller than or equal to u_{c}:

$$u(x_{\mathrm{c}}^-) = \inf\big\{u \in [0, u_{\mathrm{c}}] \;\big|\; \forall \xi \in [u, u_{\mathrm{c}}] : g_{\mathrm{R}}(\xi) = q_{\mathrm{R}}\xi + b(\xi) \ge q_{\mathrm{R}}u_{\mathrm{D}}\big\}.$$

Since $g_{\mathrm{R}}(0) = 0 < q_{\mathrm{R}}u_{\mathrm{D}}$ and $g_{\mathrm{R}}(u_{\mathrm{c}}) > q_{\mathrm{R}}u_{\mathrm{D}}$ by assumption, it is ensured that the curve $u \mapsto g_{\mathrm{R}}(u)$ and the horizontal line $u \mapsto q_{\mathrm{R}}u_{\mathrm{D}}$ always intersect on $[0, u_{\mathrm{c}}]$, and thus $u(x_{\mathrm{c}}^-)$ is well defined. Note that for a function $b(u)$ with exactly one inflection point, the infimum is taken over at most three solutions of the equation $q_{\mathrm{R}}u + b(u) = q_{\mathrm{R}}u_{\mathrm{D}}$. It is not difficult to see that the steady-state solution in the interval $(x_{\mathrm{c}}, 0)$ is given by the constant $u(x_{\mathrm{c}}^-)$, which is uniquely constructed here.

It is at this point that assumption (5.9) turns out to be convenient in order to reduce the number of possible cases occurring in the continuation of the solution into the clarification zone. There would be no difficulty associated with relaxing this assumption.

Cases 2 *and* 3 ($x_{\mathrm{c}} = 0$, $v(0^+) \ge u_{\mathrm{c}}$). The construction of the steady-state solution in the thickening zone $(0, x_{\mathrm{R}}]$ is completed. The differentiable solution profile is given by the solution of the one-sided boundary value problem (5.13).

5.3. Steady-state solution in the clarification zone.

Case 1 ($x_{\mathrm{c}} > 0$). At $x = 0$, the next flux discontinuity has to be dealt with. However, since the solution for $x > 0$ is a constant not exceeding u_{c} and since $A(u)$ is continuous across $x = 0$, we have to treat a transition between two fluxes of a hyperbolic conservation law. The entropy weak solution to this problem has been determined in several papers [19, 39, 40, 41, 46, 47, 48, 58, 75]. The basic complication is that if the fluxes adjacent to $x = 0$ are nonmonotone, then there might be several possibilities to satisfy the Rankine–Hugoniot condition if $u(0^+)$ is given, and an entropy condition is necessary to single out the unique entropy-satisfying solution. This will in general generate a multitude of cases here, depending on the flux parameters and properties of the function b and on which solution of the equation $q_{\mathrm{R}}u + b(u) = q_{\mathrm{D}}u_{\mathrm{D}}$ yields the relevant state $u(0^+)$.

All these cases can be handled by the recent theory of conservation laws with discontinuous flux. However, assumption (5.9) helps to avoid this complication since it ensures that to a given value $u(0^+)$ there corresponds a *unique* value $u(0^-)$ such that the jump condition across $x = 0$,

$$q_{\mathrm{R}}(u(0^+) - u_{\mathrm{F}}) + b\big(u(0^+)\big) = q_{\mathrm{L}}(u(0^-) - u_{\mathrm{F}}) + b\big(u(0^-)\big), \tag{5.26}$$

is satisfied. To see this, recall that the constancy of u on $(0, x_{\mathrm{c}})$ and (5.24) imply that $q_{\mathrm{R}}u(0^+) + b(u(0^+)) = q_{\mathrm{R}}u_{\mathrm{D}}$. Inserting this into (5.26), we get

$$-(q_{\mathrm{R}} - q_{\mathrm{L}})u_{\mathrm{F}} + q_{\mathrm{R}}u_{\mathrm{D}} = q_{\mathrm{L}}u(0^-) + b\big(u(0^-)\big). \tag{5.27}$$

However, the left-hand side of (5.27) is just $q_{\mathrm{L}}u_{\mathrm{E}} \le 0$. On the other hand, due to (5.9) and since we seek a solution $0 \le u(0^-) \le u_{\mathrm{c}}$, the right-hand side of (5.27) is nonnegative and is zero only for $u(0^-) = 0$. Thus, the only solution is $u_{\mathrm{E}} = 0$, $u(0^-) = 0$, which implies that $u(x) = 0$ for all $x \le 0$.

Consequently, under the assumption (5.9) the only steady-state entropy weak solutions for which the sediment level $x = x_{\mathrm{c}}$ is located strictly below the feed level $x = 0$ are solutions for which there is only clear liquid in the clarification and overflow

zones ($x \geq 0$). For these solutions, the feed and discharge concentrations u_{F} and u_{D} are linked by $(q_{\mathrm{R}} - q_{\mathrm{L}})u_{\mathrm{F}} = q_{\mathrm{R}}u_{\mathrm{D}}$.

Case 2 ($x_{\mathrm{c}} = 0$, $u(0^+) > u_{\mathrm{c}}$). Lemma 5.3 implies that $u(0^-) = u(0^+)$ if $u^+ > u_{\mathrm{c}}$. Thus, we can continue to solve (5.1) in the clarification zone $x \in (x_{\mathrm{L}}, 0)$. Integrating this ODE over the interval $(x, 0)$, we obtain the following one-sided boundary value problem for a first-order ODE:

(5.28)
$$q_{\mathrm{L}}\big(u(x) - u(0^-)\big) + b(u) - b\big(u(0^-)\big) - A(u)'|_{x=0^-} = 0 \quad \text{for } x < 0, \quad u(0) = u(0^-).$$

To determine $u'(0^-)$, we use the Rankine–Hugoniot condition (5.4) across $x = 0$,

$$\begin{aligned} &q_{\mathrm{L}}\big(u(0^-) - u_{\mathrm{F}}\big) + b\big(u(0^-)\big) - A(u)'|_{x=0^-} \\ &= q_{\mathrm{R}}\big(u(0^+) - u_{\mathrm{F}}\big) + b\big(u(0^+)\big) - A(u)'|_{x=0^+}. \end{aligned} \tag{5.29}$$

Recalling that we already know that $u(0^-) = u(0^+)$, we get

$$A(u)'|_{x=0^-} = (q_{\mathrm{L}} - q_{\mathrm{R}})\big(u(0^-) - u_{\mathrm{F}}\big) + A(u)'|_{x=0^+}. \tag{5.30}$$

On the other hand, in the present case we know that $A(u)'|_{x=0^+} = q_{\mathrm{R}}(u(0^+) - u_{\mathrm{D}}) + b(u(0^+))$. Inserting this into (5.30) and replacing $u(0^+)$ by $u(0^-)$, we get

$$A(u)'|_{x=0^-} = q_{\mathrm{L}}u(0^-) - q_{\mathrm{R}}u_{\mathrm{D}} - (q_{\mathrm{L}} - q_{\mathrm{R}})u_{\mathrm{F}} + b\big(u(0^-)\big). \tag{5.31}$$

Finally, we insert (5.31) into (5.28) and obtain the one-sided boundary value problem

$$q_{\mathrm{L}}u(x) + b(u) - A(u)' - q_{\mathrm{R}}u_{\mathrm{D}} - (q_{\mathrm{L}} - q_{\mathrm{R}})u_{\mathrm{F}} = 0, \quad x < 0; \quad v(0) = v(0^-). \tag{5.32}$$

We now define

$$\tilde{x}_{\mathrm{c}} := \inf\big\{x \in [x_{\mathrm{L}}, 0) \;\big|\; u(x) \text{ is the solution of (5.28) and } u_{\max} \geq u(x) > u_{\mathrm{c}}\big\} \tag{5.33}$$

and recall from (5.32) and (5.31) that we have

$$b\big(u(x)\big) - A(u)' = -q_{\mathrm{L}}u(x) + q_{\mathrm{R}}u_{\mathrm{D}} + (q_{\mathrm{L}} - q_{\mathrm{R}})u_{\mathrm{F}} \quad \text{for } x \in (\tilde{x}_{\mathrm{c}}, 0],$$

as well as that we obtain from (5.5) the inequality

$$\begin{aligned} &\big[\operatorname{sgn}\big(u(0^-) - k\big) - \operatorname{sgn}\big(u(x) - k\big)\big]\big(-q_{\mathrm{L}}k - b(k) + q_{\mathrm{R}}u_{\mathrm{D}} + (q_{\mathrm{L}} - q_{\mathrm{R}})u_{\mathrm{F}}\big) \leq 0 \\ &\forall x \in (\tilde{x}_{\mathrm{c}}, 0) \text{ and } \forall k \in \mathbb{R}. \end{aligned} \tag{5.34}$$

We observe that $q_{\mathrm{R}}u_{\mathrm{D}} + (q_{\mathrm{L}} - q_{\mathrm{R}})u_{\mathrm{F}} = q_{\mathrm{L}}u_{\mathrm{E}}$. Then, the solution in the interval $(\tilde{x}_{\mathrm{c}}, 0)$ is given by the solution of the one-sided boundary problem (which is a slight rearrangement of (5.32))

$$u'(x) = \frac{q_{\mathrm{L}}u(x) + b(u(x)) - q_{\mathrm{L}}u_{\mathrm{E}}}{a(u(x))}, \quad x < 0, \quad u(0) = u(0^-). \tag{5.35}$$

Due to our assumption (5.9), since $g_{\mathrm{L}}(0) = 0$, and $b(u)$ and therefore $g_{\mathrm{L}}(u)$ has exactly one inflection point, we know that $g_{\mathrm{L}}(u)$ has exactly one positive maximum $u_{\mathrm{M}} > u_{\mathrm{c}}$ and that $g_{\mathrm{L}}(u)$ is monotonically decreasing between u_{M} and $u_{\max} = 1$ with $g_{\mathrm{L}}(u_{\max}) = q_{\mathrm{L}}u_{\max} < 0$. Consequently, there exists exactly one point u^* with

$u_c < u^* < u_{\max}$ such that $q_L u^* + b(u^*) = q_L u_E$. Since the maximum of g_L is positive but $q_L u_E \le 0$, we know that $u^* > u_M$, and therefore

$$g_L(u) < g_L(u^*) \quad \text{for } u > u^*,\ u \le u_{\max}. \tag{5.36}$$

We first assume that $u(0^+) = u(0^-) > u^*$. Our immediate goal is to show that $u(0^-) > u^*$ does not lead to an admissible steady-state solution. Note that by the discussion of the solution in the thickening zone, we know that $u(0^-) < u_D$.

By the definition of u^*, we may rewrite the ODE (5.35) as

$$u'(x) = \frac{g_L(u(x)) - g_L(u^*)}{a(u(x))}, \quad x < 0, \quad u(0) = u(0^-).$$

In light of (5.36), we see that inserting $u(0^-) > u^*$ will cause the right-hand side of the ODE in (5.35) to be negative at $x = 0$, and this right-hand side remains negative if we proceed with the integration of (5.35), since we produce a solution that is monotonically decreasing (increasing upwards). This integration may be continued upwards until either $u = u_{\max} = 1$ is attained or $x = x_L$ is reached. In the first case, however, there is no valid way to continue the solution to the remaining interval $(x_L, \tilde{x}_c)$ other than setting $u = u_{\max} = 1$ for $x \in (x_L, \tilde{x}_c)$. This means that at $x = x_L$, the jump condition implies that $q_L u_{\max} = q u_E$, that is, $u_E = u_{\max}$, in contradiction with the assumption $u_E < u_{\max}$. In the other case, in which $x = x_L$ is reached by integrating (5.35), we have the following Rankine–Hugoniot condition across $x = x_L$:

$$q_L u(x_L^+) + b\big(u(x_L^+)\big) - A(u)'\big|_{w=x_L^+} = q_L u(x_L^-). \tag{5.37}$$

On the other hand, since $u(x_L^+) \ge u(0^-) > u_c$, we have $A(u(x_L^+)) > 0$, and thus, due to Lemma 5.3, $u(x_L^+) = u(x_L^-)$, and therefore (5.37) reduces to $b(u(x_L^+)) = A(u)'|_{x=x_L^+}$. However, since by assumption $u'(x_L^+) < 0$, we have that $A'(u)|_{x=x_L^+} < 0$, which in turn implies that $b(u(x_L^+)) < 0$. This is a contradiction to the nonnegativity of b. Thus, no admissible steady-state solution can be constructed if $u(0^-) > u^*$.

The case $u(0^-) = u^*$ equally leads to an inadmissible solution, since integrating (5.35) leads to the constant solution $u \equiv u^*$ on $(x_L, 0)$. Similarly to the discussion of the previous case, the jump condition (5.37) and Lemma 5.3 imply that $b(u^*) = A(u)'|_{x=x_L^+}$. However, the constancy of u along $x \in (x_L, 0)$ implies that $A(u)'|_{x=x_L^+} = 0$, and therefore $b(u^*) = 0$, in contradiction with the assumed properties of b. Another reason to reject the profiles with $u(0^-) \ge u^*$ as candidates for steady-state entropy weak solutions is the violation of the global conservation principle (5.10), since we have chosen u_D and u_F such that $u_E = 0$, but in these cases our integration yields positive values of $u(x_L^-)$, which should equal u_E.

We now look at the remaining case $u(0^-) < u^*$. Then the right-hand side of the ODE in (5.35) is always positive, which implies a monotonically increasing (decreasing upwards) solution $u(x)$ until $x = \tilde{x}_c$ is reached. This solution also satisfies the entropy condition. In fact, for any $x \in (\tilde{x}_c, 0)$ with $u(x) < u(0^-)$ and for all $k \in (u(x), u(0^-))$, condition (5.34) (which is void for all other values of k and for $u(x) = u(0^-)$) implies that $2(-q_L k - b(k) + q_L u_E) \le 0$, i.e.,

$$q_L k + b(k) - q_L u_E \ge 0 \quad \forall k \in (u(x), u(0^-)), \tag{5.38}$$

which in view of (5.35) is satisfied if $u(x)$ is a monotonically increasing solution on $(\tilde{x}_c, 0)$.

We summarize our discussion of Case 2 in the clarification zone by the following lemma.

LEMMA 5.5. *Any admissible steady-state entropy solution $u = u(x)$ of Model 1 with $u(0^-) = u(0^+) > u_c$ must satisfy $u(0^-) < u^*$, where u^* is the unique point in (u_c, u_{max}) satisfying $g_L(u^*) \equiv q_L u^* + b(u^*) = q_L u_E$. This solution is monotonically increasing on the interval $(\tilde{x}_c, 0)$, where $\tilde{x}_c$ is defined by (5.33).*

REMARK 5.3. *The statement of Lemma 5.5 also has an obvious graphical interpretation. However, this condition requires knowledge of the value $u(0^+) = u(0^-)$. Thus, it can be evaluated only after the solution in the thickening zone has been determined. Furthermore, combining this finding with Lemma 5.4 for the thickening zone, we see that in any of the Cases 1, 2, or 3, the entropy condition and jump conditions enforce that $u'(x) \geq 0$ in the compression region.*

With the present discussion, we have constructed a steady-state solution up to $\tilde{x}_c$, provided that $u(x) > u_c$ in the thickening zone $x \in (0, x_R)$. To finish the steady-state construction, let us first recall that for sake of brevity and being well aware of the incompleteness of the treatment in the present paper, we limit the discussion to those steady states for which $\tilde{x}_c > x_L$. In this case, there is a jump located at $x = \tilde{x}_c$. We now seek the constant solution value $u = u(\tilde{x}_c^-)$ in the interval $(x_L, \tilde{x}_c)$. This value must satisfy $0 \leq u(\tilde{x}_c^-) \leq u_c$. From the Rankine–Hugoniot condition that follows from (5.4), $q_L u(\tilde{x}_c^-) + b(\tilde{x}_c^-) = q_L u_c + b(u_c) - A(u)'|_{x=\tilde{x}_c^+}$, we see that the constant $u(\tilde{x}_c^-) = u_c$ is not a solution. Consequently, we look for a constant $0 \leq u(\tilde{x}_c^-) < u_c$. To this end, note that the steady-state jump condition at $x = x_L$ is $g_L(u(x_L^+)) = q_L u(x_L^+) + b(u(x_L^+)) = q_L u(x_L^-) = q_L u_E$. Taking into account that $g_L(u)$ is a nonnegative monotonically increasing function on $[0, u_c]$, while the right-hand side is a nonpositive constant, we conclude (similar as in the discussion of Case 1) that $u_E = 0$ and $u(\tilde{x}_c^-) = 0$; i.e., the solution is zero on $(x_L, \tilde{x}_c)$.

REMARK 5.4. *The last result means that the mathematical model correctly describes the elimination of the hindered settling region in steady-state operation when the sediment level (where $u = u_c$) is allowed to rise above the feed level, as drawn in Figure 1.1(b). No particles are elutriated from the compression region into the overflow. This supports the physical explanation that above the feed level, the sediment bed acts as a filter medium for the portion of the feed flow that is directed into the clarification zone.*

If we added a small amount of hydrodynamic diffusion and used a strictly positive diffusion coefficient $a(u)$ such that the resulting model were strictly parabolic, then there would be no upper limit for the integration of ODEs like (5.28), and under the preconditions of Case 2 discussed here, the quantity $\tilde{x}_c$ defined by (5.33) would always assume the value x_L (corresponding to the overflow level) with $u(x_L^+) > 0$. In other words, there would be a small but positive volume fraction of solids in the overflow. Steady-state concentration profiles for a clarifier-thickener model with a strictly positive diffusion coefficient that illustrate this situation are plotted, for example, in [81].

Finally, we mention that Lev, Rubin, and Sheintuch [64] *use a steady-state clarifier-thickener model without compression effects but with a hydrodynamic diffusion term instead. A discussion of possible concentration extrema at steady state leads them to the conclusion that the concentration must increase downwards. Our analysis shows that by applying the entropy concept to the steady-state ODE, this characterization remains valid even when hydrodynamic diffusion vanishes.*

Case 3 ($x_c = 0$, $u(0^+) = u_c$). In this case, we have $A(u(0^+)) = 0$ and due to the continuity of $A(u)$ across $x = 0$, $A(u(0^-)) = 0$, which means $u(0^-) \in [0, u_c]$. Since, in the present case, it has been possible to integrate (5.13) up to $x = 0^+$, we can

replace the Rankine–Hugoniot condition across $x = 0$, (5.29), by $q_{\mathrm{L}}u(0^-)+b(u(0^-))-A(u)'|_{x=0^-} = (q_{\mathrm{L}} - q_{\mathrm{R}})u_{\mathrm{F}} + q_{\mathrm{R}}u_{\mathrm{D}}$. On the other hand, the following entropy jump condition follows by evaluating (5.5) for $x = 0$:

$$\begin{aligned} &\operatorname{sgn}\big(u(0^+) - k\big)\big[q_{\mathrm{R}}\big(u(0^+) - k\big) + b\big(u(0^+)\big) - b(k) - A(u)'|_{x=0^+}\big] \\ &- \operatorname{sgn}\big(u(0^-) - k\big)\big[q_{\mathrm{L}}\big(u(0^-) - k\big) + b\big(u(0^-)\big) - b(k) - A(u)'|_{x=0^-}\big] \\ &\le \big|q_{\mathrm{R}}(k - u_{\mathrm{F}}) - q_{\mathrm{L}}(k - u_{\mathrm{F}})\big| \quad \forall k \in \mathbb{R}. \end{aligned} \tag{5.39}$$

Inserting (5.29), using once again (5.13) and that $u(0^+) = u_{\mathrm{c}}$, we get

$$\begin{aligned} &\operatorname{sgn}\big(u_{\mathrm{c}} - k\big)\big(-q_{\mathrm{R}}k - b(k) + q_{\mathrm{R}}u_{\mathrm{D}}\big) \\ &- \operatorname{sgn}\big(u(0^-) - k\big)\big(-q_{\mathrm{L}}k - b(k) + (q_{\mathrm{L}} - q_{\mathrm{R}})u_{\mathrm{F}} + q_{\mathrm{R}}u_{\mathrm{D}}\big) \\ &\le \big|(q_{\mathrm{R}} - q_{\mathrm{L}})(k - u_{\mathrm{F}})\big| \quad \forall k \in \mathbb{R}. \end{aligned}$$

Choosing $k = 0$ and exploiting that $b(k) = 0$, we get

$$\operatorname{sgn}(u_{\mathrm{c}})q_{\mathrm{R}}u_{\mathrm{D}} - \operatorname{sgn}\big(u(0^-)\big)\big((q_{\mathrm{L}} - q_{\mathrm{R}})u_{\mathrm{F}} + q_{\mathrm{R}}u_{\mathrm{D}}\big) \le \big|(q_{\mathrm{R}} - q_{\mathrm{L}})(-u_{\mathrm{F}})\big|, \tag{5.40}$$

which in view of $u_{\mathrm{c}} > 0$, $u_{\mathrm{F}} > 0$, and $q_{\mathrm{R}} - q_{\mathrm{L}} \ge 0$ implies that

$$q_{\mathrm{R}}u_{\mathrm{D}} + \operatorname{sgn}\big(u(0^-)\big)(q_{\mathrm{R}} - q_{\mathrm{L}})u_{\mathrm{F}} + \operatorname{sgn}\big(v(0^-)\big)q_{\mathrm{R}}u_{\mathrm{D}} \le (q_{\mathrm{R}} - q_{\mathrm{L}})u_{\mathrm{F}}. \tag{5.41}$$

If $\operatorname{sgn}(u(0^-)) = 1$, then the left-hand side of (5.41) equals $2q_{\mathrm{R}}u_{\mathrm{D}} + (q_{\mathrm{R}} - q_{\mathrm{L}})u_{\mathrm{F}}$. Thus, inequality (5.41) cannot be satisfied. The unique remaining option is $\operatorname{sgn}(u(0^-)) = 0$, i.e., $u(0^-) = 0$. It is then easily seen that the solution for $x < 0$, including also the section $x < x_{\mathrm{L}}$, vanishes identically. Thus, the solution of Case 3 is the limiting case of Cases 1 and 2 for $u(0^+) \to u_{\mathrm{c}}$.

REMARK 5.5. *The condition $u'(x) \ge 0$ is in full agreement with engineering intuition, since one expects that in a clarifier-thickener operating properly at steady state, the concentration increases downwards. In fact, in several previous papers dealing with a simpler model of an ideal continuous thickener* [13, 16]*, which basically consists only of the thickening zone of the model discussed here, the condition $u'(x) \ge 0$ was* postulated *as a separate* requirement *for the determination of admissible steady states following just from this intuition, and the graphical condition* (5.20) *was derived by using this assumption in* (5.21)*. We now clearly see that the natural requirement that a steady state should be an entropy weak solution* implies *this monotonicity property in the thickening zone, and it is therefore* unnecessary *to introduce it as an additional condition.*

Observe that in contrast to our analysis of the thickening zone, we do not apply the entropy condition to construct the restrictions on the parameters (expressed by Lemma 5.5*) in the clarification and overflow zones; rather, we exploit the jump conditions to establish these restrictions and then check that the admissible solution satisfies the entropy condition.*

5.4. Examples of steady states. Here and in the numerical examples, the flocculated suspension is characterized by the functions $b(u)$ and $\sigma_{\mathrm{e}}(u)$ given by (2.6) and (2.9), respectively, with $v_\infty = 10^{-4}\,\mathrm{m/s}$, $C = 5$, $\sigma_0 = 1.0\,\mathrm{Pa}$, $u_{\mathrm{c}} = 0.1$, and $k = 6$. The remaining parameters are $\Delta\varrho = 1500\,\mathrm{kg/m^3}$ and $g = 9.81\,\mathrm{m/s^2}$. These values are fairly realistic and are also used in [14, 16].

Moreover, we assume the bulk velocities $q_{\mathrm{R}} = 2.5 \times 10^{-6}\,\mathrm{m/s}$ and $q_{\mathrm{L}} = -1.0 \times 10^{-5}\,\mathrm{m/s}$. Thus, we are interested in steady states for which $u_{\mathrm{D}} = u_{\mathrm{F}}(q_{\mathrm{R}} - q_{\mathrm{L}})/q_{\mathrm{R}} =$

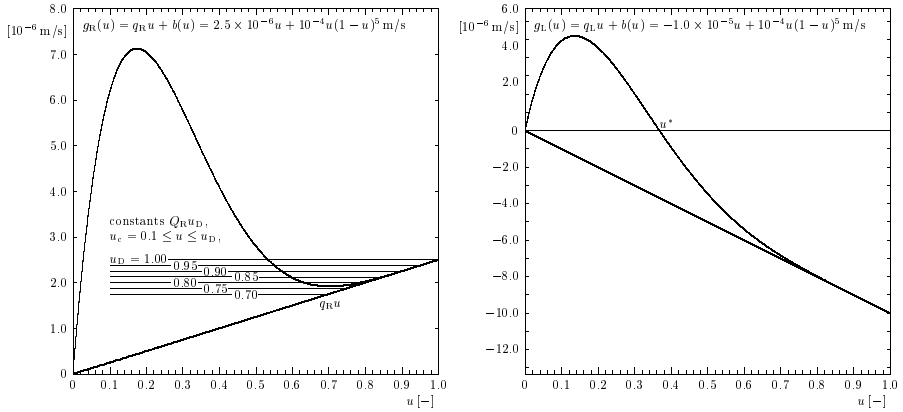


FIG. 5.1. *The flux functions $q_R u + b(u)$ (left) and $q_L + b(u)$ (right). The left diagram also shows the constant lines $q_R u_D$ for some values of u_D.*

$5u_F$, and these parameters have been chosen so that assumption (5.9) is satisfied. The relevant flux functions for the thickening and clarification zone, $g_R(u)$ and $g_L(u)$, are plotted in Figure 5.1.

We start the steady-state construction by fixing values of u_D and determining the corresponding value x_c. We limit ourselves to those values u_D that ensure that the entropy condition (5.15) is satisfied. To this end, we consider the plot of $g_R(u)$ and draw horizontal line segments $f = q_R u_D$ for a selection of values of u_D and for $u_c \le u \le u_D$, as has been done in the left plot of Figure 5.1. We see that these lines remain strictly below the graph of $g_R(u)$ for those values of u_D for which

$$q_R u_D < \min_{u_c \le u \le 1} g_R(u) = g_R(0.703) \approx 1.92 \times 10^{-6}\,\mathrm{m/s}.$$

This implies that the entropy condition (5.15) is satisfied a priori (i.e., independently of the depth of the thickening zone x_R) for all u_D with

$$u_c \le u_D < u_{D\max} := \frac{1}{q_R} \min_{u_c \le u \le 1} g_R(u) = \frac{1.92 \times 10^{-6}\,\mathrm{m/s}}{2.5 \times 10^{-6}\,\mathrm{m/s}} = 0.768.$$

For all other values of u_D, it would be necessary to determine a solution to (5.13) and to check whether this is monotone on $[x_c, x_R]$. We will not pursue this here.

Given this limitation on u_D, we choose the profiles for $u_D = 0.3, 0.35, 0.4, 0.405, 0.41, \ldots, 0.455$ for closer inspection. Solving (5.13) with a standard numerical ODE method, we obtain that for $u_D \le 0.41$, we have $x_c > 0$ and therefore steady states of Case 1, while all other values lead to candidates for Case 2. Solving the equation $q_R u(x_c^-) + b(u(x_c^-)) = q_R u_D$ numerically yields the following values of $u(x_c^-)$, which are the constant values each entropy weak solution assumes on $[0, x_c]$:

u_D	0.3	0.35	0.4	0.405	0.41
$u(x_c^-)$	0.00759	0.00892	0.01026	0.01039	0.01053

For these values of u_D, the steady-state entropy weak solution in the clarification and overflow zones is zero. Figure 5.2 includes plots of these profiles.

It remains to deal with $u_D = 0.415, 0.42, \ldots, 0.455$, the candidates for Case 2, for which the clarification zone has to be examined. We have just found out that

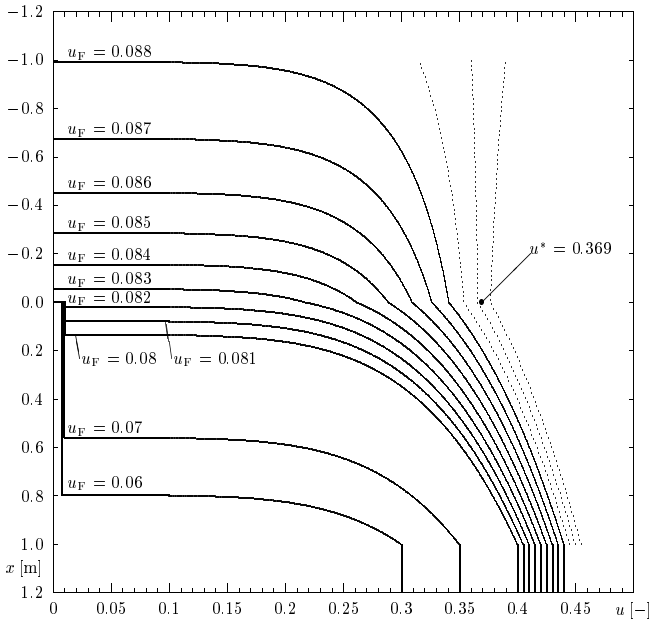


FIG. 5.2. *Steady-state concentration profiles in Vessel* 1. *The dotted curves show solutions of* (5.13) *and* (5.28) *that do not lead to admissible steady states with zero overflow concentration.*

all of these concentration values admit an entropy-satisfying steady-state solution in the thickening zone. However, with the present parameters we have $u_E = 0$, which implies that $u^* = 0.369$. This point is marked in Figures 5.1 and 5.2. We see that integrating (5.13) from $u_D = 0.455$ leads to a profile with $u(0^+) > u^*$, which does not lead to an admissible solution in the clarification zone. The tentative profile for this case is the rightmost dotted profile in Figure 5.2. For $u_D = 0.445$ and $u_D = 0.45$, we obtain admissible profiles in the clarification zone, which, however, reach the overflow level $x = x_L$ and will produce an effluent with $u_E > 0$. These profiles are no admissible entropy steady-state solutions since the global conservation principle is violated. The values $u_D = 0.415, 0.42, \ldots, 0.435$ lead to admissible steady-state profiles with $\tilde{x}_c > x_L$, and, as a consequence of our analysis, $u_E = 0$.

6. Numerical examples.

6.1. Preliminary remarks. Note that for $k \in \mathbb{N}$, as chosen here, standard calculus yields that the function $A(u)$ has the explicit representation $A(u) = \mathsf{A}(u) - \mathsf{A}(u_c)$ for $u > u_c$ with

$$\mathsf{A}(u) := -\frac{v_\infty \sigma_0}{\Delta \varrho g u_c^k}(1-u)^C u^k \sum_{j=1}^{k} c_j \left(\frac{1}{u} - 1\right)^j, \quad c_j = \prod_{l=1}^{j} \frac{k+1-l}{C+l}, \quad j = 1, \ldots, k;$$

it is straightforward to verify by differentiating $\mathsf{A}(u)$ that

$$\frac{d\mathsf{A}(u)}{du} = v_\infty u(1-u)^C \cdot \frac{1}{\Delta \varrho g u} \cdot \frac{d}{du}\big(\sigma_0\big((u/u_c)^k - 1\big)\big) = \frac{v_\infty \sigma_0}{\Delta \varrho g u_c^k}(1-u)^C k u^{k-1},$$

so that the function $A(u)$ defined here indeed satisfies (2.10).

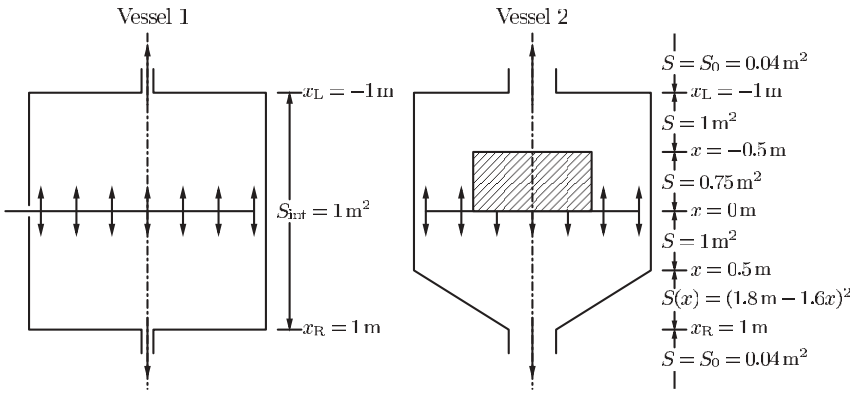


FIG. 6.1. *The cylindrical clarifier-thickener (left) and the unit with discontinuously varying cross-sectional area (right) used for numerical simulations.*

We consider two units, a cylindrical one (Vessel 1) and one with discontinuously varying cross-sectional area (Vessel 2); see Figure 6.1. The interior of Vessel 1 is $S_{\text{int}} = 1\,\text{m}^2$. (Recall that the outer pipe diameter S_0 can always be transformed away.) The piecewise constant cross-sectional area function $S(x)$ of the noncylindrical one, Vessel 2, is defined in Figure 6.1.

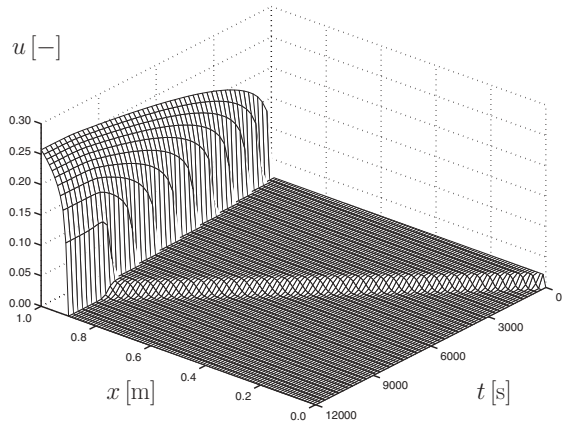
The motivation of the choice of $S(x)$ for Vessel 2 is that most clarifier-thickeners have a conically shaped bottom to facilitate transport of material to the discharge outlet and that we assume that one-quarter of the cross-sectional area in the clarification zone is occupied by installations related to the feed mechanisms. Observe that in the second case, both parameters γ_1 and γ_2 accounting for the cross-sectional area and the bulk flow, respectively, have a discontinuity at $x = 0$.

6.2. Example 1: Batch settling. To illustrate the material behavior of the suspension, we present in Figure 6.2 three simulations of the settling of an initially homogeneous suspension at initial concentrations $u_0 = 0.02$, 0.08, and 0.2 in a closed column (for which all Q's and q's vanish) of height $L = 1\,\text{m}$. We employ the explicit numerical method (3.1), $\Delta x = L/500$, and $\lambda = 20\,\text{s/m}$. In the first two cases, we have $a(u_0) = 0$, and the suspension-clear liquid interface propagates as a sharp shock and the transition between the region of initial concentration and the sediment rising from below is sharp, while in the third case transitions are continuous. In all three cases, a stationary sediment is forming. In Figure 6.2 and all subsequent three-dimensional plots, the visual grid used to represent the solution is much coarser than the computational.

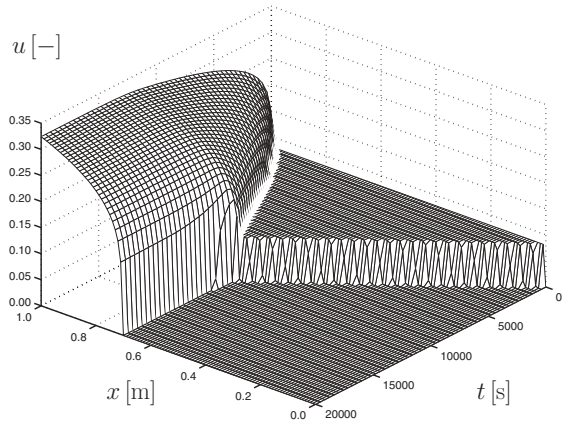
6.3. Numerical simulations of Model 1.

6.3.1. Example 2: Variation of discharge and overflow rates. The four simulations shown in Figure 6.3 have been computed using the unique feed flux $q_F u_F = (q_R - q_L)u_F$ with $q_F = 1.25 \times 10^{-5}\,\text{m/s}$ and $u_F = 0.08$ but by using four different "splits" of the feed rate into the discharge and overflow rates attained by varying the parameter $\nu \in [0, 1]$ in $q_R = \nu q_F$ and $q_L = -(1-\nu)q_F$. In all four simulations, solving the transient equations for sufficiently large times apparently leads to a stationary solution. The numerical scheme is the explicit one (3.1) with $\lambda = 40\,\text{s/m}$, and for this and all other numerical simulations of Model 1, we choose $\Delta x = 1/300\,\text{m}$.

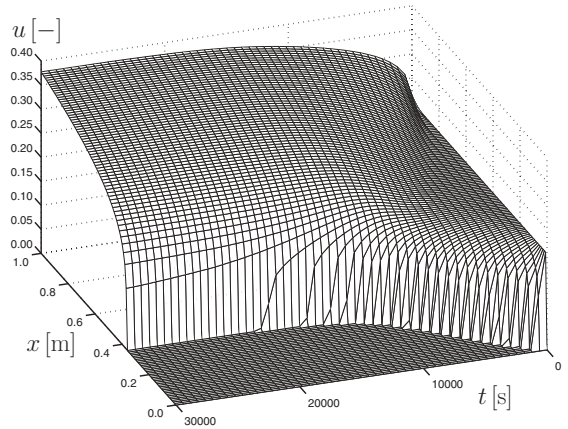
In Figure 6.3(a), we set $\nu = 1$; i.e., the vessel is closed at the top and opened



(a)



(b)



(c)

FIG. 6.2. *Example* 1: *Simulation of batch settling of an initially homogeneous suspension with* $u_0 = 0.02$ (a), $u_0 = 0.08$ (b), *and* $u_0 = 0.20$ (c).

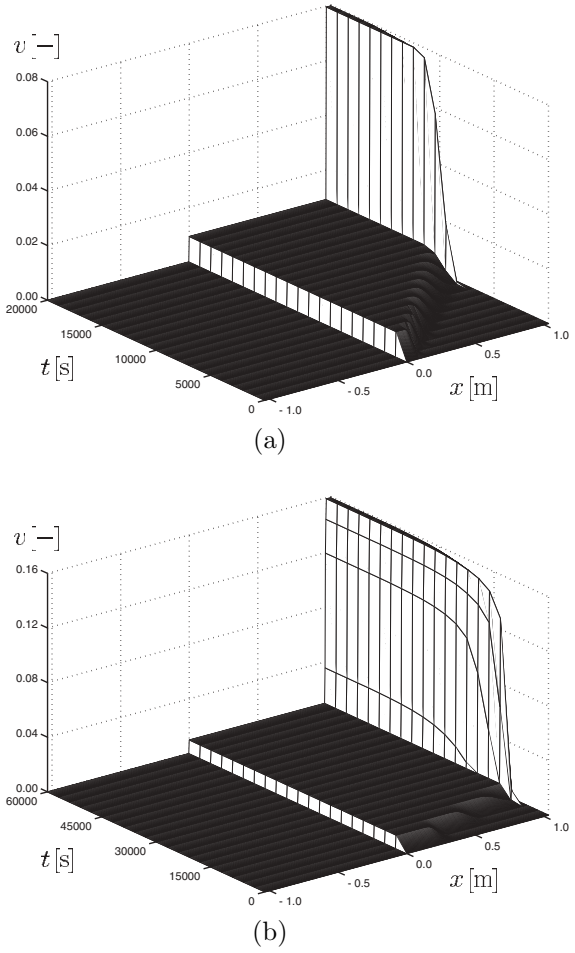
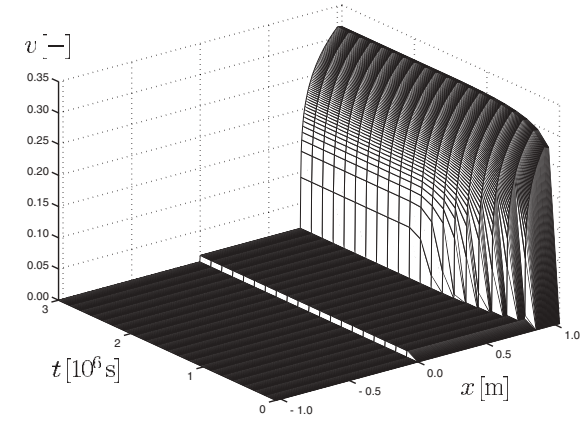


FIG. 6.3. *Example* 2: *Simulation of filling up a cylindrical clarifier-thickener with* $u_{\mathrm{F}} = 0.08$, $q_{\mathrm{F}} = 1.25 \times 10^{-5}\mathrm{m/s}$, $q_{\mathrm{R}} = \nu q_{\mathrm{F}}$, *and* $q_{\mathrm{L}} = -(1-\nu)q_{\mathrm{F}}$ *for* $\nu = 1$ (a) *and* $\nu = 0.5$ (b).

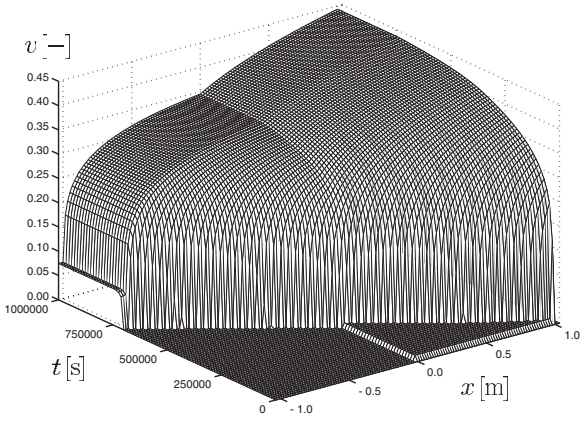
at the bottom, with the volume feed rate equaling the discharge rate. We see that the feed suspension is immediately diluted upon entering the vessel but attains its original concentration, 0.08, again when passing the discharge level. No compression region occurs.

For $\nu = 0.5$ (Figure 6.3(b)), we obtain a very thin sediment layer at the bottom, and the discharge concentration is 0.16, twice the feed concentration. The solution qualitatively agrees with that for $\nu = 0.25$ (Figure 6.3(c)). However, for $\nu = 0.25$ the final discharge concentration is $0.32 = 4u_{\mathrm{F}}$, and the sediment layer is appreciable. The stationary solutions attained in these cases correspond to steady-state solutions of Case 1 (conventional operation).

Finally, we take $\nu = 0$; i.e., the vessel is closed at its bottom. The corresponding solution is shown in Figure 6.3(d). We observe that the feed suspension is at first immediately diluted upon entering the thickening zone. The material forms a compressible sediment layer at the bottom. This layer rises at nearly constant speed, breaks into the clarification zone, and, finally, produces an overflow at constant



(c)



(d)

FIG. 6.3 (CONT'D.). *Example* 2: *Simulation of filling up a cylindrical clarifier-thickener with* $u_{\mathrm{F}} = 0.08$, $q_{\mathrm{F}} = 1.25 \times 10^{-5}\mathrm{m/s}$, $q_{\mathrm{R}} = \nu q_{\mathrm{F}}$, *and* $q_{\mathrm{L}} = -(1-\nu)q_{\mathrm{F}}$ *for* $\nu = 0.25$ (c) *and* $\nu = 0$ (d).

concentration 0.08, which is just the feed concentration. (Note that this kind of steady state is not included in the analysis of section 5.)

6.3.2. Example 3: Transitions between approximate steady states. We now utilize the examples of section 5.4 to design a long-time numerical example in which the parameters for the time-dependent Model 1 are chosen in such a way that the predetermined steady states may be attained. This example and Example 4 (for Model 2) are solved by the semi-implicit method (3.4) with $\lambda = 4000\,\mathrm{s/m}$. We consider the constant flow velocities $q_{\mathrm{R}} = 2.5 \times 10^{-6}\,\mathrm{m/s}$ and $q_{\mathrm{L}} = -1.0 \times 10^{-5}\,\mathrm{m/s}$. The feed concentration u_{F} is varied in a stepwise fashion as follows:

$$u_{\mathrm{F}}(t) = \begin{cases} 0.086 & \text{for } 0 \le t \le t_1 := 4.0 \times 10^7\,\mathrm{s}, \\ 0.08 & \text{for } t_1 < t \le t_2 := 6.0 \times 10^7\,\mathrm{s}, \\ 0.088 & \text{for } t_2 < t \le t_3 := 9.5 \times 10^7\,\mathrm{s}, \\ 0 & \text{for } t > t_3. \end{cases} \tag{6.1}$$

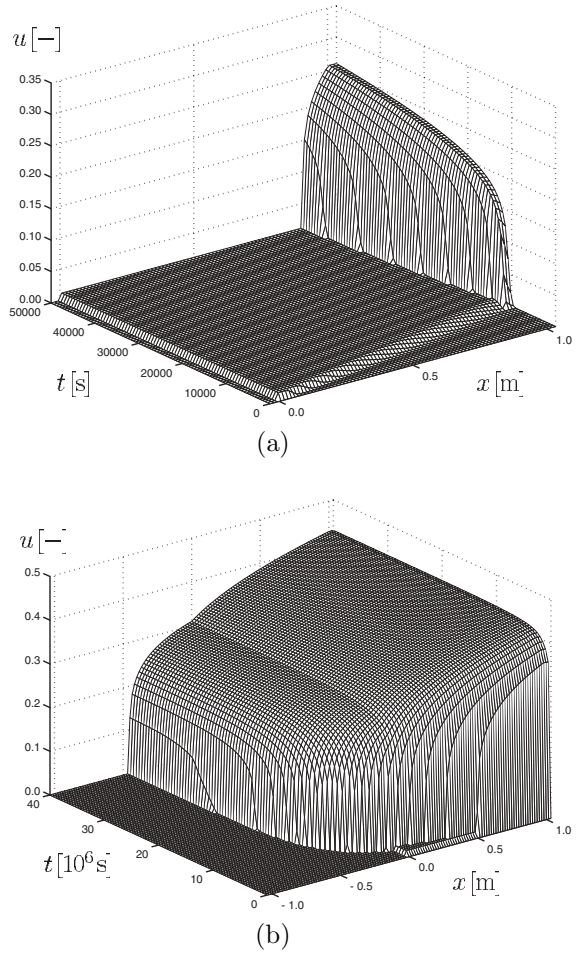
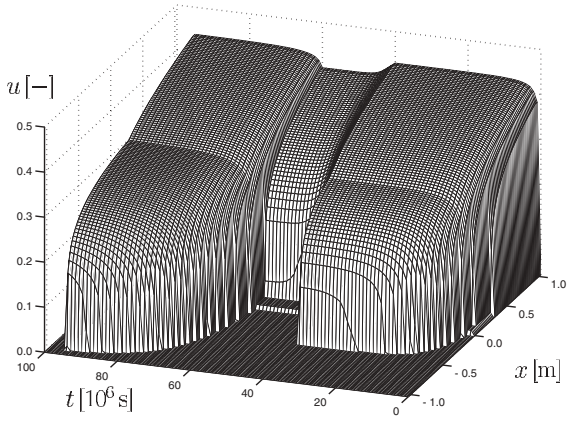


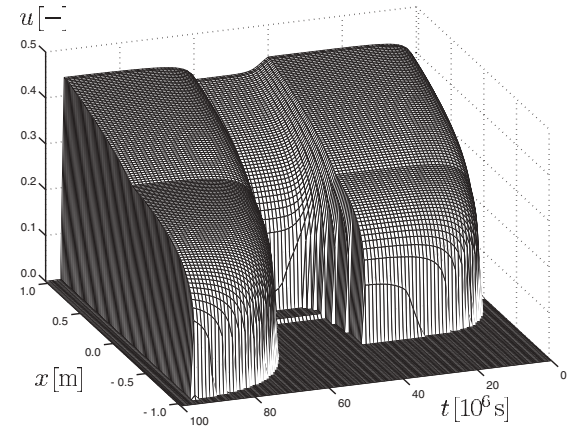
FIG. 6.4. *Example* 3: *Simulations of the fill-up and transitions between steady states in a cylindrical clarifier-thickener (Vessel* 1*): filling up a cylindrical clarifier-thickener up to steady state with* $u_{\mathrm{F}} = 0.086$ *(initial stages:* (a)*, complete process:* (b)*).*

The initial stage of the fill-up process is shown in Figure 6.4(a), while the complete solution for the first time interval $[0, t_1]$ is plotted in Figure 6.4(b). We observe that the feed propagates as a rarefaction wave into the thickening zone and that a sediment layer is built up, which rises above the feed level. The interesting point is that the numerical solution becomes stationary after the very long simulated time of about 3.0×10^7 s, which corresponds to roughly one year, and the stationary solution closely approximates the steady-state profile corresponding to the same values of q_{L}, q_{R}, and u_{F} plotted in Figure 5.2. In particular, the numerical value of the overflow concentration remains zero, and the solution value assumed at $w = x_{\mathrm{R}}$ equals 0.42997. Thus, we have reason to believe that this steady-state solution is indeed the limit attained by the entropy weak solution for these parameters, at least for $t \to \infty$; whether the steady state is reached even in finite time would be a further question.

At the simulated time $t = t_1$, we reduce u_{F} to a value that in combination with those of q_{L} and q_{R} once again corresponds to a steady state plotted in Figure 5.2 but



(c)



(d)

FIG. 6.4 (CONT'D.). *Example* 3: *Simulations of the fill-up and transitions between steady states in a cylindrical clarifier-thickener (Vessel* 1*): two different views of the complete simulation with successive changes of* u_{F} *from* 0.086 *to* 0.08, 0.088, *and* 0 *(*(c) *and* (d)*).*

this time to one of conventional operation. Figures 6.4(c) and (d) indicate that also this steady state seems to be attained by the transient solution. In particular, the hindered settling region becomes visible again. Shortly before $t = t_2$, the numerical solution value at $x = x_{\mathrm{R}}$ equals 0.40001.

At $t = t_2$, we increase u_{F} to 0.088, and we observe that the simulation converges again to the corresponding steady state of Figure 5.2. Shortly before the solution becomes stationary, at $t = t_3$, we switch off the feed by setting $u_{\mathrm{F}} = 0$, and the clarifier-thickener unit empties rapidly.

Although the operations involved in this run—filling up, transitions between steady states, and emptying of a clarifier-thickener—are typical control actions, practitioners would, of course, accelerate the fill-up by closing the unit [15, 16]. The main intention behind our example is, however, to illustrate that the model apparently converges to steady-state solutions.

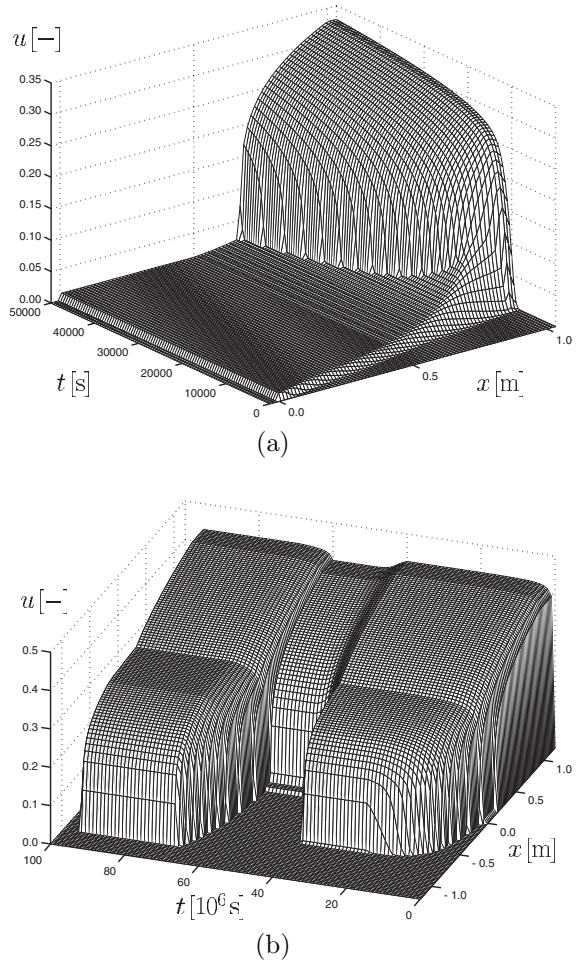


FIG. 6.5. *Example* 4: *Simulation of transitions between steady states in Vessel* 2*: plot of the initial fill-up stage* (a) *and long-time simulation* (b).

6.4. Example 4: Numerical simulation of Model 2. Finally, we repeat the simulation of Example 3; that is, we use again the function $u_{\mathrm{F}}(t)$ defined by (6.1), but now we consider Model 2, Vessel 2 drawn in Figure 6.1, and select $Q_{\mathrm{L}} = -1.0 \times 10^{-5}\mathrm{m}^3/\mathrm{s}$ and $Q_{\mathrm{R}} = 2.5 \times 10^{-6}\mathrm{m}^3/\mathrm{s}$. Model 2 enforces a CFL condition involving a factor $\max S(x)/\min S(x)$, which equals 25 here, so we have to decrease accuracy to maintain acceptable computation time. We here choose $\lambda = 100\,\mathrm{s/m}$ and $\Delta x = 1/50\,\mathrm{m}$ for the long-time run shown in Figure 6.5(b) but $\Delta x = 1/300\,\mathrm{m}$ for the short simulated period in Figure 6.5(a).

We observe that the concentration profiles slightly reflects the thickener geometry, although due to the diffusion term, this effect is less pronounced than for the same model without compression (i.e., for $A \equiv 0$); see [23, 24]. First, observe the difference between Figure 6.5(a) and the corresponding simulation for Model 1 and Vessel 1 in Figure 6.4(a). The conical cross-section of the lower part of the thickening zone causes a continuous variation of the concentration in the initial hindered settling region and

TABLE 6.1
Approximate L^1 errors for the numerical solution of Example 3.

	$t = 200000\,\mathrm{s}$		$t = 1000000\,\mathrm{s}$		$t = 2000000\,\mathrm{s}$	
$J = \frac{1\,\mathrm{m}}{\Delta x}$	L^1 error	conv. rate	L^1 error	conv. rate	L^1 error	conv. rate
10	1.429e−2		2.301e−2		2.857e−2	
20	7.092e−3	1.010	1.115e−2	1.004	1.424e−2	1.005
40	3.475e−3	1.029	5.666e−3	1.017	7.026e−3	1.019
80	1.691e−3	1.039	2.758e−3	1.038	3.425e−3	1.037
100	1.338e−3	1.048	2.176e−3	1.063	2.704e−3	1.059
160	8.004e−4	1.094	1.306e−3	1.087	1.623e−3	1.085
200	6.225e−4	1.126	1.015e−3	1.129	1.263e−3	1.124

accelerates the fill-up process.

Of course, a variable cross-sectional area $S(x)$ complicates the discussion of steady states but also offers new design opportunities [16]. Here, Vessel 2 no longer admits a steady state for $u_\mathrm{F} = 0.088$ with $u_\mathrm{E} = 0$. In fact, we observe in Figure 6.5(b) that a stationary profile is attained with nonzero overflow concentration. Rather, the numerical values attained are overflow and underflow concentrations $u(x_\mathrm{L}^-) = 0.00121$ and $u(x_\mathrm{R}^+) = 0.43293$. The stationary profile is probably a steady state with nonzero effluent concentration (which is not included in the analysis of section 5, as are not any other steady states for Model 2).

6.5. Comments on the numerical results. First of all, we mention that our numerical results, including test runs with coarser discretizations (not shown here), suggest that the scheme indeed converges to solutions for which $A(u)$ is continuous across x_L and x_R, as required by condition (D.4). However, it should be emphasized that our scheme does not possess a built-in mechanism to enforce this property. As emphasized before, a rigorous proof for the convergence of the scheme towards a solution satisfying (D.4) is still an open problem, and it might be that one even has to modify the scheme to ensure this property. This requires a deeper numerical analysis, which we defer to another paper.

Furthermore, the accuracy and convergence rate of the numerical scheme used herein may be of interest. To this end, we measured approximate L^1 errors for the simulation of Example 3 by measuring the difference $\|u^\Delta(\cdot,t) - u^\mathrm{ref}(\cdot,t)\|_{L^1(-1.1\,\mathrm{m},1.1\,\mathrm{m})}$ for a number of discretizations $(\Delta x, \Delta t = \lambda \Delta x)$ at $t = 200000\,\mathrm{s}$, $t = 1000000\,\mathrm{s}$, and $t = 2000000\,\mathrm{s}$, where u^Δ is the numerical solution obtained with $\Delta x = 1\,\mathrm{m}/J$, $J = 10, 20, 40, 80, 100, 160, 200$, and u^ref is a high-accuracy reference solution with $J = 1600$; see Table 6.1. In all cases, the semi-implicit scheme (3.4) with the parameter $\lambda = 4000\,\mathrm{s/m}$ was used.

We observe that the approximate L^1 convergence rates are slightly larger than but close to 1. This is consistent with the formal first-order accuracy of the time discretization and of the discretization of the convective fluxes. Similar approximately linear convergence has been observed for the explicit version (3.1) of the scheme applied to a slightly simpler equation that does not involve a discontinuous parameter in the diffusion term in [55], and for the application of the explicit scheme to the initial-boundary value problem of batch settling of a flocculated suspension in [18], which does not involve a discontinuous parameter at all. It should be pointed out that observed convergence rates substantially depend on the parameters and numerical examples chosen. For example, similar approximate L^1 tables for the explicit scheme (3.1) applied to the first-order clarifier-thickener model (obtained by setting $A \equiv 0$) are presented in [25]. It turns out that when approximate L^1 errors are measured at

times when the solution includes nonstationary "hyperbolic" discontinuities (shocks), then the observed L^1 convergence rate measured on a succession of grids may fall substantially below 1.

To put these observations into the proper perspective, let us mention that a theoretical estimate of the rate of convergence of the numerical scheme presented herein is outside current theory, even in the case of smooth coefficients. However, our prime motivation behind advancing the scheme (3.1) (and its semi-implicit variant (3.4)) was to use it as a constructive tool for the well-posedness analysis and to employ it for simulations to illustrate the mathematical analysis. Clearly, we do not propose (3.1) or (3.4) as the optimal scheme for simulations in practice. For that purpose the scheme should be upgraded to formal second order both in time and in space accuracy, which can be attained, for example, by combining flux correction and Strang-type operator splitting between the hyperbolic and parabolic portions of the problem. We pursue this further in [26].

Finally, there are conceivable alternative schemes for the clarifier-thickener model that seem worth exploring. For example, one could combine the very efficient front tracking method introduced in [19] for the hyperbolic portion and combine it with finite differencing of the diffusion term in an operator splitting procedure. Alternatively, the relaxation scheme used in [21] for the simulation of the first-order clarifier-thickener model could be extended to handle the second-order degenerate diffusion term accounting for sediment compressibility (as in [29]). Recent numerical schemes for strongly degenerate parabolic equations that can possibly be extended to the clarifier-thickener model also include the local discontinuous Galerkin method [33] and diffusive kinetic BGK approximations [3, 11].

Acknowledgment. We are grateful to the referees for valuable comments that resulted in a number of improvements in this paper.

REFERENCES

[1] Adimurthi, J. Jaffré, and G. D. Veerappa Gowda, *Godunov-type methods for conservation laws with a flux function discontinuous in space*, SIAM J. Numer. Anal., 42 (2004), pp. 179–208.

[2] D. Amadori, L. Gosse, and G. Guerrac, *Godunov-type approximation for a general resonant balance law with large data*, J. Differential Equations, 198 (2004), pp. 233–274.

[3] D. Aregba-Driollet, R. Natalini, and S. Tang, *Explicit diffusive kinetic schemes for nonlinear degenerate parabolic systems*, Math. Comp., 73 (2004), pp. 63–94.

[4] A. A. A. Aziz, R. G. de Kretser, D. R. Dixon, and P. J. Scales, *The characterisation of slurry dewatering*, Wat. Sci. Tech., 41 (2000), pp. 9–16.

[5] P. Baiti and H. K. Jenssen, *Well-posedness for a class of* 2×2 *conservation laws with* L^∞ *data*, J. Differential Equations, 140 (1997), pp. 161–185.

[6] N. G. Barton, C.-H. Li, and S. J. Spencer, *Control of a surface of discontinuity in continuous thickeners*, J. Austral. Math. Soc. Ser. B, 33 (1992), pp. 269–289.

[7] M. Bendahmane and K. H. Karlsen, *Renormalized entropy solutions for quasi-linear anisotropic degenerate parabolic equations*, SIAM J. Math. Anal., 36 (2004), pp. 405–422.

[8] P. Bénilan and H. Touré, *Sur l'équation générale* $u_t = a(\cdot, u, \phi(\cdot, u)_x)_x + v$ *dans* L_1. II. *Le problème d'évolution*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 12 (1995), pp. 727–761.

[9] S. Berres, R. Bürger, and K. H. Karlsen, *Central schemes and systems of conservation laws with discontinuous coefficients modeling gravity separation of polydisperse suspensions*, J. Comput. Appl. Math., 164–165 (2004), pp. 53–80.

[10] S. Berres, R. Bürger, K. H. Karlsen, and E. M. Tory, *Strongly degenerate parabolic-hyperbolic systems modeling polydisperse sedimentation with compression*, SIAM J. Appl. Math., 64 (2003), pp. 41–80.

[11] F. Bouchut, F. R. Guarguaglini, and R. Natalini, *Diffusive BGK approximations for nonlinear multidimensional parabolic equations*, Indiana Univ. Math. J., 49 (2000), pp. 723–749.

[12] F. Bouchut and F. James, *One-dimensional transport equations with discontinuous coefficients*, Nonlinear Anal., 32 (1998), pp. 891–933.

[13] R. Bürger, M. C. Bustos, and F. Concha, *Settling velocities of particulate systems:* 9. *Phenomenological theory of sedimentation processes: Numerical simulation of the transient behaviour of flocculated suspensions in an ideal batch or continuous thickener*, Int. J. Mineral Process., 55 (1999), pp. 267–282.

[14] R. Bürger and F. Concha, *Settling velocities of particulate systems:* 12. *Batch centrifugation of flocculated suspensions*, Int. J. Mineral Process., 63 (2001), pp. 115–145.

[15] R. Bürger, F. Concha, and F. M. Tiller, *Applications of the phenomenological theory to several published experimental cases of sedimentation processes*, Chem. Eng. J., 80 (2000), pp. 105–117.

[16] R. Bürger, J. J. R. Damasceno, and K. H. Karlsen, *A mathematical model for batch and continuous thickening in vessels with varying cross section*, Int. J. Mineral Process., 73 (2004), pp. 183–208.

[17] R. Bürger, S. Evje, K. H. Karlsen, and K.-A. Lie, *Numerical methods for the simulation of the settling of flocculated suspensions*, Chem. Eng. J., 80 (2000), pp. 91–104.

[18] R. Bürger and K. H. Karlsen, *On some upwind schemes for the phenomenological sedimentation-consolidation model*, J. Engrg. Math., 41 (2001), pp. 145–166.

[19] R. Bürger, K. H. Karlsen, C. Klingenberg, and N. H. Risebro, *A front tracking approach to a model of continuous sedimentation in ideal clarifier-thickener units*, Nonlinear Anal. Real World Appl., 4 (2003), pp. 457–481.

[20] R. Bürger, K. H. Karlsen, S. Mishra, and J. D. Towers, *On conservation laws with discontinuous flux*, in Trends in Applications of Mathematics to Mechanics, Y. Wang and K. Hutter, eds., Shaker Verlag, Aachen, 2005, pp. 75–84.

[21] R. Bürger, K. H. Karlsen, and N. H. Risebro, *A relaxation scheme for continuous sedimentation in ideal clarifier-thickener units*, Comput. Math. Appl., to appear.

[22] R. Bürger, K. H. Karlsen, N. H. Risebro, and J. D. Towers, *Numerical methods for the simulation of continuous sedimentation in ideal clarifier-thickener units*, Int. J. Mineral Process., 73 (2004), pp. 209–228.

[23] R. Bürger, K. H. Karlsen, N. H. Risebro, and J. D. Towers, *Monotone difference approximations for the simulation of clarifier-thickener units*, Comput. Vis. Sci., 6 (2004), pp. 83–91.

[24] R. Bürger, K. H. Karlsen, N. H. Risebro, and J. D. Towers, *On a model for continuous sedimentation in vessels with discontinuously varying cross-sectional area*, in Hyperbolic Problems: Theory, Numerics, Applications, T. Y. Hou and E. Tadmor, eds., Springer-Verlag, Berlin, 2003, pp. 397–406.

[25] R. Bürger, K. H. Karlsen, N. H. Risebro, and J. D. Towers, *Well-posedness in BV_t and convergence of a difference scheme for continuous sedimentation in ideal clarifier-thickener units*, Numer. Math., 97 (2004), pp. 25–65.

[26] R. Bürger, K. H. Karlsen, and J. D. Towers, *High Resolution Schemes for Continuous Sedimentation in Ideal Clarifier-Thickener Units*, manuscript.

[27] R. Bürger, W. L. Wendland, and F. Concha, *Model equations for gravitational sedimentation-consolidation processes*, ZAMM Z. Angew. Math. Mech., 80 (2000), pp. 79–92.

[28] J. Carrillo, *Entropy solutions for nonlinear degenerate problems*, Arch. Ration. Mech. Anal., 147 (1999), pp. 269–361.

[29] F. Cavalli, G. Naldi, and G. Toscani, *Relaxation Methods for the Sedimentation of Polydisperse Suspensions of Spheres*, preprint, Milano, Italy, 2002.

[30] J.-Ph. Chancelier, M. Cohen de Lara, and F. Pacard, *Analysis of a conservation PDE with discontinuous flux: A model of settler*, SIAM J. Appl. Math., 54 (1994), pp. 954–995.

[31] G.-Q. Chen and E. DiBenedetto, *Stability of entropy solutions to the Cauchy problem for a class of nonlinear hyperbolic-parabolic equations*, SIAM J. Math. Anal., 33 (2001), pp. 751–762.

[32] G.-Q. Chen and B. Perthame, *Well-posedness for non-isotropic degenerate parabolic-hyperbolic equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 645–668.

[33] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion problems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[34] F. Concha, A. Barrientos, and M. C. Bustos, *Phenomenological model of high capacity thickening*, in Proceedings of the 19th International Mineral Processing Congress (XIX IMPC), San Francisco, CA, 1995, pp. 75–79.

[35] R. Courant, K. Friedrichs, and H. Lewy, *Über die partiellen Differenzengleichungen der mathematischen Physik*, Math. Annalen, 100 (1928), pp. 28–74.

[36] M. G. Crandall and A. Majda, *Monotone difference approximations for scalar conservation laws*, Math. Comp., 34 (1980), pp. 1–21.

[37] M. G. Crandall and L. Tartar, *Some relations between nonexpansive and order preserving mappings,* Proc. Amer. Math. Soc., 78 (1980), pp. 385–390.
[38] R. G. De Kretser, S. P. Usher, P. J. Scales, D. V. Boger, and K. A. Landman, *Rapid filtration measurement of dewatering design and optimization parameters*, AIChE J., 47 (2001), pp. 1758–1769.
[39] S. Diehl, *On scalar conservation laws with point source and discontinuous flux function*, SIAM J. Math. Anal., 26 (1995), pp. 1425–1451.
[40] S. Diehl, *A conservation law with point source and discontinuous flux function modelling continuous sedimentation*, SIAM J. Appl. Math., 56 (1996), pp. 388–419.
[41] S. Diehl, *Dynamic and steady-state behavior of continuous sedimentation*, SIAM J. Appl. Math., 57 (1997), pp. 991–1018.
[42] S. Diehl, *On boundary conditions and solutions for ideal clarifier-thickener units*, Chem. Eng. J., 80 (2000), pp. 119–133.
[43] S. Diehl, *Operating charts for continuous sedimentation* I: *Control of steady states*, J. Engrg. Math., 41 (2001), pp. 117–144.
[44] B. Engquist and S. Osher, *One-sided difference approximations for nonlinear conservation laws*, Math. Comp., 36 (1981), pp. 321–351.
[45] P. Garrido, R. Bürger, and F. Concha, *Settling velocities of particulate systems:* 11. *Comparison of the phenomenological sedimentation-consolidation model with published experimental results*, Int. J. Mineral Process., 60 (2000), pp. 213–227.
[46] T. Gimse, *Conservation laws with discontinuous flux functions*, SIAM J. Math. Anal., 24 (1993), pp. 279–289.
[47] T. Gimse and N. H. Risebro, *Riemann problems with a discontinuous flux function*, in Proceedings of the Third International Conference on Hyperbolic Problems, Uppsala, Sweden, 1990, pp. 488–502.
[48] T. Gimse and N. H. Risebro, *Solution of the Cauchy problem for a conservation law with a discontinuous flux function*, SIAM J. Math. Anal., 23 (1992), pp. 635–648.
[49] J. M.-K. Hong, *Part* I: *An Extension of the Riemann Problems and Glimm's Method to General Systems of Conservation Laws with Source Terms. Part* II: *A Total Variation Bound on the Conserved Quantities for a Generic Resonant Nonlinear Balance Laws*, Ph.D. thesis, University of California, Davis, 2000.
[50] J. Jaffré, *Numerical calculation of the flux across an interface between two rock types of a porous medium for a two-phase flow*, in Hyperbolic Problems: Theory, Numerics, Applications, World Scientific, River Edge, NJ, 1996, pp. 165–177.
[51] E. F. Kaasschieter, *Solving the Buckley-Leverett equation with gravity in a heterogeneous porous medium*, Comput. Geosci., 3 (1999), pp. 23–48.
[52] K. H. Karlsen, C. Klingenberg, and N. H. Risebro, *A relaxation scheme for conservation laws with discontinuous coefficients*, Math. Comp., 73 (2004), pp. 1235–1259.
[53] K. H. Karlsen and N. H. Risebro, *On the uniqueness and stability of entropy solutions of nonlinear degenerate parabolic equations with rough coefficients*, Discrete Contin. Dyn. Syst., 9 (2003), pp. 1081–1104.
[54] K. H. Karlsen, N. H. Risebro, and J. D. Towers, *On a nonlinear degenerate parabolic transport-diffusion equation with a discontinuous coefficient*, Electron. J. Differential Equations, 93 (2002), pp. 1–23.
[55] K. H. Karlsen, N. H. Risebro, and J. D. Towers, *On an upwind difference scheme for degenerate parabolic convection-diffusion equations with a discontinuous coefficient*, IMA J. Numer. Anal., 22 (2002), pp. 623–664.
[56] K. H. Karlsen, N. H. Risebro, and J. D. Towers, *L^1 stability for entropy solutions of nonlinear degenerate parabolic convection-diffusion equations with discontinuous coefficients*, Skr. K. Nor. Vid. Selsk., 3 (2003), pp. 1–49.
[57] K. H. Karlsen and J. D. Towers, *Convergence of the Lax-Friedrichs scheme and stability for conservation laws with a discontinuous space-time dependent flux*, Chinese Ann. Math. Ser. B, 25 (2004), pp. 287–318.
[58] R. A. Klausen and N. H. Risebro, *Stability of conservation laws with discontinuous coefficients*, J. Differential Equations, 157 (1999), pp. 41–60.
[59] C. Klingenberg and N. H. Risebro, *Convex conservation laws with discontinuous coefficients, existence, uniqueness and asymptotic behavior*, Comm. Partial Differential Equations, 20 (1995), pp. 1959–1990.
[60] C. Klingenberg and N. H. Risebro, *Stability of a resonant system of conservation laws modeling polymer flow with gravitation*, J. Differential Equations, 170 (2001), pp. 344–380.
[61] S. N. Kružkov, *First order quasi-linear equations in several independent variables*, Math. USSR-Sb., 10 (1970), pp. 217–243.
[62] G. J. Kynch, *A theory of sedimentation*, Trans. Farad. Soc., 48 (1952), pp. 166–176.

[63] D. R. Lester, *Colloidal Suspension Dewatering Analysis*, Ph.D. thesis, Department of Chemical Engineering, University of Melbourne, Australia, 2002.

[64] O. Lev, E. Rubin, and M. Sheintuch, *Steady state analysis of a continuous clarifier-thickener system*, AIChE J., 32 (1986), pp. 1516–1525.

[65] L. Lin, J. B. Temple, and J. Wang, *A comparison of convergence rates for Godunov's method and Glimm's method in resonant nonlinear systems of conservation laws*, SIAM J. Numer. Anal., 32 (1995), pp. 824–840.

[66] L. Lin, B. Temple, and W. Jinghua, *Suppression of oscillations in Godunov's method for a resonant non-strictly hyperbolic system*, SIAM J. Numer. Anal., 32 (1995), pp. 841–864.

[67] W. K. Lyons, *Conservation laws with sharp inhomogeneities*, Quart. Appl. Math., 40 (1982/83), pp. 385–393.

[68] C. Mascia, A. Porretta, and A. Terracina, *Nonhomogeneous Dirichlet problems for degenerate parabolic-hyperbolic equations*, Arch. Ration. Mech. Anal., 163 (2002), pp. 87–124.

[69] A. Michel and J. Vovelle, *Entropy formulation for parabolic degenerate equations with general Dirichlet boundary conditions and application to the convergence of FV methods*, SIAM J. Numer. Anal., 41 (2003), pp. 2262–2293.

[70] S. Mishra, *Convergence of Upwind Finite Difference Schemes for a Scalar Conservation Law with Indefinite Discontinuities in the Flux Function*, preprint, 2002.

[71] J. Molenaar, *Entropy conditions for heterogeneity induced shocks in two-phase flow problems*, in Mathematical Modelling of Flow through Porous Media, A. P. Bourgeat, C. Carasso, S. Luckhaus, and A. Mikelic, eds., World Scientific, Singapore, 1995.

[72] D. N. Ostrov, *Viscosity solutions and convergence of monotone schemes for synthetic aperture radar shape-from-shading equations with discontinuous intensities*, SIAM J. Appl. Math., 59 (1999), pp. 2060–2085.

[73] D. N. Ostrov, *Solutions of Hamilton-Jacobi equations and scalar conservation laws with discontinuous space-time dependence*, J. Differential Equations, 182 (2002), pp. 51–77.

[74] J. F. Richardson and W. N. Zaki, *Sedimentation and fluidization: Part* I, Trans. Instn. Chem. Engrs. (London), 32 (1954), pp. 35–53.

[75] N. Seguin and J. Vovelle, *Analysis and approximation of a scalar conservation law with a flux function with discontinuous coefficients*, Math. Models Meth. Appl. Sci., 13 (2003), pp. 221–257.

[76] B. Temple, *Global solution of the Cauchy problem for a class of* 2×2 *nonstrictly hyperbolic conservation laws*, Adv. in Appl. Math., 3 (1982), pp. 335–375.

[77] J. D. Towers, *Convergence of a difference scheme for conservation laws with a discontinuous flux*, SIAM J. Numer. Anal., 38 (2000), pp. 681–698.

[78] J. D. Towers, *A difference scheme for conservation laws with a discontinuous flux: The nonconvex case*, SIAM J. Numer. Anal., 39 (2001), pp. 1197–1218.

[79] S. P. Usher, R. G. De Kretser, and P. J. Scales, *Validation of a new filtration technique for dewaterability characterization*, AIChE J., 47 (2001), pp. 1561–1570.

[80] C. J. van Duijn, M. J. de Neef, and J. Molenaar, *Effects of capillary forces on immiscible two-phase flow in strongly heterogeneous porous media*, Transp. Porous Media, 21 (1995), pp. 71–93.

[81] L. B. Verdickt, T. V. Voitovich, S. Vandewalle, K. Lust, I. Y. Smets, and J. F. Van Impe, *Role of the diffusion coefficient in one-dimensional convection-diffusion models for sedimentation/thickening in secondary settling tanks*, Math. Comp. Mod. Dyn. Syst., to appear.

[82] A. I. Volpert, *Generalized solutions of degenerate second-order quasilinear parabolic and elliptic equations*, Adv. Differential Equations, 5 (2000), pp. 1493–1518.

[83] A. I. Vol'pert and S. I. Hudjaev, *Cauchy's problem for degenerate second order quasilinear parabolic equations*, Math. USSR-Sb., 7 (1969), pp. 365–387.

[84] Z. Wu and J. Yin, *Some properties of functions in* BV_x *and their applications to the uniqueness of solutions for degenerate quasilinear parabolic equations*, Northeastern Math. J., 5 (1989), pp. 395–422.

[85] Y. C. Yortsos and J. Chang, *Capillary effects in steady-state flow in heterogeneous cores*, Transp. Porous Media, 5 (1990), pp. 399–420.

SIAM J. APPL. MATH.
Vol. 65, No. 3, pp. 941–963

EXACT SOLUTIONS FOR THE EVOLUTION OF A BUBBLE IN STOKES FLOW: A CAUCHY TRANSFORM APPROACH*

DARREN CROWDY† AND MICHAEL SIEGEL‡

Abstract. A Cauchy transform approach to the problem of determining the free surface evolution of a single bubble in Stokes flow is developed. A number of exact solutions to a class of problems have been derived in the literature using conformal mapping theory, and these solutions are retrieved and further generalized using the new formulation. Certain quantities which are conserved by the dynamics are also identified, the existence of which had not previously been pointed out. A principal purpose of this paper is to use the new formulation to understand when it is possible to externally specify the evolution of the bubble area in such classes of exact solution. It is found to be possible only for certain types of far-field boundary conditions.

Key words. bubble, Stokes flow, complex variables, Cauchy transform

AMS subject classifications. 76D27, 30E25

DOI. 10.1137/S0036139903430847

1. Introduction.

It is a well-known yet remarkable fact that large classes of exact solutions can be found for unsteady two-dimensional (2-D) Stokes flow with a free surface, both with and without surface tension. The solutions follow from the application of powerful complex variable methods; most often these methods involve the use of a conformal mapping $z(\zeta, t)$ to reformulate the free boundary problem as a boundary value problem on a fixed domain in the ζ-plane (assumed in this paper to be a unit disk). The free boundary evolution is then conveniently described by the functional form of the map $z(\zeta, t)$.

Among the first results from the application of complex variable methods to free boundary problems for 2-D Stokes bubbles are the steady bubble solutions of Richardson [18, 19]. Antanovskii [2] later constructed exact unsteady solutions in the case when the asymptotic form of the far-field flow is given by an mth order irrotational straining flow. Independently, Tanveer and Vasconcelos [23] derived explicit unsteady solutions in the form of polynomial mappings in the case when the far-field flow is purely linear, e.g., pure straining flow or simple shear flow with $m = 1$. They also constructed new exact solutions for an expanding/contracting bubble in a quiescent flow. Antanovskii [3] also constructed explicit steady solutions in the case of nonlinear[1] and rotational far-field conditions. The latter solutions were applied as a simple 2-D model for flow in Taylor's four roller mill. Siegel [22] later generalized these results to include explicit unsteady solutions for certain nonlinear rotational far-field conditions, including the time-dependent evolution of the solutions in [3]. Additionally, there is a large amount of literature describing related developments for the evolution of viscous blobs in 2-D Stokes flow (see, e.g., Howison and Richardson [14] and the

*Received by the editors July 1, 2003; accepted for publication (in revised form) July 15, 2004; published electronically March 31, 2005.
http://www.siam.org/journals/siap/65-3/43084.html

†Department of Mathematics, Imperial College of Science and Technology, London SW7 2AZ, England (d.crowdy@imperial.ac.uk). The research of this author was partially supported by EPSRC grant GR/R40104/01.

‡Department of Mathematical Sciences, NJIT, Newark, NJ 07102 (misieg@m.njit.edu). The research of this author was supported in part by NSF grant DMS-0104350.

[1] A precise definition of the term "nonlinear" in this context is given in section 3.3.1.

references therein). Complex variable methods have also been applied to study the evolution of bubbles in Hele–Shaw flow. Cummings, Howison, and King [10] provide a comparison of the developments in Hele–Shaw flow and 2-D Stokes flows. Aside from their intrinsic mathematical interest, these investigations have led to improved understanding of the formation of cusp singularities in free surface flow. The exact 2-D solutions are also useful as an important component of the leading-order solution to three-dimensional flow in slender geometries [9].

The aforementioned exact solutions for 2-D Stokes flow and Hele–Shaw flow all take the form of rational conformal mappings which may be written as

$$z(\zeta,t)=\frac{a_0(t)+a_1(t)\zeta+\cdots+a_N(t)\zeta^N}{\zeta(1+b_1(t)\zeta+\cdots+b_M(t)\zeta^M)}. \tag{1}$$

For convenience we have chosen $b_0(t)=1$ (this is always possible through a redefinition of the other coefficients); if $b_i(t)\equiv 0$, then the mapping reduces to a simpler polynomial form. Note that the extra ζ term in the denominator of (1) is due to the mapping of the inside of the unit disk to the exterior of the bubble. When the dynamics preserves the form (1), the free boundary evolution reduces from an infinite-dimensional dynamical system (namely, the original governing PDEs) to a finite system of ODEs from which one can compute the $N+M+1$ parameters of the conformal mapping from given initial data and external flow.[2]

In essence, demonstration of the existence of a solution of the form (1) for a given free boundary problem involves two key steps. First, it must be shown that the form (1) is preserved in time; i.e., if $z(\zeta,0)$ is a rational function of the form (1) for some M,N, then, as long as the solution exists, $z(\zeta,t)$ remains a rational function with the same M,N. In this paper we shall refer to this requirement as the closure condition. In particular, this implies that the number and type of (pole) singularities in the complex plane (i.e., $|\zeta|>1$) is invariant with time. Second, it is generally required that solutions $z(\zeta,t)$ do not generate any flow singularities (i.e., sources or sinks) in the *finite* fluid domain, although we do allow sources or sinks at infinity, corresponding to expanding/contracting bubble area.

For the solutions described in [2, 23] the closure condition is automatically satisfied, i.e., without any restrictions on the map coefficients $a_i,\ b_i$. The governing ODEs for the coefficients $a_i,\ b_i$ therefore come strictly from the second requirement, which gives $N+M$ conditions for $N+M+1$ unknowns. As a final condition, one is free to specify the time rate of change of the bubble area or, in other words, to specify the existence of a time-dependent source or sink fixed at infinity.

In contrast, for the solutions derived in [22] the closure condition imposes a constraint on the map coefficients $a_i,\ b_i$, in addition to the $N+M$ constraints supplied by the second condition above. Thus, seemingly, one is not free to arbitrarily specify the bubble area. It therefore seems rather fortuitous that the constant area condition employed in [22] is correct, i.e., that the dynamics actually preserves bubble area. However, it is not at all clear from the discussion there how to ascertain if or when this is the case.

In this paper, we provide a general discussion of when it is possible to obtain exact solutions to a very broad class of problems which encompasses and expands the class investigated in [2, 22, 23]. Additionally, we specify precisely when it is possible to externally control the bubble area and, in cases for which the area is not controllable,

[2]Solutions to such a system of ODEs are commonly referred to as exact solutions.

provide a simple means of ascertaining its time dependence. Our analysis involves a different approach from the one given in [2, 22, 23], namely, that emphasis is placed on consideration of the Cauchy transform which we define here as

$$\frac{1}{2\pi i}\oint_{\partial D(t)} \frac{\bar{z}'}{z'-z}dz', \tag{2}$$

where $\partial D(t)$ denotes the boundary of the fluid domain $D(t)$. When z is inside the bubble, the line integral (2) defines an analytic function, say $C(z,t)$. The analytic continuation of $C(z,t)$ outside the bubble (and into the fluid region $D(t)$) contains a great deal of information about the bubble shape. In many respects, $C(z,t)$ is a more natural mathematical object to consider; indeed, given the functional form of $C(z,t)$ at each instant it is possible to reconstruct the relevant conformal mapping. The evolution equation for $C(z,t)$ also has a convenient mathematical form which is easy to analyze. In particular, using the new formulation presented here it is possible to resolve such issues as the question of bubble area evolution, a question which often involves formidable calculation in the usual conformal map approach. It is appropriate to mention that the Cauchy transform has been used to great effect by many previous authors in the study of Hele–Shaw free boundary problems [12, 20, 21, 24].

The rest of this paper is organized as follows. In section 2 we introduce the Cauchy transform formulation for the problem of Stokes flow for a single bubble and prove its equivalence to the usual Stokes flow formulation. We also demonstrate the existence of certain conserved quantities which are useful in constructing exact solutions. In section 3 we use the new formulation to retrieve the exact solutions which have been previously presented in the literature and show how new classes of solutions may be derived. In particular, with the new approach the closure condition is easily verified, in contrast to the much more involved calculations that are necessary using the conformal map approach. More importantly, we use our formulation to investigate when it is possible to externally specify the evolution of bubble area. Some concluding remarks are presented in section 4.

2. Mathematical formulation.

2.1. Stokes flow problem. Consider the quasi-steady evolution of a single bubble in an ambient Stokes flow. The fluid inside the bubble is assumed to have zero viscosity, implying that it is a passive fluid with spatially constant pressure, which for convenience is set to zero. We denote the fluid region exterior to the bubble by $D(t)$, and the bubble is denoted by $D_c(t)$. The flow is allowed to be singular at infinity, although we do not consider any flow singularities (such as sources or sinks) in the finite flow domain. Figure 1 gives a schematic. In view of the incompressibility of the flow, it is convenient to introduce a streamfunction $\psi(x,y)$ which satisfies

$$\mathbf{u} = \nabla^{\perp}\psi. \tag{3}$$

It is easily seen that

$$\nabla^4\psi = 0 \quad \text{in } D(t). \tag{4}$$

We assume that surface tension acts on the bubble boundary so that the stress condition is

$$-pn_j + 2e_{jk}n_k = \kappa n_j, \tag{5}$$

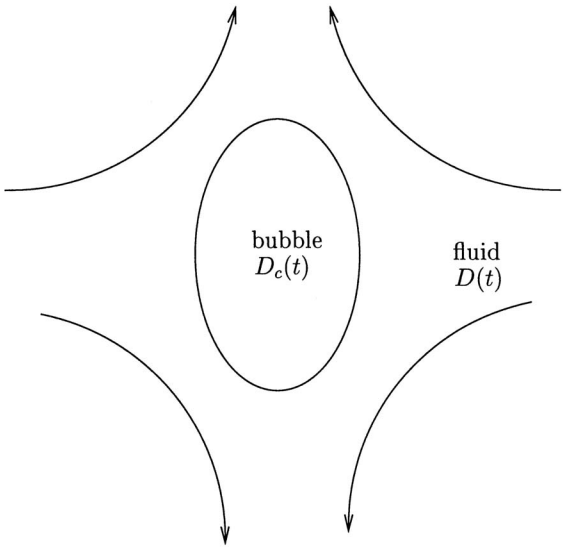


FIG. 1. *Schematic illustrating the bubble region $D_c(t)$ and the fluid region $D(t)$. There is a singular flow field at infinity.*

where κ is the surface curvature (assumed positive for a convex surface), p is the pressure exterior to the bubble, $\mathbf{n}$ is a unit normal pointing outward from the bubble, and

$$e_{jk} = \frac{1}{2}\left(\frac{\partial u_j}{\partial x_k} + \frac{\partial u_k}{\partial x_j}\right)$$

is the rate of strain tensor. In (5) we have assumed that velocities are nondimensionalized by σ/μ, where σ is the surface tension, lengths are nondimensionalized by R, the undeformed bubble radius, and p is nondimensionalized by σ/R. Additionally, time is nondimensionalized by $R\mu/\sigma$. The kinematic condition is that

$$\mathbf{u}\cdot\mathbf{n} = V_n \tag{6}$$

at each point on the interface.

The problem is now reformulated as a problem in analytic function theory following the formulation of Tanveer and Vasconcelos [23]. The general solution of (4) at each instant has the form

$$\psi = \mathrm{Im}[\bar{z}f(z,t) + g(z,t)], \tag{7}$$

where $z = x + iy$ and the overbar denotes complex conjugate. Here $f(z,t)$ and $g(z,t)$ are the *Goursat functions* which are analytic everywhere in the fluid region $D(t)$. In terms of the Goursat functions, the following relations can easily be established:

$$\begin{aligned} p - i\omega &= 4f'(z,t), \\ u + iv &= -f(z,t) + z\overline{f'(z,t)} + \overline{g'(z,t)}, \\ e_{11} + ie_{12} &= z\overline{f''(z,t)} + \overline{g''(z,t)}, \end{aligned} \tag{8}$$

where ω is the vorticity and u, v are the x- and y-components of velocity. Defining s to be the arclength traversed in a counterclockwise direction around the bubble boundary $z(s,t)$, the stress boundary condition can be written in the form

$$f(z,t) + z\overline{f'(z,t)} + \overline{g'(z,t)} = -i\frac{z_s}{2} \quad \text{on } \partial D(t). \tag{9}$$

Using the second equation of (8) and (9), the kinematic condition can be written as

$$\operatorname{Im}\left[(z_t + 2f)\,\bar{z}_s\right] = -\frac{1}{2} \quad \text{on } \partial D(t). \tag{10}$$

Equations (9) and (10) are supplemented by boundary conditions on $f(z,t)$ and $g'(z,t)$ at infinity. In this paper we consider far-field conditions of the form

$$f(z,t) = f_n z^n + \cdots + f_0 + O\left(\frac{1}{z}\right), \tag{11}$$

$$g'(z,t) = g_m z^m + \cdots + g_0 + O\left(\frac{1}{z}\right). \tag{12}$$

2.2. Evolution of the Cauchy transform. Given the above formulation, it is instructive to consider the evolution of the Cauchy transform $C(z,t)$ introduced in (2). The following result is central to the subsequent developments in this paper.

THEOREM 2.1 (evolution of the Cauchy transform). *The Stokes flow problem described in section* 2.1 *is equivalent to the equation*

$$\frac{\partial C(z,t)}{\partial t} + \frac{\partial I(z,t)}{\partial z} = R(z,t) \tag{13}$$

together with (10) *and boundary condition* (11). *Here* $C(z,t)$ *and* $I(z,t)$ *are defined for* $z \in D_c(t)$ *(i.e., inside the bubble) by*

$$\begin{aligned} C(z,t) &= \frac{1}{2\pi i}\oint_{\partial D(t)} \frac{\bar{z}'}{z'-z}dz', \\ I(z,t) &= \frac{1}{2\pi i}\oint_{\partial D(t)} \frac{-2f(z',t)\bar{z}'}{z'-z}dz' \end{aligned} \tag{14}$$

and where the forcing term $R(z,t)$ *is given by*

$$R(z,t) = 2\left(g_m z^m + g_{m-1}z^{m-1} + \cdots + g_0\right). \tag{15}$$

Remark. The viscous sintering problem, which involves the evolution of fluid drops (rather than bubbles), has been investigated using the Cauchy transform approach by Crowdy [5]. There, the Cauchy transform of the domain was defined by the area integral

$$C(z,t) \equiv \frac{1}{\pi}\int\int_{D(t)} \frac{dx'dy'}{z'-z}. \tag{16}$$

In the case of a bounded fluid region, (16) is equivalent to our current definition (2), by the complex form of Green's theorem [1]. Note also that, in the viscous sintering problem, the inhomogeneous term $R(z,t)$ is absent since there are no singularities present to drive the flow; i.e., the flow is driven purely by surface tension. For the

Stokes bubbles considered here, the term $R(z,t)$ represents the forcing due to the singular behavior of the flow at infinity.

Remark. In the viscous sintering problem, [5] provides a proof of the forward result for Theorem 2.1, i.e., that a solution to the Stokes flow problem satisfies (13) (with $R(z,t)=0$). This paper provides the first proof of the "backward" result, and hence the equivalence of the two formulations. Although the proof here is for bubbles, it may easily be modified to the case of fluid drops.

Proof of Theorem 2.1. We first prove the "backward" result; i.e., we assume the existence of a solution $C(z,t)$, $f(z,t)$ to (10), (13) (with $f(z,t)$ analytic in $D(t)$ and having far-field behavior (11)) and show that (9) is satisfied for an appropriately defined $g'(z,t)$ that is analytic in the fluid domain $D(t)$ and satisfies (12). By direct differentiation we have

$$\frac{\partial C(z,t)}{\partial t} = -\frac{1}{2\pi i}\oint_{\partial D(t)} \frac{z_t' \bar{z}' dz'}{(z'-z)^2} + \frac{1}{2\pi i}\oint_{\partial D(t)} \frac{(\bar{z}' dz')_t}{z'-z}. \tag{17}$$

Next, the kinematic condition (10) is rewritten in the form

$$(\bar{z}dz)_t = z_t d\bar{z} + \bar{z}(dz_t) + 2f d\bar{z} - 2\bar{f}dz + ids. \tag{18}$$

Substituting (18) into (17) gives the equation

$$\begin{aligned}\frac{\partial C(z,t)}{\partial t} &= \frac{1}{2\pi i}\oint_{\partial D(t)} \left[\frac{-z_t' \bar{z}' dz'}{(z'-z)^2} + \frac{{z'}_t d\bar{z}' + \bar{z}'(dz')_t}{z'-z}\right] \\ &\quad + \frac{1}{2\pi i}\oint_{\partial D(t)} \frac{2f(z',t)d\bar{z}' - 2\overline{f}(z',t)dz' + ids}{z'-z}.\end{aligned} \tag{19}$$

The terms in square brackets represent a total (spatial) differential of ${z'}_t\bar{z}'/(z'-z)$ which is assumed to be single-valued. Therefore the first integral term is zero. After replacing $\partial C/\partial t$ using (13), we obtain

$$\begin{aligned}R(z,t) &+ \frac{1}{2\pi i}\frac{\partial}{\partial z}\oint_{\partial D(t)} \frac{2f(z',t)\bar{z}'}{z'-z}\, dz' \\ &- \frac{1}{2\pi i}\oint_{\partial D(t)} \frac{2f(z',t)d\bar{z}' - 2\overline{f}(z',t)dz' + ids}{z'-z} = 0.\end{aligned} \tag{20}$$

But

$$\begin{aligned}\frac{1}{2\pi i}\frac{\partial}{\partial z}\oint_{\partial D(t)} \frac{f(z',t)\bar{z}'}{z'-z}\, dz' &= \frac{1}{2\pi i}\oint_{\partial D(t)} \frac{f(z',t)\bar{z}'}{(z'-z)^2}\, dz' \\ &= \frac{1}{2\pi i}\oint_{\partial D(t)} \frac{f_{z'}(z',t)\bar{z}' + f(z',t)\bar{z}'_{z'}}{z'-z}\, dz',\end{aligned} \tag{21}$$

where the subscript z' denotes partial differentiation, and the latter equality follows after integration by parts. Substituting (21) into (20) then yields

$$R(z,t) + \frac{1}{2\pi i}\oint_{\partial D(t)} \frac{\left(2f_{z'}(z',t)\bar{z}' + 2\overline{f}(z',t)\right) - ids}{z'-z} = 0. \tag{22}$$

It is convenient to introduce the notation

$$\tilde{f}(z,t) = f(z,t) - (f_n z^n + \cdots + f_0)$$

so that $\tilde{f}(z,t)$ represents the component of $f(z,t)$ that decays to zero as $z \to \infty$. By the well-known properties of Cauchy integrals, we have for $z \in D_c(t)$

$$\frac{1}{2\pi i} \oint_{\partial D(t)} \frac{\tilde{f}(z',t)}{z'-z}\, dz' = 0, \tag{23}$$

$$\frac{1}{2\pi i} \frac{\partial}{\partial z} \oint_{\partial D(t)} \frac{\tilde{f}(z',t)}{z'-z}\, dz' = 0, \tag{24}$$

$$\frac{1}{2\pi i} \oint_{\partial D(t)} \frac{R(z',t)}{z'-z}\, dz' = R(z,t). \tag{25}$$

Equation (22) is now modified by adding twice the conjugate of (23) and the product of $2\bar{z}$ with (24) to it. After employing (25), the modified equation is written as

$$\overline{\Phi}(z,t) + \bar{z}\Phi'(z,t) + \Psi(z,t) = 0 \tag{26}$$

for $z \in D_c(t)$, where

$$\Phi(z,t) = \frac{1}{2\pi i} \oint_{\partial D(t)} \frac{2\tilde{f}(z',t)}{z'-z}\, dz', \tag{27}$$

$$\Psi(z,t) = \frac{1}{2\pi i} \oint_{\partial D(t)} \frac{R(z',t) + 2f_{z'}(z',t)\bar{z}' + 2\overline{f}(z',t)}{z'-z}\, dz',$$

$$- \frac{1}{2\pi i} \oint_{\partial D(t)} \frac{ids}{z'-z}. \tag{28}$$

Additionally, upon taking the limit of (26) as $z \to \tau$, a point on the boundary of $D_c(t)$, we obtain the relation

$$\overline{\Phi}(\tau,t) + \bar{\tau}\Phi'(\tau,t) + \Psi(\tau,t) = 0. \tag{29}$$

The functions $\Phi(z,t)$ and $\Psi(z,t)$ are analytic in $D_c(t)$, and the boundary condition (29) is identical to the one satisfied by the Goursat functions in the plane problem of the theory of elasticity for the region $D_c(t)$, under the assumption that the boundary of the region is free from the action of external forces. It follows from the theorem of uniqueness [15] of the solution to the plane problem of elasticity that

$$\Phi(z,t) = i\alpha z + \beta,$$
$$\Psi(z,t) = -\bar{\beta},$$

where α is a real and β is a complex constant. But from (23), $\Phi(z) = 0$ in $D_c(t)$, implying that $\alpha = \beta = 0$. Hence

$$\Phi(z,t) = 0, \qquad \Psi(z,t) = 0. \tag{30}$$

From the second identity of (30) it is concluded that the following function may be analytically continued to $D(t)$:

$$\chi(z,t) = g_m z^m + \cdots + g_0 + \overline{f}(z,t) + \bar{z}f'(z,t) - \frac{i}{2z_s}.$$

Upon associating $g'(z,t)$ with $g_m z^m + \cdots + g_0 - \chi(z,t)$, it is concluded that (10)–(13) determine $f(z,t)$ and $g'(z,t)$ as analytic functions in $D(t)$ that satisfy the boundary condition (9) and far-field condition (12). This proves the "backward" result.

We next consider the "forward" problem; i.e., we show that the Stokes problem (9)–(12) implies (13). First, take the conjugate of (9) and use the fact that $z_s\bar{z}_s = 1$ to write the stress condition in the form

$$2\bar{f}dz + 2\bar{z}df + 2dg = ids.$$

This equation is combined with the kinematic condition (10) to give

$$(\bar{z}dz)_t = 2dg + 2d(\bar{z}f) + \bar{z}dz_t + z_t d\bar{z}. \tag{31}$$

Substituting (31) into (17) then yields the equation

$$\frac{\partial C(z,t)}{\partial t} = \frac{1}{2\pi i}\oint_{\partial D(t)} \frac{2g'(z',t)dz'}{z'-z} + \frac{2d(\bar{z}'f)}{z'-z} + \left[-\frac{z_t'\bar{z}'dz'}{(z'-z)^2} + \frac{\bar{z}'(dz')_t}{z'-z} + \frac{z_t'd\bar{z}'}{z'-z}\right].$$

The terms in square brackets represent a total (spatial) differential and therefore give zero total contribution to the integral. The first integral term on the right-hand side is rewritten using

$$\frac{1}{2\pi i}\oint_{\partial D(t)} \frac{2g'(z',t)}{z'-z}\,dz' = 2(g_m z^m + \cdots + g_0) = R(z,t).$$

Finally, using integration by parts on the second integral term and rearranging, we obtain (13), which completes the proof. □

2.3. Analytic continuation inside the fluid region. The functions $C(z,t)$, $I(z,t)$, and $R(z,t)$ defined by the integrals (14) and (15) are all analytic inside the bubble. We now consider the analytic continuations of these functions inside the fluid domain $D(t)$, i.e., exterior to the bubble. By the continuation principle, (13) is also the equation relating these (analytically continued) functions inside $D(t)$.

It is assumed that the bubble boundary $\partial D(t)$ is an analytic curve. This implies that there exists a (unique) function (known as the *Schwarz function* of the curve [11]) analytic inside an annular domain containing the curve $\partial D(t)$ which satisfies the equation

$$\bar{z} = S(z,t) \tag{32}$$

everywhere on the curve $\partial D(t)$. But, by the Plemelj formulae,

$$S(z,t) = C(z,t) - C_i(z,t), \tag{33}$$

$$-2f(z,t)S(z,t) = I(z,t) - I_i(z,t), \tag{34}$$

$$2g'(z,t) = R(z,t) - R_i(z,t), \tag{35}$$

where, for $z \in D(t)$, the functions $C_i(z,t)$, $I_i(z,t)$, and $R_i(z,t)$ are given by the integrals

$$C_i(z,t) = \frac{1}{2\pi i}\oint_{\partial D(t)} \frac{\bar{z}'}{z'-z}dz', \tag{36}$$

$$I_i(z,t) = \frac{1}{2\pi i}\oint_{\partial D(t)} \frac{-2f(z',t)\bar{z}'}{z'-z}dz', \tag{37}$$

$$R_i(z,t) = \frac{1}{2\pi i}\oint_{\partial D(t)} \frac{2g'(z',t)}{z'-z}dz'. \tag{38}$$

The formulae (33)–(35) provide expressions for the analytic continuations of $C(z,t)$, $I(z,t)$, and $R(z,t)$ into the fluid domain $D(t)$.

2.4. Conservation of finite poles of $C(z,t)$. The following result will be useful in constructing exact solutions to the Stokes flow problem (10)–(13).

Let $C(z,t)$, $f(z,t)$ be solutions to (10)–(13). If $C(z,0)$ initially has a pole at a finite point $z_j(0)$ inside the fluid domain, then (provided the solution exists) $C(z,t)$ continues to have a pole at the point $z_j(t)$, where $z_j(t)$ satisfies the ODE

$$\dot{z}_j(t) = -2f(z_j(t),t). \tag{39}$$

To derive the result, first note that the complex form of Green's theorem [1] can be used to write $I(z,t)$ in the form

$$I(z,t) = -\frac{1}{2\pi i}\oint_{\partial D(t)} \frac{2f(z',t)-2f(z,t)}{z'-z}\ \bar{z}'dz' - 2f(z,t)C(z,t). \tag{40}$$

Define the function $\Sigma(z,t)$ as

$$\Sigma(z,t) = -\frac{1}{2\pi i}\oint_{\partial D(t)} \frac{2f(z',t)-2f(z,t)}{z'-z}\ \bar{z}'dz'. \tag{41}$$

The singularity of the integrand in (41) is removable, and $\Sigma(z,t)$ is therefore analytic in $D(t)$. Thus,

$$I(z,t) = \Sigma(z,t) - 2f(z,t)C(z,t), \tag{42}$$

which, when substituted into (13), yields the following PDE for the Cauchy transform:

$$\frac{\partial C}{\partial t} - 2f(z,t)\frac{\partial C}{\partial z} - 2\frac{\partial f(z,t)}{\partial z}C(z,t) + \frac{\partial \Sigma}{\partial z}(z,t) = R(z,t). \tag{43}$$

This equation also governs the analytic continuation of $C(z,t)$ inside $D(t)$. But inside $D(t)$, (43) has the form of a first-order linear equation for $C(z,t)$ with coefficients that are known a priori to be analytic in $D(t)$. Thus, provided they exist, solutions for $C(z,t)$ will have the same analytic structure inside $D(t)$ as solutions of a first-order linear PDE with analytic coefficients. Using the well-known theory of such equations, it is deduced that the pole singularities are preserved and move on characteristics. In this case,

$$-\dot{z}_j(t) - 2f(z_j(t),t) = 0, \tag{44}$$

and the result is demonstrated.

Remark. Note that the above result says nothing about any singularities of $C(z,t)$ at infinity. In the present application, the coefficient function $f(z,t)$ is singular at infinity (and, in the language of ODEs [13], is therefore a *fixed singularity* of (43)).

Moreover, depending on the far-field flow, $R(z,t)$ may also be singular at infinity. The behavior of $C(z,t)$ at $z \to \infty$ must be examined by a local analysis of (13) in each case. Such an analysis will be seen to be crucial in determining whether $C(z,t)$ can consistently preserve a rational function form under evolution, i.e., satisfy the closure condition. When this is the case, the pole at infinity does not move into the finite complex plane. These issues are discussed further in section 3.

Remark. The results of this section are closely related to the results of section 4.5 of Cummings, Howison, and King [10], where rather different arguments are used.

2.5. Circulation theorem and conserved quantities. In the context of the viscous sintering problem, Crowdy [5] derived a circulation theorem which provides the existence of conserved quantities associated with certain choices of initial conditions. This theorem can be extended to the case of a bubble in an infinite flow.

Let the curve γ_j be a *fixed* closed curve surrounding the isolated pole singularity $z_j(t) \in D(t)$. Because $z_j(t)$ moves at finite speed (see (39)), γ_j can always be chosen so that it continues to enclose the (moving) singularity $z_j(t)$, at least for sufficiently short times. Consider the circulation-type quantity

$$\oint_{\gamma_j} C(z,t)dz. \tag{45}$$

Because γ_j is assumed fixed, we have

$$\frac{d}{dt}\oint_{\gamma_j} C(z,t)dz = \oint_{\gamma_j} \frac{\partial C(z,t)}{\partial t}dz = \oint_{\gamma_j} \left(-\frac{\partial I(z,t)}{\partial z} + 2g'(z,t) + R_i(z,t)\right) dz. \tag{46}$$

But $R_i(z,t)$ is analytic everywhere inside γ_j, as is $g'(z,t)$ if the flow has no singularities in the finite plane. Use of Cauchy's theorem then implies that

$$\frac{d}{dt}\oint_{\gamma_j} C(z,t)dz = -\left[I(z,t)\right]_{\gamma_j}, \tag{47}$$

where the square bracket denotes the change in $I(z,t)$ around the contour γ_j. But, by (33) and (34), while $I(z,t)$ has a simple pole at $z_j(t)$, it is single-valued so that the right-hand side of (47) vanishes. This implies that

$$\oint_{\gamma_j} C(z,t)dz = A_j(t), \qquad j = 1, \ldots, N, \tag{48}$$

are constants of the motion, i.e.,

$$A_j(t) = A_j(0). \tag{49}$$

Note that in (48) we have used the fact that the simple pole of $C(z,t)$ at $z_j(t)$ is preserved in time.

The existence of the above conserved quantities has not been pointed out in any previous studies of bubbles in ambient Stokes flows. It is, however, closely related to a similar result that has been observed in the context of evolving viscous blobs [5, 17, 6].

3. Exact solutions. The advantage of considering the Cauchy transform $C(z,t)$ is that its evolution equation is particularly simple. In certain situations, it will be seen that $C(z,t)$ can retain a rational function form under evolution. In this case, the corresponding solutions will be called *exact* in the sense that the evolution depends on just a *finite set* of time-evolving parameters. As an example of the simplicity of the evolution equations, suppose $C(z,0)$ is rational with a finite distribution of simple pole singularities. Suppose too that it is known that $C(z,t)$ preserves this functional form under evolution. Then, if $C(z,0)$ has the form

$$C(z,0) = \sum_{j=1}^{N} \frac{A_j(0)}{z - z_j(0)}, \tag{50}$$

then, by the results above, $C(z,t)$ has the form

$$C(z,t) = \sum_{j=1}^{N} \frac{A_j(0)}{z - z_j(t)}, \tag{51}$$

where

$$\dot{z}_j(t) = -2f(z_j(t),t), \qquad j = 1,\dots,N. \tag{52}$$

With the singularities under control at all regular points of the evolution equation (13), the question of whether $C(z,t)$ can consistently preserve a given rational function form must be determined by a local analysis of (13) at the fixed singularity at infinity. This will be illustrated in the context of the examples considered in section 3.1.

Even supposing that $C(z,t)$ indeed evolves as a rational function, it still remains to reconstruct the corresponding time-evolving domain from knowledge of the Cauchy transform. This can be done using conformal maps, but it is important to observe that in the case of unbounded domains such as here, knowledge of the Cauchy transform does not uniquely determine the unbounded domain. Rather, it determines it only up to a real degree of freedom which can be associated with the freedom to specify the bubble area. The appendix provides a discussion of this point in terms of the inverse problem of 2-D potential theory.

None of the authors Antanovskii [2], Siegel [22], or Tanveer and Vasconcelos [23] use the above formulation but instead make direct use of a conformal map formulation in a parametric ζ-plane. It is instructive to retrieve the conformal mapping solutions of [2, 22, 23] using the above perspective. In this way, certain advantages of the Cauchy transform formulation will become apparent. In particular, we gain important insight into the bubble area evolution in each case and see exactly when it is possible to specify it externally.

We note that, in the process of retrieving the conformal mapping solutions of [2, 22, 23], the values of $f(z_j(t),t)$ are computed with the aid of the conformal map. One can bypass the introduction of a conformal map and compute f from the kinematic condition (10) by using an alternative representation of the boundary (e.g., algebraic curves; see [8]). Although this may require the use of a boundary integral numerical calculation, one still has the advantage of a finite/exact representation of the interface, since the preserved rational function form of the Cauchy transform implies that the shape can be described by a small finite number of parameters.

3.1. Exact solutions of Tanveer and Vasconcelos [23].

3.1.1. The Cauchy transform. Tanveer and Vasconcelos assume that the far-field form of $f(z,t)$ and $g'(z,t)$ are of the form

$$f(z,t) \sim f_1 z + \mathcal{O}(1), \tag{53}$$

$$g'(z,t) \sim g_1 z + \mathcal{O}(1). \tag{54}$$

Specifically, in the notation of Tanveer and Vasconcelos [23],

$$f_1 = \frac{1}{4}\left(\frac{p_\infty(t)}{\mu} - i\omega_0\right) \tag{55}$$

and

$$g_1 = \frac{1}{2}\left(\alpha_0 - i\beta_0\right). \tag{56}$$

$p_\infty(t)$ and ω_0 are the fluid pressure and vorticity in the far-field, respectively, while α_0 and β_0 characterize the strain rates of a linear straining flow at infinity.

As an illustrative example, we take the first case considered by Tanveer and Vasconcelos [23], that is, a bubble in a shear flow of the form

$$\mathbf{u} = (\Gamma y, 0). \tag{57}$$

This corresponds to the far-field values $\alpha_0 = 0$ and $\beta_0 = -\omega_0 = \Gamma$ in the notation of (55) and (56), or equivalently

$$f_1 = \frac{i\Gamma}{4}, \qquad g_1 = -\frac{i\Gamma}{2} \tag{58}$$

in the notation of (53) and (54).

Tanveer and Vasconcelos [23] show the existence of exact solutions that are polynomial maps, i.e., of the form (1) with $b_i = 0$. To retrieve these, let us seek solutions in which $C(z,t)$ has the rational function form

$$C(z,t) = A(t)z; \tag{59}$$

that is, the only singularity of $C(z,t)$ in the fluid region is a single simple pole singularity at infinity. It is necessary to check that this is a consistent solution of (13). To do this, we analyze (13) in the neighborhood of infinity and find that we require

$$\frac{\partial}{\partial t}\left(A(t)z\right) + \frac{\partial}{\partial z}\left(-2f_1 z(A(t)z)\right) + o(z) = 2g_1 z + o(z). \tag{60}$$

This is consistent, provided $A(t)$ satisfies

$$\dot{A}(t) - 4f_1 A(t) = 2g_1, \tag{61}$$

where this equation comes from equating coefficients of z in (60).

3.1.2. Conformal mapping. We now consider the conformal maps for which the corresponding Cauchy transform $C(z,t)$ has the form (16). Consider, for example, the mapping from the unit ζ-disc given by

$$z(\zeta,t) = \frac{a}{\zeta} + b\zeta, \tag{62}$$

where a and b are functions of time and where a can be assumed real (using a rotational degree of freedom of the Riemann mapping theorem). Note that, on the unit ζ-circle,

$$\bar{z} = a\zeta + \frac{\bar{b}}{\zeta}. \tag{63}$$

From (62) we have that

$$\frac{1}{\zeta} = \frac{z}{a} - \frac{b}{a}\zeta. \tag{64}$$

Using (64) in (63) we obtain

$$\bar{z} = \frac{\bar{b}}{a}z + \left(a - \frac{|b|^2}{a}\right)\zeta, \tag{65}$$

which is valid on the unit ζ-circle. By comparison with (33) we make the identifications

$$C(z,t) = \frac{\bar{b}}{a}z, \qquad C_i(z,t) = -\left(a - \frac{|b|^2}{a}\right)\zeta(z,t), \tag{66}$$

where we have also used the fact that $C_i(z,t)$ decays at infinity. This shows that all conformal maps of the form (62) have corresponding Cauchy transforms of the form (16) with

$$A = \frac{\bar{b}}{a}. \tag{67}$$

Note that a and b can be multiplied by any real constant (corresponding to changing the area of the elliptical bubble) and the Cauchy transform (16) remains unchanged. A degree of freedom is therefore available, and this is used up, following Tanveer and Vasconcelos [23], by specifying the area of the bubble to be fixed in time, i.e.,

$$a^2 - |b|^2 = R^2, \tag{68}$$

where R is constant. (Alternatively, one can specify the bubble area to be an arbitrary function of time, corresponding to the presence of a source or sink at infinity.) This gives one equation relating a and b as derived by Tanveer and Vasconcelos [23]. The second equation obtained by them is

$$\frac{d(ab)}{dt} = -(2I_0 + i\Gamma)ab + i\Gamma a^2, \tag{69}$$

where $I_0 = I(0,t)$ and

$$I(\zeta,t) = \frac{1}{4\pi i}\oint_{|\zeta'|=1} \frac{d\zeta'}{\zeta'}\left[\frac{\zeta+\zeta'}{\zeta'-\zeta}\right]\frac{1}{|z_\zeta(\zeta',t)|}. \tag{70}$$

This equation must, of course, be equivalent to (61). To see this, note that upon applying the Poisson integral formula [1] to the kinematic condition it can be shown that

$$z_t(\zeta,t) + 2f(z(\zeta,t),t) = \zeta\left[I(\zeta,t) + \frac{i\omega_0}{2}\right] z_\zeta(\zeta,t) \tag{71}$$

(see Tanveer and Vasconcelos [23]). Substituting the conformal map (62) into this equation and equating powers of ζ^{-1} provides

$$\dot{a} + 2f_1 a = -a\left(I_0 + \frac{i\omega_0}{2}\right). \tag{72}$$

Eliminating f_1 in (61) using (72), as well as using (67) and (58), gives the required (69).

3.2. Exact solutions of Antanovskii [2]. Antanovskii [2] assumes the far-field asymptotic form of $f(z,t)$ and $g'(z,t)$ to be

$$f(z,t) \sim f_1 z + \mathcal{O}(1/z), \tag{73}$$

$$g'(z,t) \sim g_m z^m + \mathcal{O}(1/z), \tag{74}$$

where $m \geq 1$ is some positive integer and f_1 is some real time-dependent function that is independent of z. This gives rise to a situation in which the far-field flow is irrotational and given by an mth order straining flow.

Anticipating the form of the singularity in $C(z,t)$ at infinity in order to satisfy the closure condition, we seek solutions in which $C(z,t)$ is an mth order polynomial, i.e.,

$$C(z,t) = A_m z^m + A_{m-1} z^{m-1} + \cdots + A_1 z. \tag{75}$$

Analyzing the singularity of (13) at infinity gives the equations

$$\begin{aligned}
\dot{A}_m - 2(m+1)f_1 A_m &= 2g_m,\\
\dot{A}_{m-1} - 2m f_1 A_{m-1} &= 0,\\
\dot{A}_{m-2} - 2(m-1)f_1 A_{m-2} &= 0,\\
&\cdots\\
\dot{A}_1 - 2f_1 A_1 &= 0.
\end{aligned} \tag{76}$$

First, it is immediately clear that if $A_{m-1}(0) = A_{m-2}(0) = \cdots = A_1(0) = 0$, then $A_{m-1}(t) = A_{m-2}(t) = \cdots = A_1(t) = 0$. It therefore remains only to satisfy the equation for A_m, viz.,

$$\dot{A}_m - 2(m+1)f_1 A_m = 2g_m. \tag{77}$$

3.2.1. Conformal mapping. Now consider the class of conformal maps given by

$$z(\zeta,t) = \frac{a}{\zeta} + b\zeta^m, \tag{78}$$

where a is again assumed to be real. On the unit ζ-circle we have

$$\overline{z}(\zeta^{-1},t) = a\zeta + \frac{\overline{b}}{\zeta^m}. \tag{79}$$

But from (78),

$$\frac{1}{\zeta} = \frac{z}{a} - \frac{b}{a}\zeta^m, \tag{80}$$

which, when substituted into (79), gives

$$S(z(\zeta)) = \overline{z}(\zeta^{-1},t) = a\zeta + \overline{b}\left(\frac{z}{a} - \frac{b}{a}\zeta^m\right)^m \tag{81}$$

from which, by comparison with (33), we deduce that

$$C(z,t) = \frac{\overline{b}}{a^m} z^m. \tag{82}$$

Thus, all maps of the form (78) yield Cauchy transforms of the form $C(z,t) = A_m z^m$ with

$$A_m = \frac{\overline{b}}{a^m}. \tag{83}$$

Note again that b and a can be multiplied by a real number without changing $C(z,t)$, while the only equation to be satisfied is (77). Thus, the relevant evolution equations for the two parameters a and b are (77) along with an area evolution equation which may be arbitrarily specified. These can be shown to be equivalent to those given in equation (37) of Antanovskii [2].

3.3. Exact solutions of Siegel [22].

3.3.1. The Cauchy transform. Siegel [22] assumes the far-field asymptotic form of $f(z,t)$ and $g'(z,t)$ to be

$$f(z,t) \sim f_3 z^3 + f_1 z + \mathcal{O}(1/z), \tag{84}$$

$$g'(z,t) \sim g_1 z + \mathcal{O}(1/z). \tag{85}$$

f_3 and g_1 are externally specifiable and will be taken to be constant in time (g_1 gives a measure of an imposed irrotational straining flow at infinity, while f_3 produces a rotational far-field component). Such boundary conditions in which $f(z)$ is a nonlinear function of z in the far field will be referred to as *nonlinear* far field conditions. Siegel [22] makes the choices

$$g_1 = 1, \qquad f_3 = \frac{\epsilon}{2}. \tag{86}$$

Antanovskii [3] derived exact *steady* solution for the bubble shape subject to the above far-field conditions. The analysis of Siegel [22] essentially generalizes the results of Antanovskii [3] to the case where the bubble evolves in a quasi-steady, time-dependent manner.

Suppose we seek solutions to (13) in which $C(z,t)$ evolves as

$$C(z,t) = \sum_{j=1}^{N} \frac{A_j(t)}{z - z_j(t)} \tag{87}$$

for all times. It is clear that

$$C(z,t) \sim \frac{B(t)}{z} \quad \text{as } z \to \infty, \tag{88}$$

where

$$B(t) = \sum_{j=1}^{N} A_j(t). \tag{89}$$

By the circulation theorem, all the quantities $\{A_j(t)\}$ are constants of the motion. This implies, by (89), that $B(t)$ is also a constant of the motion. Moreover, all the poles $\{z_j(t)\}$ move according to (52). The Cauchy transform is determined at each instant by these equations.

As mentioned earlier, it is important to verify that the solution (87) is a consistent solution of (13) at the fixed singularity at the point at infinity. To see this, note from (36) that

$$C_i(z,t) \sim -\frac{1}{2\pi i}\left(\oint_{\partial D(t)} \bar{z}' dz'\right)\frac{1}{z} + \mathcal{O}(1/z^2), \tag{90}$$

while, from (37),

$$I_i(z,t) \sim \mathcal{O}\left(\frac{1}{z}\right) \tag{91}$$

as $z \to \infty$. But the area of the bubble (denoted $\mathcal{A}$) is precisely

$$\mathcal{A} = \frac{1}{2i}\oint_{\partial D(t)} \bar{z}' dz'. \tag{92}$$

Therefore, using (33) and (88),

$$S(z) = \frac{B(t) + \mathcal{A}/\pi}{z} + \mathcal{O}\left(\frac{1}{z^2}\right) \tag{93}$$

as $z \to \infty$. Inside the fluid domain $D(t)$, (13) takes the form

$$\frac{\partial C(z,t)}{\partial t} + \frac{\partial}{\partial z}\left[-2f(z,t)S(z,t) + I_i(z,t)\right] = 2g'(z,t) + R_i(z,t). \tag{94}$$

A local analysis of this equation as $z \to \infty$ therefore implies that

$$\frac{\partial}{\partial z}\left(-2f_3 z^3\left(\frac{B}{z} + \frac{\mathcal{A}/\pi}{z}\right)\right) + \mathcal{O}(1/z) = 2g_1 z + \mathcal{O}(1/z) \tag{95}$$

so that, equating coefficients of the singularity at $\mathcal{O}(z)$, we have

$$-2(\mathcal{A}/\pi + B)f_3 = g_1. \tag{96}$$

This is a necessary condition if the solution (87) is to be a consistent solution of (13).

It is crucial to note that if $C(z,t)$ is determined at each instant, then, in contrast to the previous examples of Tanveer and Vasconcelos [23] and Antanovskii [2], (96) represents an *additional* constraint on the free boundary evolution. In this case, there is no freedom to externally specify the bubble area evolution. Rather, it is determined implicitly by the exact solution itself. Since A has been shown to be a constant of the motion, (96) implies that, for solutions of the form (87), the area $\mathcal{A}$ of the bubble is necessarily fixed in time if f_3 and g_1 are constant (independent of time), which has been assumed to be the case, i.e.,

$$\mathcal{A} = -\pi \left(B + \frac{g_1}{2f_3} \right). \tag{97}$$

The initial area of the bubble therefore dictates the value of B.

In summary, solutions of (13) in which $C(z,t)$ is rational with a finite set of simple pole singularities (and for which $f(z,t)$ and $g'(z,t)$ have the far-field form (53) and (54)) are admitted, and, in these solutions, the bubble area remains constant. If we seek solutions in which the bubble area is not fixed in time (but varies, for example, at some externally specified rate Q), the solutions will not be such that the Cauchy transform has the far-field behavior (88) for all times $t > 0$. In that case, the functional form of the Cauchy transform will not be of the proposed simple rational character and will not lend itself to exact solutions.

3.3.2. Conformal maps. The corresponding conformal maps from a unit ζ-disk to fluid domains whose Cauchy transforms have a finite distribution of simple pole singularities with the far-field behavior (88) are given by rational functions of the form

$$z(\zeta,t) = \frac{C}{\zeta} \left(\frac{\prod_{j=1}^{N} (\zeta - \eta_j(t))}{\prod_{j=1}^{N} (\zeta - \zeta_j(t))} \right), \tag{98}$$

where C can be assumed real (to use up the rotational degree of freedom in the Riemann mapping theorem). It follows that

$$\overline{z}(\zeta^{-1},t) = C\zeta \left(\frac{\prod_{j=1}^{N} (1 - \zeta\bar{\eta}_j(t))}{\prod_{j=1}^{N} (1 - \zeta\bar{\zeta}_j(t))} \right). \tag{99}$$

It is clear that because $S(z,t) = \bar{z}$, this class of domains is such that the far-field asymptotic form of $C(z,t)$ is of the form in (88). It is also clear that

$$z_j(t) = \overline{z(\zeta_j,t)} \tag{100}$$

so that the N equations (39) can be viewed as providing evolution equations for the parameters $\{\zeta_j(t) | j = 1,\ldots,N\}$. The N constants of motion derived from the circulation theorem (by taking a contour γ_j around each distinct pole $z_j(t)$) provide N additional equations for the parameters $\{\eta_j(t) | j = 1,\ldots,N\}$. Finally, the parameter C is determined by condition (96). Equivalently, one could determine C by ensuring that the bubble area is conserved under evolution.

Siegel [22] considers one of the above solutions in detail. This solution has a mapping given by the special choice

$$z(\zeta,t) = \frac{1}{\zeta} \frac{\gamma_0 + \gamma_1\zeta^2}{1 - \gamma_2\zeta^2}. \tag{101}$$

It is straightforward to see that this map corresponds to domains with the Cauchy transform

$$C(z,t) = \frac{E(t)}{z - z_0(t)} + \frac{E(t)}{z + z_0(t)}, \tag{102}$$

where

$$z_0(t) = z(\sqrt{\gamma_2}, t). \tag{103}$$

In this case,

$$C(z,t) \sim \frac{2E(t)}{z} \quad \text{as } z \to \infty. \tag{104}$$

The relevant equations of motion for this solution are

$$\dot{z}_0(t) = -2f(z_0(t), t), \tag{105}$$

$$E(t) = E(0), \tag{106}$$

$$-\left(\frac{\mathcal{A}}{\pi} + 2E\right) = \frac{g_1}{2f_3}, \tag{107}$$

where the last equation is just (96). Equations (105) and (106) determine $C(z,t)$ at each instant, while (107) additionally constrains the bubble area to be fixed in time. This additional constraint is absent in the solutions of [23] and [2].

The three equations obtained by Siegel [22] are

$$\dot{\gamma}_2 = -2\gamma_2 \left[I(\sqrt{\gamma_2}, t) - \epsilon \frac{\gamma_0^2}{\gamma_2} \right], \tag{108}$$

$$\gamma_1 = \frac{\gamma_2}{\epsilon \gamma_0}, \tag{109}$$

$$\gamma_0 = \left[\frac{2(1+\gamma_1^2)}{c_1 + (c_1^2 - c_2)^{1/2}} \right]^{1/2}, \tag{110}$$

where condition (110) derives from the fact that the bubble area is taken to be constant and equal to π. c_1 and c_2 are defined as

$$c_1 = 1 - \epsilon\gamma_1^2 \left(2(1-\epsilon) + \epsilon\gamma_1^2\right) \quad \text{and} \quad c_2 = 4\epsilon^2\gamma_1^2(1+\gamma_1^2)\left[3 + \epsilon(2+\epsilon)\gamma_1^2\right], \tag{111}$$

while

$$I(\zeta, t) = \frac{1}{2\pi i} \oint_{|\zeta'|=1} \frac{d\zeta'}{\zeta'} \left[\frac{\zeta' + \zeta}{\zeta' - \zeta}\right] \left\{ \frac{1}{|z_\zeta|} + \mathrm{Re}\left[\frac{\epsilon\gamma_0^2}{\zeta^2}\right] \right\}. \tag{112}$$

We now indicate how (105)–(107) are equivalent to (108)–(110). First we use the fact, as derived by Siegel [22], that

$$z_t(\zeta, t) + 2f(z(\zeta,t),t) = \zeta z_\zeta(\zeta,t)\left(-\frac{\epsilon\gamma_0^2}{\zeta^2} + I(\zeta,t)\right) \tag{113}$$

to eliminate $f(z_0(t),t)$ from (105). This reproduces exactly (108). Next recall from (93) that as $z \to \infty$,

$$S(z) \sim \left(2E + \frac{\mathcal{A}}{\pi}\right)\frac{1}{z}. \tag{114}$$

But

$$\begin{aligned} S(z(\zeta,t)) = \overline{z(\zeta,t)} = &= \zeta\left(\frac{\gamma_0\zeta^2+\gamma_1}{\zeta^2-\gamma_2}\right) \\ &\sim -\frac{\gamma_1}{\gamma_2}\zeta \quad \text{as} \quad \zeta \to 0 \\ &\sim -\left(\frac{\gamma_1\gamma_0}{\gamma_2}\right)\frac{1}{z} \quad \text{as} \quad z \to \infty, \end{aligned} \tag{115}$$

where we have used that fact that $\zeta \sim \frac{\gamma_0}{z}$ as $z \to \infty$. Together, (114), (115), (86), and (107) yield Siegel's second equation (109). Finally, the equation for $S(z(\zeta,t),t)$ in (115) also yields the following equation for $E(t)$ in terms of the conformal mapping parameters, viz.,

$$E(t) = \frac{\gamma_0\gamma_2+\gamma_1}{2} z_\zeta(\sqrt{\gamma_2},t) = \frac{\gamma_0\gamma_2+\gamma_1}{2\gamma_2(1-\gamma_2^2)^2}\left(\gamma_1\gamma_2+\gamma_1\gamma_2^3-\gamma_0+3\gamma_0\gamma_2^2\right). \tag{116}$$

Using (116), (86), and the fact that $\mathcal{A} = \pi$ in (107) combine, after some further algebra, to retrieve Siegel's final equation (110).

Remark. The Cauchy transform formulation is crucial in proving that the bubble area is conserved for all maps of the form (98). This is because the bubble area $\mathcal{A}$ appears explicitly in the form of the far-field asymptotics of $C(z,t)$. To establish this general result using the direct conformal mapping approach of [23, 22] would be exceedingly difficult.

Remark. If the conformal map $z(\zeta,t)$ of the form (98) has N poles, the Cauchy transform formulation also immediately implies the existence of precisely N conserved quantities, one associated with each of the N poles.

3.4. A final example. We now give an example to show that the previous result is rather special and that, for general nonlinear flow conditions at infinity, while exact solutions to the problem exist, it is not generally possible to externally specify the bubble area evolution in these solutions.

A natural generalization of the far-field conditions considered in the previous three subsections is

$$f(z,t) \sim f_3 z^3 + f_1 z + \mathcal{O}(z^{-1}), \tag{117}$$

$$g'(z,t) \sim g_3 z^3 + g_1 z + \mathcal{O}(z^{-1}). \tag{118}$$

These far-field conditions are again nonlinear and are exactly those considered by Antanovskii [4] in his studies of the formation of cusped bubbles, although his analysis is restricted to the derivation of classes of exact *steady* solutions. The special case $g_3 = 0$ reduces to the far-field conditions considered in Antanovskii [3] and Siegel [22] (as well as in section 3.3). For this reason, one might expect the bubble area to again be constant under evolution. However, this is not the case, as will now be shown.

Suppose that we seek exact solutions in which $C(z,t)$ has the rational function form

$$C(z,t) = A_0(t)z + \sum_{k=1}^{N} \frac{A_k(t)}{z - z_k(t)} \tag{119}$$

for all times. As $z \to \infty$,

$$C(z,t) \sim A_0(t)z + \left(\sum_{k=1}^{N} A_k(t)\right) \frac{1}{z} + \ldots, \tag{120}$$

and this far-field form is forced by (13), as will be shown. First, it is known immediately that the points $\{z_k(t)|k = 1, 2, \ldots, N\}$ must satisfy

$$\dot{z}_k(t) = -2f(z_k(t), t), \qquad k = 1, \ldots, N, \tag{121}$$

while the circulation theorem implies that

$$A_k(t) = A_k(0), \qquad k = 1, \ldots, N. \tag{122}$$

It remains to determine the evolution of $A_0(t)$, as well as to ensure that (119) is a consistent solution of (13) at $z \to \infty$. Analyzing (13) as $z \to \infty$, at $\mathcal{O}(z^3)$ we get

$$-8f_3 A_0(t) = 2g_3, \tag{123}$$

which implies that $A_0(t)$ must also be constant. $C(z,t)$ is now completely determined, but we expect to be able to specify another degree of freedom associated with the bubble area $\mathcal{A}$. But at $\mathcal{O}(z)$, we get another equation having the form

$$-4\left(f_1 A_0 + f_3 \mathcal{A}\right) = 2g_1. \tag{124}$$

Equation (124) is a further constraint on the solution and can be thought of as an equation governing the bubble area $\mathcal{A}$. It is clear from (124) that the bubble area is not constant this time. Rather, how the bubble area evolves is governed by the exact solution itself. To see this, recall that f_1 (which is related to the far-field pressure evaluated in the near field of the bubble) is a time-evolving quantity whose evolution is governed by (113) and is not externally controllable.

Note that if g_3 is taken equal to zero (so that the far-field conditions of Siegel [22] are retrieved), then necessarily $A_0(t) = 0$ (by (123)), and then $\mathcal{A}$ turns out to be constant (by (124)).

4. Conclusion. We have introduced a Cauchy transform formulation of the problem of Stokes flow for a single bubble and proven its equivalence to the usual formulation of free surface Stokes flow. The new formulation has been used to derive a very broad class of exact solutions (namely, the class for which the Cauchy transform takes the form of a rational function), generalizing the set of solutions which have heretofore appeared in the literature. Indeed, it is surmised that the class of solutions discussed here is maximal, i.e., that it includes all cases for which the evolution is reducible to a finite-dimensional system of ODEs. We use our formulation to investigate when it is possible to externally specify the evolution of bubble area. This issue is extremely difficult to address using other approaches. In general, it is found that when the Goursat function $f(z)$ has a nonlinear far-field

behavior it is *not* possible to find exact solutions *and* specify the bubble area evolution. Instead the bubble area in such cases is determined by the exact solution. In the case of pure linear strain (so that $g'(z) \sim g_1 z$ as $z \to \infty$) it happens that the bubble area is constant, and this situation corresponds to a physically interesting case. However, this occurrence is somewhat coincidental.

Concerning generalizations, it is possible to extend the present formulation to the case of a more general compressible bubble with an externally specified equation of state relating its internal pressure (say, $p_B(t)$) to its area. Pozrikidis [16] has considered such problems using numerical boundary integral methods, while Crowdy [7] has generalized the solutions of Tanveer and Vasconcelos [23] to this case.

Appendix. Inverse problem of potential theory. In this appendix we explain why, in the case of a simply connected unbounded domain, knowledge of $C(z,t)$ determines the domain $D(t)$ up to a single real degree of freedom associated with specification of the bubble area.

Consider a smooth family of bounded, time-evolving, simply connected planar domains $D(t)$ in some time interval $t \in [0,T)$. Define the harmonic moments of these domains to be the integrals of a basis of all functions harmonic in $D(t)$ at time t. Suppose all the moments of $D(t)$ for $t \in [0,T)$ are known. It is a well-known result of the inverse problem of 2-D potential theory that the domains $D(t)$ can be uniquely reconstructed from knowledge of all these harmonic moments. Varchenko and Etingof [24] discuss this result in detail. For a bounded domain, if the Cauchy transform is defined as

$$C(z,t) = \frac{1}{\pi}\int\int_{D(t)} \frac{dx'dy'}{z'-z} = \frac{1}{2\pi i}\oint_{\partial D(t)} \frac{\bar{z}'dz'}{z'-z}, \tag{125}$$

then it is a generating function for the harmonic moments because, Laurent expanding for large $|z|$,

$$C(z,t) = \sum_{n=0}^{\infty} \frac{M_n}{z^{n+1}}, \tag{126}$$

where

$$M_n = \frac{1}{\pi}\int\int_{D(t)} z'^n dx'dy'. \tag{127}$$

The harmonic functions

$$\left\{ \mathrm{Re}\left[z^n\right],\ \mathrm{Im}\left[z^n\right] \middle| n = 0,1,2,\dots \right\} \tag{128}$$

span the space of functions harmonic in $D(t)$. The real and imaginary parts of the set of complex moments (127) generate all the harmonic moments of $D(t)$.

Exactly the same result pertains to the case of unbounded domains $D(t)$, except that if the moments are defined in terms of area integrals over $D(t)$, some of them do not exist, owing to the unboundedness of the domain. This is the reason for our choice of defining the Cauchy transform, from the outset, as the line integral

$$\frac{1}{2\pi i}\oint_{\partial D} \frac{\bar{z}'dz'}{z'-z}. \tag{129}$$

If $z \in D_c$, the Cauchy transform defines an analytic function $C(z,t)$, say. Assume D_c contains the origin. Then $C(z,t)$ has a Taylor expansion

$$C(z,t) = \sum_{n=0}^{\infty} M_n z^n, \tag{130}$$

where

$$M_n = \frac{1}{2\pi i} \oint_{\partial D} \frac{\bar{z}' dz'}{z'^{n+1}}. \tag{131}$$

Suppose now that $C(z,t)$ is known. This is equivalent to knowledge of the moments $M_n, n = 0, 1, \ldots$, from which it is possible to infer the values of the harmonic moments associated with the set of functions

$$\left\{ \mathrm{Re}\left[\frac{1}{z^{n+1}}\right],\ \mathrm{Im}\left[\frac{1}{z^{n+1}}\right] \middle| n = 0, 1, , \ldots \right\}. \tag{132}$$

This is a subspace of codimension one in the space of functions harmonic in $D(t)$ because it excludes the constant function 1, which is also harmonic in $D(t)$. Thus, knowledge of $C(z,t)$ is not quite enough to determine all harmonic moments—just one more moment is needed. Generalizing the set (131), the "moment" corresponding to the constant function 1 is

$$\frac{1}{2\pi i} \oint_{\partial D} \bar{z}' dz', \tag{133}$$

which is proportional to the area of the bubble. Thus, in the case of an unbounded simply connected domain $D(t)$, the Cauchy transform $C(z,t)$ determines the domain up to a single real degree of freedom associated with the area of the complement of $D(t)$ (here, the area of the bubble).

Acknowledgments. The authors wish to thank B. Gustafsson and S. Tanveer for useful discussions.

REFERENCES

[1] M. Ablowitz and A. S. Fokas, *Complex Variables*, Cambridge University Press, London, 1997.
[2] L. K. Antanovskii, *Quasi-steady deformation of a two-dimensional bubble placed within a potential viscous flow*, Meccanica, 29 (1994), pp. 27–42.
[3] L. K. Antanovskii, *Formation of a pointed drop in Taylor's four-roller mill*, J. Fluid Mech., 327 (1996), pp. 325–341.
[4] L. K. Antanovskii, *A plane inviscid incompressible bubble placed within a creeping viscous flow: Formation of a cusped bubble*, European J. Mech. B. Fluids, 13 (1994), pp. 491–509.
[5] D. Crowdy, *On a class of geometry-driven free boundary problems*, SIAM J. Appl. Math., 62 (2002), pp. 945–964.
[6] D. G. Crowdy and S. Tanveer, *A theory of exact solutions for plane viscous blobs*, J. Nonlinear Sci., 8 (1998), pp. 261–279.
[7] D. Crowdy, *Compressible bubbles in Stokes flow*, J. Fluid Mech., 476 (2003), pp. 345–356.
[8] D. Crowdy and J. Marshall, *Constructing multiply connected quadrature domains*, SIAM J. Appl. Math., 64 (2004), pp. 1334–1359.
[9] L. J. Cummings and P. D. Howell, *On the evolution of non-axisymmetric viscous fibres with surface tension, inertia, and gravity*, J. Fluid Mech., 389 (1999), pp. 361–389.
[10] L. J. Cummings, S. D. Howison, and J. R. King, *Two-dimensional Stokes and Hele-Shaw flows with free surfaces*, European J. Appl. Math., 10 (1999), pp. 635–680.

[11] P. J. Davis, *The Schwarz Function and Its Applications*, Carus Math. Monogr., The Mathematical Association of America, Buffalo, NY, 1974.
[12] V. M. Entov, P. I. Etingof, and D. Ya. Kleinbock, *On nonlinear interface dynamics in Hele-Shaw flows*, European J. Appl. Math., 6 (1995), pp. 399–420.
[13] E. Hille, *Ordinary Differential Equations in the Complex Plane*, Wiley-Interscience, New York, 1976.
[14] S. D. Howison and S. Richardson, *Cusp development in free boundaries, and two-dimensional slow viscous flows*, European J. Appl. Math., 6 (1995), pp. 441–454.
[15] S. G. Mikhlin, *Integral Equations*, Pergamon Press, New York, 1957.
[16] C. Pozrikidis, *Expansion of a compressible bubble in Stokes flow*, J. Fluid Mech., 442 (2001), pp. 171–189.
[17] S. Richardson, *Two-dimensional slow viscous flows with time-dependent free boundaries driven by surface tension*, European J. Appl. Math., 3 (1992), pp. 193–207.
[18] S. Richardson, *Two-dimensional bubbles in slow viscous flow*, J. Fluid Mech., 33 (1968), pp. 476–493.
[19] S. Richardson, *Two-dimensional bubbles in slow viscous flows. Part* 2, J. Fluid Mech., 58 (1973), pp. 115–127.
[20] S. Richardson, *Hele-Shaw flows with time-dependent free boundaries involving injection through slits*, Stud. Appl. Math., 87 (1992), pp. 175–194.
[21] S. Richardson, *Some Hele-Shaw flows with time-dependent free boundaries*, J. Fluid Mech., 102 (1981), pp. 263–278.
[22] M. Siegel, *Cusp formation for time-evolving bubbles in two-dimensional Stokes flow*, J. Fluid Mech., 412 (2000), pp. 227–257.
[23] S. Tanveer and G. L. Vasconcelos, *Time-evolving bubbles in two-dimensional Stokes flow*, J. Fluid Mech., 301 (1995), pp. 325–344.
[24] A. Varchenko and P. I. Etingof, *Why the Boundary of a Round Drop Becomes a Curve of Order Four*, Univ. Lecture Ser. 3, AMS, Providence, RI, 1992.

SIAM J. APPL. MATH.
Vol. 65, No. 3, pp. 964–982

DYNAMICS OF TWO-STRAIN INFLUENZA WITH ISOLATION AND PARTIAL CROSS-IMMUNITY*

M. NUÑO†, Z. FENG‡, M. MARTCHEVA§, AND C. CASTILLO-CHAVEZ¶

Abstract. The time evolution of the influenza A virus is linked to a nonfixed landscape driven by interactions between hosts and competing influenza strains. Herd-immunity, cross-immunity, and age-structure are among the factors that have been shown to support strain coexistence and/or disease oscillations. In this study, we put two influenza strains under various levels of (interference) competition. We establish that cross-immunity and host isolation lead to periodic epidemic outbreaks (sustained oscillations) in this multistrain system. We compute the isolation reproductive number for each strain ($\Re_i$) independently, as well as for the full system ($\Re_q$), and show that when $\Re_q < 1$, both strains die out. Subthreshold coexistence driven by cross-immunity is possible even when the isolation reproductive number of one strain is below 1. Conditions that guarantee a winning type or coexistence are established in general. Oscillatory coexistence is established via Hopf bifurcation theory and confirmed via numerical simulations.

Key words. influenza, multiple strains, cross-immunity, isolation, stability, bifurcation, oscillations, coexistence

AMS subject classifications. 92D30, 92D25, 34C25, 34C60

DOI. 10.1137/S003613990343882X

1. Introduction.

Several studies have focused on the identification of mechanisms capable of supporting multiple-strain coexistence for diseases that provide permanent or temporary immunity [19, 18]. Although there is still limited understanding on the role of cross-immunity (form of interference competition) between strains of a given virus, host variability (behavioral and immunological) is known to play a key role in maintaining virus diversity. Influenza epidemics and pandemics are closely linked to two types of mechanisms that maintain viral genetic diversity: antigenic "drift," the driver of strain heterogeneity, and antigenic "shift," the generator of subtype variability [28].

In 1918, the "Spanish Flu" pandemic caused the largest number of flu-related deaths worldwide in a single season [28]. More than 500,000 people died in the United States with 20–50 millions deaths worldwide. The "Asian Flu," a result of an antigenic shift in the hemmaglutinin and neuraminidase surface proteins, was responsible for about 70,000 deaths in the United States in 1969 [9]. The most recent and least lethal "pandemic," the "Hong Kong" pandemic, is attributed to the appearance of the H3N2 subtype [9].

*Received by the editors December 17, 2003; accepted for publication (in revised form) August 11, 2004; published electronically March 31, 2005. The research of the second and third authors was supported by NSF grants DMS-0314575, DMS-0137687, and DMS-0406119. The research of the third and fourth authors was supported with grants directed toward the Mathematical and Theoretical Biological Institute (MTBI) by the following institutions: National Science Foundation (NSF), National Security Agency (NSA), and Alfred P. Sloan Foundation.

http://www.siam.org/journals/siap/65-3/43882.html

†Department of Biological Statistics and Computational Biology, Cornell University, Warren Hall, Ithaca, NY 14853-7801 (man16@cornell.edu).

‡Department of Mathematics, Purdue University, West Lafayette, IN 47907.

§Department of Mathematics, University of Florida, Gainesville, FL 32611-8105.

¶Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287-1804.

The main focus of this paper is on the identification of competitive outcomes (mediated by cross-immunity) that result from the interactions between two strains of influenza A in a population where sick individuals may be isolated. Single-strain susceptible-infected-quarantined-recovered (SIQR) models with vital dynamics can generate sustained oscillations [15, 20]. The introduction of a second strain increases the competition for susceptibles, a process mediated by cross-immunity in our setting. Will such competition preclude the possibility of sustained multistrain oscillations? We show that coexistence of both strains in the oscillatory regime is not uncommon and that oscillatory dynamics are possible for reasonable values of influenza parameters [17, 11, 28].

Our paper is structured as follows. Section 2 introduces the general two-strain model; section 3 carries out the local stability analysis of the disease-free state; section 4 shows that periodic solutions can arise via a Hopf bifurcation; section 5 illustrates our theoretical results; section 6 summarizes our findings and collects some conclusions and thoughts.

2. Two-strain model. Theoretical work on two-strain models that incorporate the effects of interference competition in the context of communicable diseases goes back (at least) to the work of Dietz [13]. His work has been extended in the context of influenza [5, 1, 6]. None of these extensions considered the role of isolation. The study of mechanisms capable of generating sustained oscillations in single-strain epidemic models has received some attention in the last decades [19, 18]. Feng [14] and Feng and Thieme [15] showed that the introduction of an isolation class, in an otherwise standard SIR epidemiological model, is enough to generate sustained oscillations in single-strain models, but the region of parameter space where such oscillations are possible is unrealistic. Castillo-Chavez et al. [5, 6] provide support for the hypothesis that age-structure (age-dependent survival) and cross-immunity are enough to generate multistrain sustained oscillations in two-strain models without isolation. Here we show that cross-immunity in a two-strain system with isolation classes generate sustained oscillations within a region of parameter space that is consistent with the "flu" [11, 29, 24]. Furthermore, we identify the dependence of these regions on cross-immunity levels. The description of the two-strain model requires the division of the population into ten different classes: susceptibles (S), infected with strain i (I_i, primary infection), isolated with strain i (Q_i), recovered from strain i (R_i, as a result of primary infection), infected with strain i (V_i, secondary infection), given that the population had recovered from strains $j \neq i$, and recovered from both strains (W). The population is assumed to mix randomly, except that the mixing is impacted by the process of quarantine/isolation [14, 15, 20, 8]. Using the flow diagram in Figure 1, we arrive at the model

$$
\begin{aligned}
\frac{dS}{dt} &= \Lambda - \sum_{i=1}^{2} \beta_i S \frac{(I_i + V_i)}{A} - \mu S, \\
\frac{dI_i}{dt} &= \beta_i S \frac{(I_i + V_i)}{A} - (\mu + \gamma_i + \delta_i) I_i, \\
\frac{dQ_i}{dt} &= \delta_i I_i - (\mu + \alpha_i) Q_i, \\
\frac{dR_i}{dt} &= \gamma_i I_i + \alpha_i Q_i - \beta_j \sigma_{ij} R_i \frac{(I_j + V_j)}{A} - \mu R_i, \qquad j \neq i,
\end{aligned} \tag{1}
$$

$$\frac{dV_i}{dt} = \beta_i \sigma_{ij} R_j \frac{(I_i + V_i)}{A} - (\mu + \gamma_i) V_i, \qquad j \neq i$$

$$\frac{dW}{dt} = \sum_{i=1}^{2} \gamma_i V_i - \mu W,$$

$$A = S + W + \sum_{i=1}^{2} (I_i + V_i + R_i),$$

where A denotes the population of nonisolated individuals and $\frac{\beta_i S(I_i+V_i)}{A}$ models the rate at which susceptibles become infected with strain i. That is, the ith $(i \neq j)$ incidence rate is assumed to be proportional to both the number of susceptibles and the available proportion of i-infectious individuals, $\frac{(I_i+V_i)}{A}$. The parameter σ_{ij} is a measure of the cross-immunity provided by a prior infection with strain i to exposure with strain j $(i \neq j)$. Data from epidemiological studies conducted in Houston and Seattle [27, 17] generate rough measures of cross-immunity. From these studies it is clear that $\sigma_{ij} \in [0, 1]$. Model (1) includes the models in [5, 6]. The absence of the Q classes in earlier work precludes the possibility of sustained oscillations (see [5, 6]). Isolation classes are not introduced after the V-classes to simplify the analysis and because often symptoms are less severe in these classes.

3. Disease invasion and stability. System (1) can support four equilibria. Analysis of the local stability of the trivial equilibrium (absence of disease) helps identify conditions under which the "flu" can invade. We assume (sections 3 and 4)

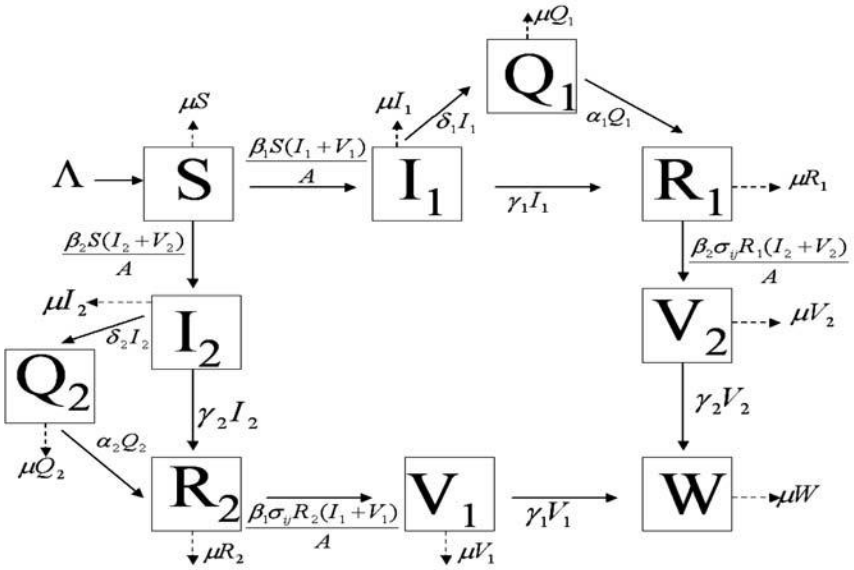


FIG. 1. *Schematic diagram of disease dynamics when the host is exposed to two cocirculating strains.* Λ *is the rate at which individuals are born into the population,* β_i *denotes the transmission coefficient for strain* i, μ *is the per capita mortality rate,* δ_i *is the per capita isolation rate for strain* i, γ_i *denotes the per capita recovery rate from strain* i, α_i *is the per capita rate at which individuals leave the isolated class as a result of infection with strain* i, *and* σ_{ij} *is the relative susceptibility to strain* j *for an individual previously infected with and recovered from strain* i $(i \neq j)$. $\sigma_{ij} = 0$ *corresponds to total cross-immunity, while* $\sigma_{ij} = 1$ *indicates no cross-immunity.*

that $\sigma_{12} = \sigma_{21} = \sigma$ (as it was done in [5, 6]). This approach strongly limits the generality of our analysis, but the general case has turned out to be too difficult. The case where $|\sigma_{12} - \sigma_{21}| = \varepsilon$ (a small positive number) is explored numerically. Quantitative results do not seem to change when ε is small enough. Flu-related mortality is low, and hence it is ignored. This is not a limiting factor over appropriate time scales. From our model (time scale for which demographic factors can be ignored)

$$\frac{d}{dt}N = \Lambda - \mu N,$$

where $N = S + W + \sum_{i=1}^{2}(I_i + V_i + Q_i + R_i)$. Hence, $N(t) \to \frac{\Lambda}{\mu}$ as $t \to \infty$, and the results in [7] allow us to assume, without loss of generality, that $N(0) = \frac{\Lambda}{\mu}$. Hence, we set

$$N(t) \equiv \frac{\Lambda}{\mu} \equiv S + W + \sum_{i=1}^{2}(I_i + V_i + Q_i + R_i) = A + Q \qquad \text{for all } t,$$

where $Q = Q_1 + Q_2$ and $A = N - Q$.

The *isolation* reproductive number $\Re_q$, the average number of secondary infections generated by the simultaneous introduction of both strains in a fully susceptible population, is a function of the independent capacity of each strain to invade. Hence, $\Re_q = \max\{\Re_1, \Re_2\}$, where

$$\Re_i = \frac{\beta_i}{\mu + \gamma_i + \delta_i}.$$

Here, β_i is the maximal effective transmission rate and $(\mu + \gamma_i + \delta_i)^{-1}$ is the average window of opportunity (effective infectious period) for transmission. It follows that E_0, the disease-free state, is locally asymptotically stable when $\Re_q < 1$ and an unstable (saddle) whenever $\Re_i > 1$ for either $i = 1$ or $i = 2$ (for details see Appendix A).

4. Nontrivial equilibria and sustained oscillations. Hethcote and Levin's 1989 survey of mathematical models [18], Feng [14], Feng and Thieme [15], Hethcote [19], and the recent comprehensive literature review of Hethcote and Levin [18] provide a solid perspective on what is known about the mechanisms that are capable of supporting sustained oscillations in epidemic models. Nonstructured two-strain SIR models with cross-immunity appear to be incapable of supporting them [1], but the addition of a third strain reverses the situation [26].

Here we focus on the role of isolation, a mechanism capable of generating sustained oscillations even in a single-strain model. The "flu" may survive in three states: either strain 1 or 2 survives or both strains coexist. Here, we carry only out the analysis in the symmetric cross-immunity case ($\sigma_{12} = \sigma_{21} = \sigma$). We let $U = (S, I_1, Q_1, R_1, V_1, I_2, Q_2, R_2, V_2, W)$ denote the state variables and focus on the analysis of the stability of the boundary equilibria, namely $E_1 = (\tilde{S}, \tilde{I}_1, \tilde{Q}_1, \tilde{R}_1, 0, 0, 0, 0, 0, 0)$. Setting $V_1 = I_2 = Q_2 = R_2 = V_2 = W = 0$ in (1) leads to the following relationships:

$$\begin{aligned} &\frac{\tilde{S}}{\tilde{A}} = \frac{1}{\Re_1}, \qquad &&\frac{\tilde{I}_1}{\tilde{A}} = \mu(\mu + \alpha_1)\phi, \\ &\frac{\tilde{Q}_1}{\tilde{A}} = \mu\delta_1\phi, \qquad &&\frac{\tilde{R}_1}{\tilde{A}} = (\gamma_1(\mu + \alpha_1) + \alpha_1\delta_1)\,\phi, \end{aligned} \tag{2}$$

where

$$\phi = \frac{(1 - \frac{1}{\Re_1})}{(\mu + \gamma_1)(\mu + \alpha_1) + \alpha_1 \delta_1} \tag{3}$$

and

$$\tilde{A} = \frac{1}{\mu(1 + \mu \delta_1 \phi)}.$$

E_1 exists (entries are positive) and is unique if and only if $\Re_1 > 1$. Letting $A = N - Q$ and $S = A - \sum_{i=1}^{2}(I_i + V_i + R_i) - W$ allows the elimination of the S equation. The I_i equations become

$$\frac{dI_i}{dt} = \beta_i \left(1 - \frac{W + \sum_{i=1}^{2}(I_i + R_i + V_i)}{A}\right)(I_i + V_i) - (\mu + \gamma_i + \delta_i) I_i.$$

The Jacobian at E_1, $\tilde{J}$, is given by the 9×9 (without S) matrix

$$\tilde{J} = \begin{pmatrix} G_1 & * & * & * \\ 0 & -(\mu + \gamma_1) & * & 0 \\ 0 & 0 & G_2 & 0 \\ 0 & * & * & -\mu \end{pmatrix},$$

where

$$G_1 = \begin{pmatrix} -\beta_1 \frac{\tilde{I}_1}{\tilde{A}} & -\beta_1 \frac{\tilde{I}_1}{\tilde{A}}(1 - \frac{1}{\Re_1}) & -\beta_1 \frac{\tilde{I}_1}{\tilde{A}} \\ \delta_1 & -(\mu + \alpha_1) & 0 \\ \gamma_1 & \alpha_1 & -\mu \end{pmatrix},$$

and

$$G_2 = \begin{pmatrix} \beta_2 \frac{\tilde{S}}{\tilde{A}} - (\mu + \gamma_2 + \delta_2) & 0 & 0 & \beta_2 \frac{\tilde{S}}{\tilde{A}} \\ \delta_2 & -(\mu + \alpha_2) & 0 & 0 \\ \gamma_2 & \alpha_2 & -\mu - \beta_1 \sigma \frac{\tilde{I}_1}{\tilde{A}} & 0 \\ \beta_2 \sigma \frac{\tilde{R}_1}{\tilde{A}} & 0 & 0 & \beta_2 \sigma \frac{\tilde{R}_1}{\tilde{A}} - (\mu + \gamma_2) \end{pmatrix}.$$

"*" represents a nonzero block matrix.

G_2 has two negative eigenvalues, plus the roots of the equation

$$\lambda^2 - c_1 \lambda + c_2 = 0, \tag{4}$$

where

$$\begin{aligned} c_1 &= (\mu + \gamma_2 + \delta_2)\left(\frac{\Re_2 - \Re_1}{\Re_1}\right) + \beta_2 \sigma \frac{\tilde{R}_1}{\tilde{A}} - (\mu + \gamma_2), \\ c_2 &= -(\mu + \gamma_2 + \delta_2)\left[\beta_2 \sigma \frac{\tilde{R}_1}{\tilde{A}} + (\mu + \gamma_2)\left(\frac{\Re_2 - \Re_1}{\Re_1}\right)\right]. \end{aligned} \tag{5}$$

Hence $c_1 < 0$ and $c_2 > 0$ guarantee the local asymptotic stability (l.a.s.) of E_1.

$$c_1 < 0 \Longleftrightarrow F_1(\Re_1, \Re_2) := (\mu + \gamma_2 + \delta_2)\left(\frac{\Re_2}{\Re_1} - 1 + \sigma \Re_2 \frac{\tilde{R}_1}{\tilde{A}}\right) - (\mu + \gamma_2) < 0, \tag{6}$$

$$c_2 > 0 \Longleftrightarrow F_2(\Re_1, \Re_2) := \sigma(\mu + \gamma_2 + \delta_2)\Re_2 \frac{\tilde{R}_1}{\tilde{A}} + (\mu + \gamma_2)\left(\frac{\Re_2}{\Re_1} - 1\right) < 0, \tag{7}$$

where $\tilde{R}_1/\tilde{A}$ is given in (2). In the case of full immunity ($\sigma = 0$), the conditions in (6) and (7) hold if and only if $\Re_2 < \Re_1$. That is, when $\sigma = 0$, E_1 has l.a.s., as long as $\Re_1 > 1$ and $\frac{\Re_1}{\Re_2} > 1$ and E_1 is unstable ($\sigma = 0$) when $\frac{\Re_2}{\Re_1} > 1$. As cross-immunity between strains diminishes ($0 < \sigma \uparrow 1$), alternative conditions are needed to ensure that (6) and (7) hold. To find these conditions we rewrite F_1 in terms of F_2,

$$F_1(\Re_1, \Re_2) = F_2(\Re_1, \Re_2) + \delta_2\left(\frac{\Re_2}{\Re_1} - 1\right) - (\mu + \gamma_2),$$

and observe that $F_1 \leq F_2$ when $\Re_2 < \Re_1$. Alternatively, the introduction of

$$f(\Re_1) \equiv \frac{\Re_1}{1 + \sigma(\Re_1 - 1)\left(1 + \frac{\delta_2}{\mu+\gamma_2}\right)\left(1 - \frac{\mu(\mu+\alpha_1)}{(\mu+\gamma_1)(\mu+\alpha_1)+\alpha_1\delta_1}\right)} \tag{8}$$

implies that $F_2 < 0$ if and only if $\Re_2 < f(\Re_1)$ ($0 < f(\Re_1) < \Re_1$). Therefore, $\Re_2 < f(\Re_1)$ implies that $F_1 \leq F_2 < 0$. Hence, all eigenvalues of G_2 have negative real part when $\frac{\Re_2}{f(\Re_1)} < 1$, and E_1 is unstable when $\frac{\Re_2}{f(\Re_1)} > 1$. Similarly, the use of

$$g(\Re_2) \equiv \frac{\Re_2}{1 + \sigma(\Re_2 - 1)\left(1 + \frac{\delta_1}{\mu+\gamma_1}\right)\left(1 - \frac{\mu(\mu+\alpha_2)}{(\mu+\gamma_2)(\mu+\alpha_2)+\alpha_2\delta_2}\right)} \tag{9}$$

implies that the eigenvalues corresponding to the system when strain 2 has become established are all negative whenever $\Re_1 < g(\Re_2)$. The boundary endemic equilibria for strain 2 (E_2) are stable when $\frac{\Re_1}{g(\Re_2)} < 1$ and unstable when $\frac{\Re_1}{g(\Re_2)} > 1$.

Conditions that guarantee "coexistence" equilibria are formulated in terms of the (conditional) "invasion" reproductive numbers for strains 2 and 1 ($\Re_2^1$ and $\Re_1^2$). $\Re_2^1$ is defined as the number of secondary infections generated by a "typical" strain-2-infected individual in a population where strain 1 is endemic (E_1). From conditions (6) and (7) (which ensure the stability of E_1) we find that

$$\Re_2^1 = \frac{\beta_2}{\mu + \gamma_2 + \delta_2}\frac{\tilde{S}}{\tilde{A}} + \frac{\beta_2\sigma}{\mu + \gamma_2}\frac{\tilde{R}_1}{\tilde{A}}.$$

Similarly, the (conditional) invasion reproductive number of strain 1 under the assumption that strain 2 is endemic is given by

$$\Re_1^2 = \frac{\beta_1}{\mu + \gamma_1 + \delta_1}\frac{\tilde{S}}{\tilde{A}} + \frac{\beta_1\sigma}{\mu + \gamma_1}\frac{\tilde{R}_2}{\tilde{A}}.$$

The condition $\Re_2 < f(\Re_1)$ is equivalent to the condition $\Re_2^1 < 1$, while the condition $\Re_1 < g(\Re_2)$ is equivalent to the condition $\Re_1^2 < 1$. $\Re_i^j$ $(i, j = 1, 2\ i \neq j)$ is in fact the result of two additive contributions: $\beta_i/(\mu + \gamma_i + \delta_i)$ gives the number of secondary cases that a "typical" strain-i-infected individual will generate in the fully susceptible proportion of the population $\tilde{S}/\tilde{A}$, while $\beta_i\sigma/(\mu + \gamma_i)$ is the number of secondary cases that a "typical" strain-i-infected individual will generate in the "cross-immune" proportion of the susceptible population $\tilde{R}_i/\tilde{A}$. Note that whenever $\Re_i > 1$ and $\Re_j^i < 1$, the boundary equilibrium E_i is locally stable.

4.1. Multiple and subthreshold coexistence. Both strains coexist if their basic reproductive numbers are above 1 (see Figures 2(a), (b), (d)), but subthreshold coexistence is possible (see Figure 2(c)). In order to see this (in system (1)) let $S/A = s$, $I_i/A = i_i$, $Q_i/A = q_i$, $R_i/A = r_i$, $V_i/A = v_i$, $W/A = w$, and $n = \frac{N}{A}$ with $\sigma_{12} = \sigma_{21} = \sigma$. The equilibrium conditions for the rescaled system are

$$\beta_1 s(i_1 + v_1) + \beta_2 s \frac{\Re_1}{\Re_2} \frac{i_2}{i_1}(i_1 + v_1) + \mu s = \mu(1 + \eta_1 i_1 + \eta_2 i_2), \tag{10}$$

$$\beta_1 s(i_1 + v_1) = (\mu + \gamma_1 + \delta_1) i_1, \tag{11}$$

$$\frac{(i_1 + v_1)}{(i_2 + v_2)} = \frac{\Re_2}{\Re_1} \frac{i_1}{i_2}, \tag{12}$$

$$\beta_2 \sigma r_1 \frac{\Re_1}{\Re_2} \frac{i_2}{i_1}(i_1 + v_1) + \mu r_1 = (\gamma_1 + \kappa_1) i_1, \tag{13}$$

$$\beta_1 \sigma r_2 (i_1 + v_1) = (\mu + \gamma_1) v_1, \tag{14}$$

$$\beta_1 \sigma r_2 (i_1 + v_1) + \mu r_2 = (\gamma_2 + \kappa_2) i_2, \tag{15}$$

$$\beta_2 \sigma r_1 \frac{\Re_1}{\Re_2} \frac{i_2}{i_1}(i_1 + v_1) + (\mu + \gamma_2) i_2 = (\mu + \gamma_2)(i_2 + v_2), \tag{16}$$

where

$$\kappa_i = \frac{\alpha_i \delta_i}{\mu + \alpha_i},$$

$$\eta_i = \frac{\delta_i}{\mu + \alpha_i}.$$

Expressions (10) and (11) can be solved for s, that is,

$$s = \frac{\mu(1 + \eta_1 i_1 + \eta_2 i_2)}{\beta_1(i_1 + v_1) + \beta_2(i_2 + v_2) + \mu} = \frac{(\mu + \gamma_1 + \delta_1) i_1}{\beta_1(i_1 + v_1)}. \tag{17}$$

From (17) it follows that

$$\frac{\mu \beta_1 (i_1 + v_1)(1 + \eta_1 i_1 + \eta_2 i_2)}{\beta_1(i_1 + v_1) + \beta_2(i_2 + v_2) + \mu} = (\mu + \gamma_1 + \delta_1) i_1. \tag{18}$$

Equation (14) and its symmetric analogue are solved for r_1 and r_2. In fact,

$$r_1 = \frac{(\mu + \gamma_2) v_2}{\beta_2 \sigma (i_2 + v_2)}$$

and

$$r_2 = \frac{(\mu + \gamma_1) v_1}{\beta_1 \sigma (i_1 + v_1)}.$$

From (13) and the relationship $\beta_2 \sigma (i_2 + v_2) r_1 = (\mu + \gamma_2) v_2$ we have that

$$(\gamma_1 + \kappa_1) i_1 - (\mu + \gamma_2) v_2 = \mu r_1.$$

Using (12) and substituting the above expression for r_1 helps rewrite (13) as

$$(\gamma_1 + \kappa_1) i_1 - (\mu + \gamma_2) \frac{\Re_1 i_2 (i_1 + v_1)}{\Re_2 i_1} + (\mu + \gamma_2) i_2 = \frac{\mu(\mu + \gamma_2)}{\beta_2 \sigma} - \frac{\mu(\mu + \gamma_2) i_2}{\beta_2 \sigma (i_2 + v_2)}. \tag{19}$$

Solving for $(i_2 + v_2)$ in (13) and using (19) leads to

(20)
$$(\gamma_1 + \kappa_1)i_1 - (\mu + \gamma_2)\frac{\Re_1 i_2(i_1 + v_1)}{\Re_2 i_1} + (\mu + \gamma_2)i_2 = \frac{\mu(\mu + \gamma_2)}{\beta_2\sigma}\left(1 - \frac{\Re_2 i_1}{\Re_1(i_1 + v_1)}\right).$$

In a similar manner, using r_2 and the expression $\beta_1\sigma r_2(i_1 + v_1) = (\mu + \gamma_1)v_1$ leads to the reduced system

$$\frac{\mu\beta_1(1 + \eta_1 i_1 + \eta_2 i_2)(i_1 + v_1)}{\beta_1(i_1 + v_1) + \frac{\beta_2\Re_1(i_1+v_1)i_2}{\Re_2 i_1} + \mu} = (\mu + \gamma_1 + \delta_1)i_1,$$

$$(\gamma_1 + \kappa_1)i_1 - (\mu + \gamma_2)i_2\left(\frac{\Re_1(i_1 + v_1)}{\Re_2 i_1} - 1\right) = \frac{\mu(\mu + \gamma_2)}{\beta_2\sigma}\left(1 - \frac{\Re_2 i_1}{\Re_1(i_1 + v_1)}\right), \tag{21}$$

$$(\gamma_2 + \kappa_2)i_2 - (\mu + \gamma_1)v_1 = \frac{\mu(\mu + \gamma_1)v_1}{\beta_1\sigma(i_1 + v_1)}.$$

From the first equation in (17) we get that

$$\begin{aligned}(i_1 + v_1) &= \frac{\mu(\mu + \gamma_1 + \delta_1)i_1}{\beta_1[\mu(1 + \eta_1 i_1 + \eta_2 i_2) - i_1(\mu + \gamma_1 + \delta_1) - i_2(\mu + \gamma_2 + \delta_2)]},\\ &= \frac{\mu(\mu + \gamma_1 + \delta_1)i_1}{\beta_1[\triangle(i_1, i_2)]},\end{aligned} \tag{22}$$

where

$$\triangle(i_1, i_2) = \mu(1 + \eta_1 i_1 + \eta_2 i_2) - i_1(\mu + \gamma_1 + \delta_1) - i_2(\mu + \gamma_2 + \delta_2).$$

The substitution of (22) into the second equation in (21) and the use of its symmetric analogue gives a system of equations (in terms of i_1 and i_2 only). The system is

$$\begin{aligned}&\big(\triangle(i_1, i_2)\big)^2\frac{(\mu + \gamma_2)\Re_2}{\beta_2\sigma} + \triangle(i_1, i_2)\left[(\gamma_1 + \kappa_1)i_1 + (\mu + \gamma_2)i_2 - \frac{\mu(\mu + \gamma_2)}{\beta_2\sigma}\right]\\ &\quad - \frac{\mu(\mu + \gamma_2)i_2}{\Re_2} = 0,\\ &\big(\triangle(i_1, i_2)\big)^2\frac{(\mu + \gamma_1)\Re_1}{\beta_1\sigma} + \triangle(i_1, i_2)\left[(\gamma_2 + \kappa_2)i_2 + (\mu + \gamma_1)i_1 - \frac{\mu(\mu + \gamma_1)}{\beta_1\sigma}\right]\\ &\quad - \frac{\mu(\mu + \gamma_1)i_1}{\Re_1} = 0.\end{aligned} \tag{23}$$

Positive solutions of (23) are only candidates for coexistence equilibria, as we must check that the corresponding values (s, q_i, r_i, v_i, and w) are positive. Numerical simulations show that such solutions exist in the ranges $0 \le i_1 \le 1$ and $0 \le i_2 \le 1$ for parameter values that are reasonable for the "flu." Figure 2(a) ($\Re_1 > 1$ and $\Re_2 > 1$) shows one such intersection in the positive quadrant. Subthreshold coexistence equilibrium is also possible for $\Re_1 < 1$ and $\Re_2 > 1$ (see Figure 2(c)). As we increase the basic reproductive number of both strains and allow varying levels of cross-immunity ($\sigma = 0.5$ and $\sigma = 0.6$), Figures 2(b) and 2(d) show that two intersections (in the positive quadrant) are possible. That is, multiple coexistence equilibria exist (it was verified that all classes are positive). This possibility is absent from prior influenza models. For the parameter values in Table 1, only a single coexistence equilibrium is biologically reasonable. Figures 2(a), (c) provide additional examples where coexistence and subthreshold coexistence are possible.

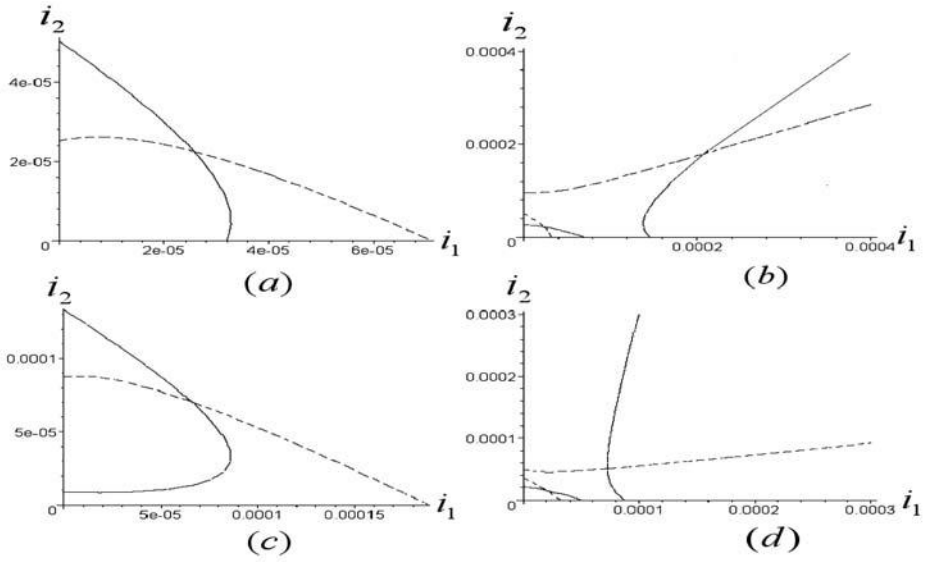


FIG. 2. *Positive solutions of* (23) *are provided to illustrate the existence of multiple endemic states and subthreshold coexistence (section* 4.1*). The horizontal axis is the fraction* ($i_1 = I_1/A$) *of individuals infected with strain* 1*, and the vertical axis depicts the fraction* ($i_2 = I_2/A$) *of individuals infected with strain* 2. (a) $\Re_1 = 2, \Re_2 = 1.99$*, and* $\sigma = 0.9$. (b) $\Re_1 = 2, \Re_2 = 1.99$*, and* $\sigma = 0.5$. (c) $\Re_1 = 0.75, \Re_2 = 2.9$*, and* $\sigma = 0.9$. (d) $\Re_1 = 2.83, \Re_2 = 2.93$*, and* $\sigma = 0.6$.

TABLE 1

The parameter values used for the numerical simulations are provided here. The initial conditions are given by $s(0) = 0.4$, $i_1(0) = 0.199$, $r_1(0) = r_2(0) = 0.2$, *and* $i_2(0) = q_1(0) = q_2 = w(0) = 0$ *(only one strain present initially) and* $s(0) = 0.4$, $i_1(0) = 0.199$, $q_1(0) = 0.1$, $r_1(0) = r_2(0) = 0.2$, $i_2(0) = 0.001$, *and* $q_1(0) = q_2(0) = w(0) = 0$ *(both strains present initially).*

Parameters	Definition	Values
$\Re_i$	Number of secondary cases generated by a primary case infected with strain i	(0.75, 4.5)
σ_{ij}	Cross-immunity against strain j following an infection with strain i	(0.008, 0.8)
Λ	Rate at which individuals are born into the population	0.00004
α_i	Rate at which individuals leave isolation ($1/\alpha_i$ = days)	(1, 15)
δ_i	Rate of isolation with strain i, $i = 1, 2$ ($1/\delta_i$ = days)	(1, 6)
γ_i	Recovery rate from strain i, $i = 1, 2$ ($1/\gamma_i$ = days)	(5, 7)
β_i	Transmission coefficient for strain i, $i = 1, 2$	(0.4, 2.2)
μ	Mortality rate	0.00004

4.2. Sustained oscillations. A detailed study of the nature of the eigenvalues of matrix G_1 makes use of the identity

$$\beta_1 \frac{\tilde{S}}{\tilde{A}} - (\mu + \gamma_1 + \alpha_1) = -\beta_1 \frac{\tilde{I}_1}{\tilde{A}}$$

and the fact that $\Re_1$ does not depend on α_1. The dependence of f on α_1 is in the order of μ, and this observation is used in the study of the characteristic equation associated with G_1 as we search for the possibility of sustained oscillations. The characteristic

equation is

$$\omega^3 + a_1\omega^2 + a_2\omega + a_3 = 0, \tag{24}$$

where

$$\begin{aligned}
a_1 &= 2\mu + \alpha_1 + \Re_1(\mu + \gamma_1 + \delta_1)\mu(\mu + \alpha_1)\phi,\\
a_2 &= \mu(\mu + \alpha_1)\left[1 + \Re_1(\mu + \gamma_1 + \delta_1)\left(2\mu + \alpha_1 + \gamma_1 + \delta_1\left(1 - \frac{1}{\Re_1}\right)\right)\phi\right],\\
a_3 &= \left[\mu^2 + \alpha_1\mu + \delta_1\alpha_1 + \gamma_1\mu + \gamma_1\alpha_1 + \delta_1\mu\left(1 - \frac{1}{\Re_1}\right)\right]\Re_1(\mu + \gamma_1 + \delta_1)\mu(\mu + \alpha_1)\phi.
\end{aligned} \tag{25}$$

Since a_1, a_2, and a_3 are all positive ($\Re_1 > 1$), then the cubic equation in (24) has either three negative or one negative root and possibly two complex conjugate roots. Differences in epidemiological and demographic time scales are used to tease out the nature of the roots of (24). The average life expectancy ($1/\mu$) is in the order of decades, while the infective ($1/\delta_i$ or $1/\gamma_i$) and isolation periods ($1/\alpha_i$) are just a few days. That is, μ is much smaller than δ_i, γ_i, and α_i. Following early approaches [14, 15, 23], we carry out an asymptotic expansion on the coefficients of (24) using μ. From (25), it is clear that a_i are analytic functions of $\mu > -\varepsilon$ for some $\varepsilon > 0$. Hence,

$$\begin{aligned}
a_1 &= \alpha_1 + (\Re_1^* + 1)\mu + O(\mu^2),\\
a_2 &= \left[(\alpha_1 + \gamma_1)\Re_1^* - \gamma_1 + \frac{\delta_1(\Re_1^* - 1)^2}{\Re_1^*}\right]\mu + O(\mu^2),\\
a_3 &= \alpha_1(\delta_1 + \gamma_1)(\Re_1^* - 1)\mu + O(\mu^2),
\end{aligned} \tag{26}$$

where $\Re_1^*$ denotes $\Re_1(\mu)$ evaluated at $\mu = 0$, that is, $\Re_1^* = \Re_1(0)$. The continuous dependence of the roots on μ is acknowledged by letting $\omega_i = \omega_i(\mu)$ ($i = 1, 2, 3$) denote the roots of (24) (for a fixed value of μ). In the limiting case, $\mu = 0$, a_2 and a_3 are zero, while $a_1 = \alpha_1$ (from (26)). The characteristic polynomial in this limiting case is simply

$$\omega^3 + \alpha_1\omega^2 = 0,$$

which has the roots $\omega_1(0) = -\alpha_1$ and $\omega_2(0) = \omega_3(0) = 0$. Hence, by continuity, $\omega_1(\mu) = -\alpha_1 + O(\mu)$ is a negative real root of (24) for small $\mu > 0$. In order to use arguments similar to those found in [14, 15, 23] or in Kato [23, II, §1, section 2], it is assumed that the roots $\omega_2(\mu)$ and $\omega_3(\mu)$ have expansions of the form

$$\omega(\mu) = \sum_{j=1}^{\infty} \xi_j \nu^j, \qquad \nu = \mu^{\frac{1}{2}}. \tag{27}$$

The formal substitution of (27) into (24) (neglecting higher-order terms in ν) yields

$$\begin{aligned}
&\left[\xi_1^2\alpha_1 + \alpha_1(\delta_1 + \gamma_1)(\Re_1^* - 1)\right]\nu^2\\
&\quad + \left[\xi_1^3 + 2\xi_1\xi_2\alpha_1 + \left((\gamma_1 + \alpha_1)\Re_1^* - \gamma_1 + \frac{\delta_1(\Re_1^* - 1)^2}{\Re_1^*}\right)\xi_1\right]\nu^3 + O(\nu^4) = 0.
\end{aligned}$$

Hence,

$$\xi_1^2 = -(\gamma_1 + \delta_1)(\Re_1^* - 1) \quad \text{and} \quad \xi_2 = -\frac{1}{2\alpha_1}\left(\xi_1^2 - \gamma_1 + (\gamma_1 + \alpha_1)\Re_1^* + \frac{\delta_1(\Re_1^* - 1)^2}{\Re_1^*}\right).$$

From the fact that $\Re_1^* > 1$ we have that

$$\xi_1 = \pm i\sqrt{(\gamma_1+\delta_1)\left(\Re_1^*-1\right)} \quad \text{and} \quad \xi_2 = -\frac{1}{2\alpha_1}\left(\alpha_1\Re_1^* + \delta_1\left(\frac{1}{\Re_1^*}-1\right)\right). \tag{28}$$

That is, the three roots of (24) have expressions of the form

$$\omega_1(\nu) = -\alpha_1 + O(\nu^2) \tag{29}$$

and

$$\omega_{2,3}(\nu) = \pm i\left((\gamma_1+\delta_1)\left(\Re_1^*-1\right)\right)^{\frac{1}{2}}\nu - \frac{1}{2\alpha_1}\left(\alpha_1\Re_1^* + \delta_1\left(\frac{1}{\Re_1^*}-1\right)\right)\nu^2 + O(\nu^3). \tag{30}$$

We select α_1 as our implicit bifurcation parameter ($1/\alpha_1$ is the isolation period for strain 1). We observe that $\alpha_1 = \alpha_1(\nu)$ is a function of ν that satisfies the equation $\xi_2(\alpha_1(0)) = 0$. Hence

$$\alpha_1(0) = \frac{\delta_1}{\Re_1^*}\left(1 - \frac{1}{\Re_1^*}\right).$$

The use of functions $\omega_{2,3} = \omega_{2,3}(\alpha_1,\nu)$ and $H(\alpha_1,\nu) = \frac{1}{\nu^2}\,\mathbb{R}\,\omega_{2,3}(\alpha_1,\nu)$ (where $\mathbb{R}\omega_{2,3}(\alpha_1,\nu)$ denotes the real part of the roots of (24) as given in (30)) imply that $H(\alpha_1(0),0) = \xi_2(\alpha_1(0)) = 0$. The implicit function theorem guarantees the existence of a critical function $\alpha_c(\nu) = \frac{\delta_1}{\Re_1^*}(1-\frac{1}{\Re_1^*}) + O(\nu)$, such that $H(\alpha_c(\nu),\nu) = 0$ for small ν. Clearly, $\alpha_c(\nu) > 0$, as long as $\Re_1^* > 1$. Furthermore, since

$$\frac{\partial H}{\partial \alpha_1}(0) = -\frac{1}{2\delta_1}\frac{{\Re_1^*}^3}{(\Re_1^*-1)} < 0,$$

nonresonance holds [21], that is, as the frequency of strain 1 approaches that of strain 2 (or vice versa). Solutions remain bounded. The use of α_1 as a bifurcation parameter shows that the roots $\omega_{2,3}$ cross the imaginary axis from left to right whenever α_1 crosses α_c from right to left. That is, the crossing is transversal. Hence, a Hopf bifurcation occurs near the critical point $\alpha_c = \delta_1(\Re_1^*-1)/(\Re_1^*)^2$. We collect these results in the following theorem.

THEOREM 1. *There are two functions: $f(\Re_1)$ as defined in* (8), *and $\alpha_c(\mu)$ defined for small $\mu > 0$ by*

$$\alpha_c(\mu) = \frac{\delta_1}{\Re_1^*}\left(1 - \frac{1}{\Re_1^*}\right) + O\left(\mu^{\frac{1}{2}}\right),$$

with the following properties: (i) *The boundary endemic equilibrium E_1 is locally asymptotically stable if $\Re_2 < f(\Re_1)$ and $\alpha_1 < \alpha_c(\mu)$, and unstable if $\Re_2 > f(\Re_1)$ or $\alpha_1 > \alpha_c(\mu)$.* (ii) *When $\Re_2 < f(\Re_1)$, periodic solutions arise at $\alpha_1 = \alpha_c(\mu)$ via Hopf bifurcation for small enough $\mu > 0$. The period can be approximated by*

$$T = \frac{2\pi}{|\Im\omega_{2,3}|} \approx \frac{2\pi}{((\gamma_1+\delta_1)(\Re_1^*-1))^{\frac{1}{2}}\,\mu^{\frac{1}{2}}}$$

or (using (2)*) by*

$$T \approx \frac{2\pi}{(\gamma_1+\delta_1)^{\frac{1}{2}}\left(\frac{\hat{I}_1}{\hat{A}}\right)^{\frac{1}{2}}\mu^{\frac{1}{2}}},$$

where $\hat{I}_1/\hat{A}$ *denotes* $\beta_1\tilde{I}_1/\mu\tilde{A}$ *evaluated at* $\mu = 0$ *and* $|\Im\omega_{2,3}|$ *refers to the imaginary roots calculated in* (30).

The latter expression for T allows one to compare the period of this model with the quasi periods obtained from models which do not include an isolation class [5, 6]. Since we have focused on the symmetric case, an analogous result for the second boundary equilibrium E_2 can be stated immediately. That is, the boundary endemic equilibrium E_2 is locally asymptotically stable if $\Re_1 < g(\Re_2)$ and $\alpha_2 < \alpha_c(\mu)$. It becomes unstable if $\Re_1 > g(\Re_2)$ or $\alpha_2 > \alpha_c(\mu)$. A summary of the stability results as presented in Theorem 1 for strain 1 is obtained for strain 2 by replacing the parameter indices 1's with 2's and replacing $f(\Re_1)$ with $g(\Re_2)$. Functions $f(\Re_1)$ and $g(\Re_2)$ help in the characterization of the stability and coexistence regions for strains 1 and 2. Changes in the regions of stability for either a single or for both strains can be illustrated as the coefficients of cross-immunity are varied. For instance, from (8) we can compute the value of σ at which

$$f'(\Re_1) \equiv \frac{\partial f(\Re_1, \sigma)}{\partial \Re_1}\Big|_{\sigma_1^*} = 0, \tag{31}$$

namely

$$\sigma_1^* = \frac{1}{\left(1 + \frac{\delta_2}{\mu+\gamma_2}\right)\left(1 - \frac{\mu(\mu+\alpha_1)}{(\mu+\gamma_1)(\mu+\alpha_1)+\alpha_1\delta_1}\right)}.$$

Hence, for all $\Re_1 > 1$,

$$f'(\Re_1) > (<, =)\ 0, \qquad f(\Re_1) > (<, =)\ 1 \qquad \text{if } \sigma < (>, =)\ \sigma_1^*.$$

These properties are easily verified, since (from (8))

$$f(\Re_1) = \frac{\Re_1}{1 + \frac{\sigma}{\sigma_1^*}(\Re_1 - 1)} \quad \text{and} \quad f'(\Re_1) = \frac{1 - \frac{\sigma}{\sigma_1^*}}{\left(1 + \frac{\sigma}{\sigma_1^*}(\Re_1 - 1)\right)^2}.$$

From the facts that $f(\Re_1) < \Re_1$ and $f(1) = 1$ we see that Figure 3 captures the properties of the curve $\Re_2 = f(\Re_1)$. Similar curve "boundary" features can be studied using threshold value σ_2^* (interchanging the subscripts 1 and 2 in the expression of σ_1^*) and the function $\Re_1 = g(\Re_2)$ (also shown in Figure 3). The special case when both strains are identical, $\sigma_1^* = \sigma_2^* = \sigma^*$, is implicit in Figure 3. $\Re_2 < f(\Re_1)$ is a necessary condition for the stability of strain 1 (either a stable boundary endemic equilibrium E_1 or the equilibrium associated with strain-1 oscillations). Hence, E_1 is unstable when $\Re_2 > f(\Re_1)$. Similarly, E_2 is unstable when $\Re_1 > g(\Re_2)$. Hence, coexistence is expected when $\Re_2 > f(\Re_1)$ and $\Re_1 > g(\Re_2)$.

Next, the cases $\sigma_2^* < \sigma < \sigma_1^*$ and $\sigma_1^* < \sigma < \sigma_2^*$ are considered. $f(\Re_1)$ and $g(\Re_2)$ are increasing and decreasing functions of σ correspondingly for $\sigma_2^* < \sigma < \sigma_1^*$ and decreasing and increasing (respectively) for $\sigma_1^* < \sigma < \sigma_2^*$ (Figure 4(b)). Hence, the stability region for strain 1 (Figure 4(a), region I) may be significantly larger than that of strain 2 (Figure 4(a), region II) for $\sigma_2^* < \sigma < \sigma_1^*$. For $\sigma_1^* < \sigma < \sigma_2^*$, the stability region of strain 1 may be noticeably smaller than that of strain 2. The changes in the relative sizes of these stability regions seem to cause strong cross-immunity when it is conferred by strain i ($\sigma_i \downarrow 0$) against an infection with strain j (largely reduced susceptibility to alternative strain infections). The possibility that strain j

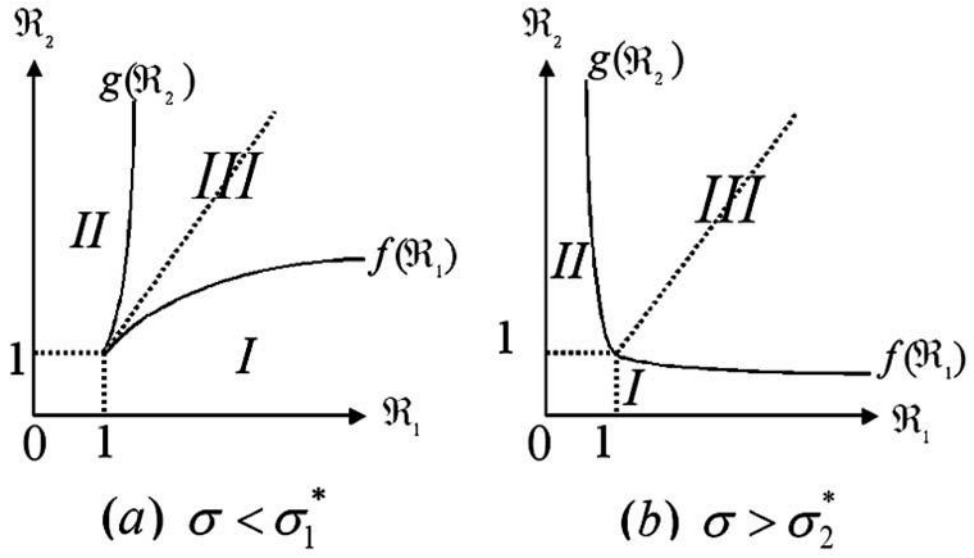


FIG. 3. *Bifurcation diagram in the* $(\Re_1, \Re_2)$*-plane. The curves* $\Re_2 = f(\Re_1)$ *(for* $\Re_1 > 1$*) and* $\Re_1 = g(\Re_2)$ *(for* $\Re_2 > 1$*) divide the region* $\mathbf{R}^2_+ - \{(\Re_1, \Re_2) \mid \Re_1 < 1$ *and* $\Re_2 < 1\}$ *into three subregions:* I, II, III*. When the parameters are in region* I *(*II*), only strain* 1 *(strain* 2*) will be maintained (a stable boundary equilibrium or sustained oscillations of a single strain). In region* III*, both strains will be maintained (a stable boundary equilibrium or sustained oscillations of both strains).*

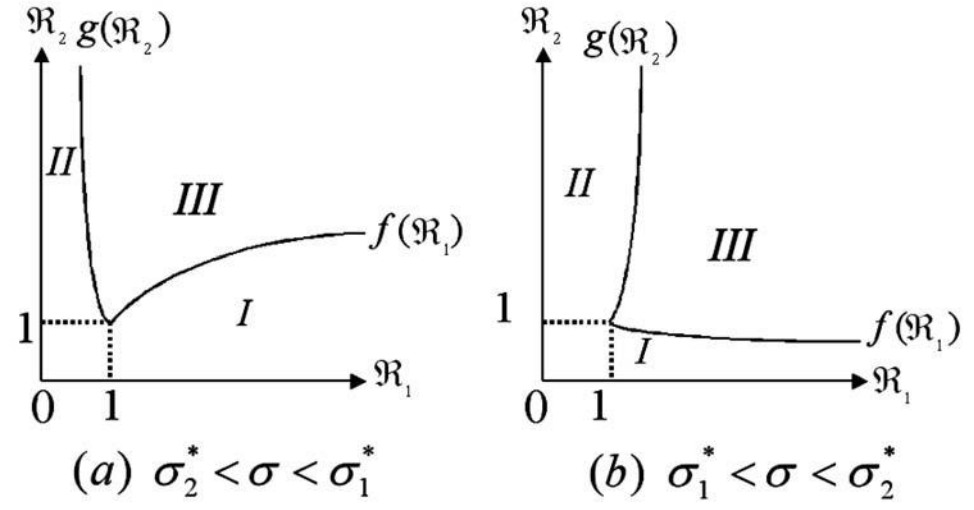


FIG. 4. *Bifurcation diagram in the* $(\Re_1, \Re_2)$*-plane for the case when* $\sigma_1^* \neq \sigma_2^*$*. The curves* $\Re_2 = f(\Re_1)$ *(for* $\Re_1 > 1$*) and* $\Re_1 = g(\Re_2)$ *(for* $\Re_2 > 1$*) divide the region* $\mathbf{R}^2_+ - \{(\Re_1, \Re_2) \mid \Re_1 < 1$ *and* $\Re_2 < 1\}$ *into three subregions:* I, II, III*. The meanings of these regions are the same as those in Figure* 3*.*

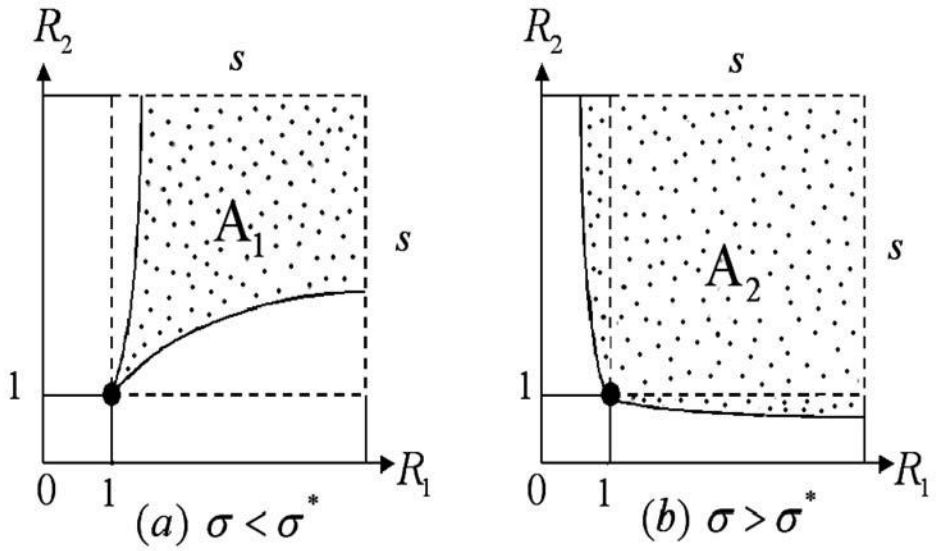


FIG. 5. *A_1 and A_2 are the regions used to approximate the areas pertaining to section III in Figures 2(a)–(b). In order to approximate these areas we enclose regions A_1 and A_2 with a square with dimensions s and corresponding area s^2 (dashed region); s is chosen so that $1 + s = \Re_1$ (and $\Re_2 \approx \Re_1$) using the parameters in Table 1 and two values of σ^* (see the text).*

may become established under these conditions can be small. Likewise, weaker levels of cross-immunity to strain j after an infection with strain i ($\sigma_i \uparrow 1$) will support relatively larger regions of stability for strain j.

The stability regions for strain 1 (I) and strain 2 (II) in the $(\Re_1, \Re_2)$-plane ($\sigma < \sigma_1^*$ and $\sigma > \sigma_2^*$, $\sigma_1^* = \sigma_2^* = \sigma^*$) are illustrated in Figures 3(a)–(b). We show that as the levels of cross-immunity decrease, that is, as the values of σ get closer to 1 (from Figure 3(a) to Figure 3(b)), the region of stability corresponding to each individual strain is reduced significantly (regions I and II). Simultaneously, an increase in the region of multiple strain coexistence (III) can be observed as cross-immunity is weakened. It seems that as strains become antigenically distinct; that is, when cross-immunity against each other is weak, coexistence is more likely. Strong levels of cross-immunity ($\sigma \downarrow 0$) support the survival of a single strain; that is, in this case competition for susceptibles between strains is "fierce" ("competitive exclusion"). The strain with the highest ability to invade the host (largest $\Re_q$) is the most likely to become established (driving the other strain to extinction [4]).

Using Figures 5(a)–(b), a rough estimate for the "probability" of multiple strain coexistence is computed as a function of cross-immunity. The areas of both regions A_1 and A_2 (previously depicted by region III) by delineating the regions of interest with functions $f(\Re_1)$ and $g(\Re_2)$ are "approximately" computed. The area of A_1 in Figure 5(a) is enclosed by a square region with dimensions s, where $1+s=\Re_1$ and $\Re_2 \approx \Re_1$. Similarly, the area A_2 in Figure 5(b) is estimated. The selected value of σ^* ($\sigma^* = 0.33$) used corresponds to the one derived using the parameters provided in Table 1. A value of $\sigma = 0.0008$ is used in Figure 5(a) and $\sigma = 0.8$ in Figure 5(b). Letting A_1 ($\sigma = 0.0008$) and A_2 ($\sigma = 0.8$) in Figures 5(a)–(b) be the calculated areas corresponding to region III, we find that the quotient A_1/A_2 is small (0.0055562).

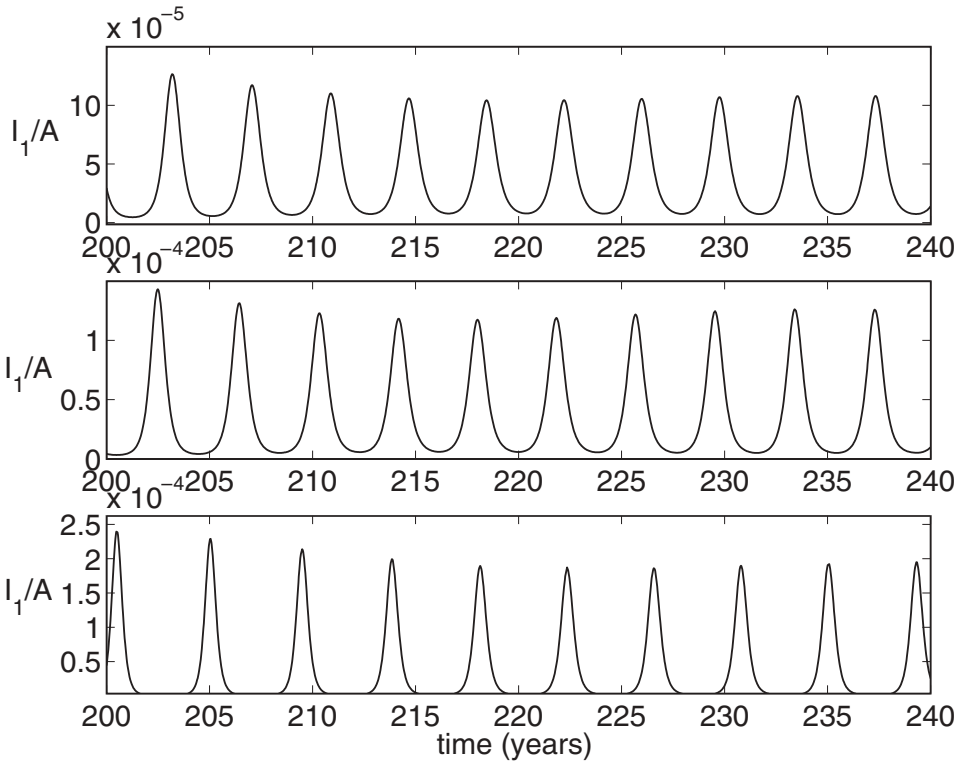


FIG. 6. *Numerical integration of the model equations. The fraction of the infective individuals (nonisolated) with strain* 1 (I_1/N) *is shown for increasing periods of isolation. The length of the isolation period has been chosen (from top to bottom) to be* 3 *days,* 7 *days, and* 15 *days. Cross-immunity between strains is intermediate* ($\sigma = 0.5$).

Hence, the coexistence of antigenically similar strains (sharing strong levels of cross-immunity) seems less likely than when cross-immunity is weak.

5. Numerical results. In this section we explore the model equations numerically as the levels of cross-immunity and isolation are varied. In the first set of simulations, we study the symmetric case ($\sigma_{12} = \sigma_{21} = \sigma$). We explore the role of cross-immunity and host isolation in supporting sustained oscillations for a single and/or both strains where $\sigma \in (0.01, 0.8)$ and $\frac{1}{\alpha}$ is either 1 day, 3 days, or 15 days. In the second set of simulations, we explore the case where $\sigma_{12} \neq \sigma_{21}$. Average life expectancy is fixed at 70 years; infected individuals recover from infection in 5–7 days; individuals are isolated for 1–15 days. The parameters used in simulations are listed in Table 1.

Case 1. $\sigma_{12} = \sigma_{21} = \sigma$. The robustness of multiple strain coexistence begins from the assumption that both strains are present in the population ($s(0) = 0.4$, $i_1(0) = 0.199$, $r_1(0) = r_2(0) = 0.2$, $i_2(0) = 0.001$, and $q_1(0) = q_2(0) = w(0) = 0$). Simulations are conducted using varying levels of cross-immunity ($\sigma = 0.01$, 0.33, 0.5, and 0.8) and isolation periods (1 day, 3 days, and 15 days). Figure 6 shows that for intermediate cross-immunity ($\sigma = 0.5$) the periods between outbreaks is approximately 4 years with an amplitude ranging from 1.1×10^{-4} to 1.5×10^{-4}. Figure 7 shows that strong cross-immunity gives periods of approximately 3 years and amplitude of 2×10^{-4}. As the levels of cross-immunity range from intermediate

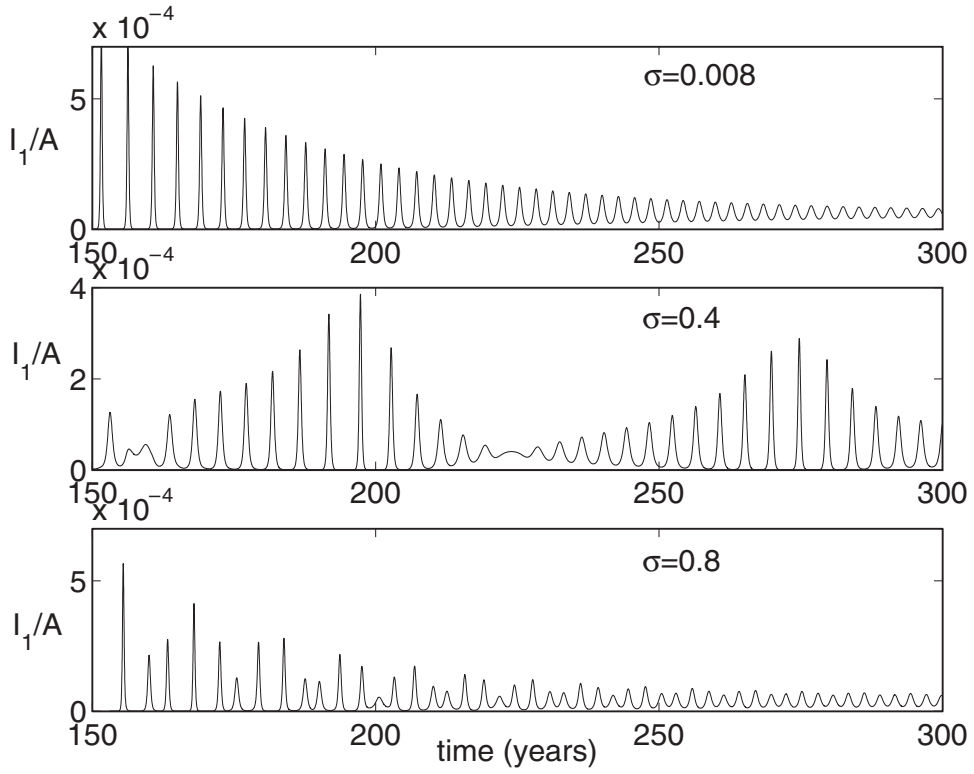


FIG. 7. *Numerical integration of the model equations. The fraction of the infective individuals (nonisolated) with strain* 1 (I_1/N) *is shown. The isolation period is fixed at* 3 *days, while cross-immunity levels are varied (from top to bottom:* 0.008 *(strong),* 0.4 *(intermediate), and* 0.8 *(weak).*

to weak, the periods become more irregular, and the amplitude ranges vary (4×10^{-5} to 4×10^{-4}, $\sigma = 0.4$ and 1–5.5×10^{-4}, $\sigma = 0.8$).

Case 2. $\sigma_{12} \neq \sigma_{21}$. Briefly, we study the effect of isolation for asymmetric cross-immunity by allowing strains to become antigenically distinct with increasing ε (that is, $|\sigma_{12} - \sigma_{21}| = \varepsilon$) (see also [12]). We assume a 3-day isolation period. Figure 8 illustrates the interactions that arise between nonsymmetric strains as their difference in cross-immunity increases, $\varepsilon \in (0.01, 0.03)$. The periods between oscillations for strain 1 (solid) vary from 10–11 years with increasing ε and decreasing levels of cross-immunity. Similarly, the periods between oscillations corresponding to strain 2 (dashed) vary from 10–13 years. The amplitude with highest peak for strain 1 (3.8×10^{-4}) is attained at $\varepsilon = 0.02$, whereas that of strain 2 is observed at $\varepsilon = 0.02$ and $\varepsilon = 0.03$ (3×10^{-4}). Figure 8 shows that for intermediate coupled strains ($\sigma_{ij} \approx 0.33$), the system goes through cycles with approximate periods of 10–13 years, where each cycle may contain minor outbreaks followed by a period with very low disease levels (1–2 $\times 10^{-5}$).

6. Discussion. "Flu" epidemic patterns include yearly outbreaks (antigenic drift), the explosive onset of outbreaks, the rapid termination of local epidemics (despite an "abundance" of susceptible individuals), and potentially major pandemics (antigenic shift). The continuous generation (most likely from random mutations in the NS gene) of new "flu" strains ("minor" genetic changes) and the sudden generation of subtypes (radical genetic changes) and their impact on the history of

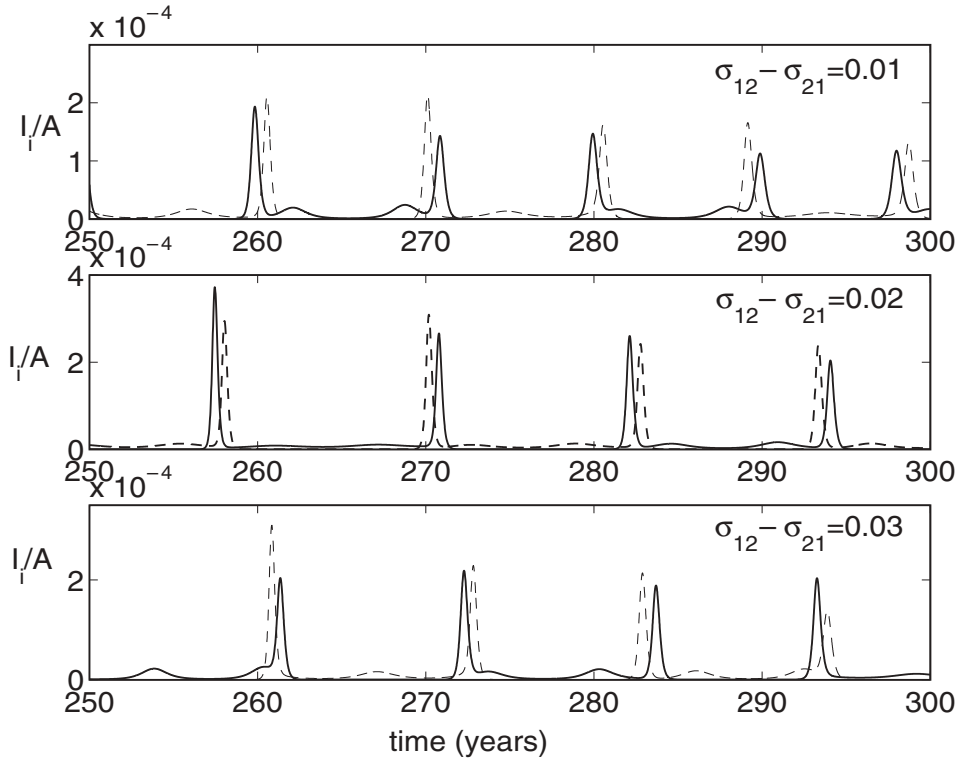


FIG. 8. *Numerical integration of the model equations. The fraction of the infective individuals (nonisolated) with strain* 1 *(solid) and strain* 2 *(dashed) is shown. Differences in cross-immunity levels between strains* 1 *and* 2 $(\sigma_{12} - \sigma_{21})$ *increase (from top to bottom):* 0.01, 0.02, *and* 0.03. *For example, cross-immunity for strains* 1 *and* 2 *correspondingly are given by* $\sigma_{12} = 0.36$ *and* $\sigma_{21} = 0.33$ *(bottom panel).*

acquired (age-dependent) immunity of host populations make the study of influenza dynamics and its control challenging and fascinating [2, 16].

The focus of this article is on the time evolution of influenza A in a nonfixed landscape driven by tight coevolutionary interactions (that is, interactions where the fate of the host and the parasite are intimately connected; see [25]) between human hosts and competing strains. The process is mediated by intervention (behavioral changes) and cross-immunity. In other words, the nature of the invading landscape (susceptible host) changes dynamically from behavioral changes (isolation, short time scale) and past immunological experience (cross-immunity, long time scale).

The "partial" herd-immunity generated by past history of invasions on the host population can have a huge impact on the quantitative dynamics of the "flu" at the population level. The assumption that $\sigma_{12} = \sigma_{21} = \sigma$ for $i \neq j$ naturally results in a dynamic landscape that is not too different (in the oscillatory regime) than the one observed on single-strain models with isolation [15, 20]. That is, a lack of heterogeneity in cross-immunity results in a system "more or less" driven (in the oscillatory regime) by the process of isolation. However, small variations in isolation (Figure 6) leads to radically quantitatively distinct epidemics in the oscillatory regime. This modeling framework (see also [5, 6]) can "asses" the impact of antigenically similar ($|\sigma_{ij} - \sigma_{ji}| \to 0$) and antigenically distinct strains ($|\sigma_{ij} - \sigma_{ji}| > \epsilon$).

In all cases, sustained oscillations with periods that are consistent with influenza epidemics/pandemics are possible [11, 28]. These results are consistent with those obtained in single-strain models [14] (i.e., sustained oscillations are preserved), except that the oscillations are now possible for "realistic" isolation periods. The introduction of a second strain enhances the possibilities. Numerical simulations illustrate various outcomes, including competitive exclusion, coexistence, and subthreshold coexistence. The interepidemic periods range from 2 to 10–13 years, depending on the levels of cross-immunity. Strong intermediate asymmetric cross-immunity leads to interepidemic periods in the range of 10–13 years. Symmetric cross-immunity reduces the range to 1–3 years. The results of intermediate (symmetric) cross-immunity are consistent with those found in [5, 6]. Documented evidence on the cocirculation of strains belonging to the same subtype [11, 28] appears to be consistent with these results.

Our results show that multiple strain coexistence is highly likely for antigenically distinct (weak cross-immunity) strains and not for antigenically similar under symmetric cross-immunity ("competitive exclusion" principle [4]). As the levels of cross-immunity weaken, the likelihood of subthreshold coexistence ($\Re_i^j < 1$) increases. However, "full" understanding of the evolutionary implications that result from human host and influenza virus interactions may require the study of systems that incorporate additional mechanisms such as seasonality in transmission rates, age-structure, individual differences in susceptibility or infectiousness, and the possibility of coinfections. Thacker [28] notes that the observed seasonality of influenza in temperate zones may be the key to observed patterns of recurrent epidemics. Superinfection may also be a mechanism worth consideration, even though studies in [22] show that it is only moderately possible for young individuals to become infected with two different strains in one "flu" season.

The recent flu epidemic [3] which has invaded all 50 states (2003–2004) and our experiences with the recent SARS epidemic [10] are a source of concern. While isolation and quarantine [8] seem effective [10], they can "destabilize" "flu" dynamics (oscillations) and generate some level of uncertainty. The results in this paper suggest the need to explore the long-term impact of current U.S. vaccination policies on the levels of cross-immunity generated by herd-immunity in the case of the flu. Whether or not they increase or reduce the likelihood of a future major outbreak is a question worth considering.

Appendix A. The local stability of the disease-free state follows from the study of the eigenvalues of the Jacobian matrix J of system (1) at E_0. The 10×10 Jacobian matrix J is partitioned after arranging the variables, that is, after rewriting system (1) as $dM/dt = F(M)$, where $M = (S, I_1, I_2, Q_1, Q_2, R_1, R_2, V_1, V_2, W)$. The corresponding eigenvalues are given by $\lambda_i = -\mu$ for $i = 1, 2, 3, 4$; $\lambda_i = \beta_i - (\mu + \gamma_i + \delta_i)$ for $i = 1, 2$; $\lambda_i = -(\mu + \alpha_i)$ for $i = 1, 2$; $\lambda_i = (\mu + \gamma_i)$.

Acknowledgments. The authors are sincerely grateful to Gerardo Chowell, Laura Jones, Prasith Baccam, Catherine Macken, and Christopher Kribs for their beneficial comments and suggestions. The authors would also like to thank LANL (Center for Nonlinear Studies) for their support during the visit of Dr. Castillo-Chavez as a Ulam Scholar.

REFERENCES

[1] V. Andreasen, J. Lin, and S. A. Levin, *The dynamics of co-circulating influenza strains conferring partial cross- immunity*, J. Math. Biol., 35 (1997), pp. 825–842.

[2] azcentral.com, *Flu Is Confirmed in All* 50 *States; Vaccine Dwindles*, http://www.azcentral.com/news/articles/1213flu-outbreak13.html (2003).
[3] D. A. Buonagurio, S. Nakada, J. D. Parvin, M. Krystal, P. Palese, and W. M. Fitch, *Evolution of human influenza A viruses over* 50 *years: Rapid, uniform rate of change in NS gene*, Science, 232 (1986), pp. 980–982.
[4] H. J. Bremermann and H. R. Thieme, *A competitive exclusion principle for pathogen virulence*, J. Math. Biol., 27 (1989), pp. 179–190.
[5] C. Castillo-Chavez, H. W. Hethcote, V. Andreasen, S. A. Levin, and W. M. Liu, *Cross-immunity in the dynamics of homogeneous and heterogeneous populations*, in Mathematical Ecology, World Scientific, Teaneck, NJ, 1988, pp. 303–316.
[6] C. Castillo-Chavez, H. W. Hethcote, V. Andreasen, S. A. Levin, and W. M. Liu, *Epidemiological models with age structure, proportionate mixing, and cross-immunity*, J. Math. Biol., 27 (1989), pp. 233–258.
[7] C. Castillo-Chavez and H. R. Thieme, *Asymptotically autonomous epidemic models*, Mathematical Population Dynamics: Analysis of Heterogeneity, Vol. 1: Theory of Epidemics, O. Arino, D. Axelrod, M. Kimmel, and M. Langlais, eds., Wuerz, Winnepeg, ON, Canada, 1995, pp. 33–50.
[8] C. Castillo-Chavez, C. W. Castillo-Garsow, and A. A. Yakubu, *Mathematical models of isolation and quarantine*, JAMA, 290 (2003), pp. 2876–2877.
[9] CDC (Centers for Disease Control), *The Influenza (Flu) Viruses*, http://www.cdc.gov/ncidod/diseases/flu/viruses.htm (2003).
[10] G. Chowell, P. W. Fenimore, M. A. Castillo-Garsow, and C. Castillo-Chavez, *SARS outbreak in Ontario, Hong Kong and Singapore: The role of diagnosis and isolation as a control mechanism*, J. Theor. Biol., 24 (2003), pp. 1–8.
[11] R. B. Couch and J. A. Kasel, *Immunity to influenza in man*, Ann. Rev. Micro., 31 (1983), pp. 529–549.
[12] J. H. P. Dawes and J. R. Gog, *The onset of oscillatory dynamics in models of multiple disease strains*, J. Math. Biol., 45 (2002), pp. 471–510.
[13] K. Dietz, *Epidemiological interference of virus populations*, J. Math. Biol., 8 (1979), pp. 291–300.
[14] Z. Feng, *Multi-Annual Outbreaks of Childhood Diseases Revisited the Impact of Isolation*, Ph.D. thesis, Arizona State University, Tempe, AZ, 1994.
[15] Z. Feng and H. R. Thieme, *Recurrent Outbreaks of childhood diseases revisited: The impact of isolation*, J. Math. Biosci., 128 (1995), pp. 93–130.
[16] W. M. Fitch, R. M. Bush, C. A. Bender, and N. J. Cox, *Long term trends in the evolution of H(3) HA*1 *human influenza type A*, Proc. Nat. Acad. Sci. U.S.A., 94 (1997), pp. 7712–7718.
[17] J. P. Fox, C. E. Hall, M. K. Cooney, and H. M. Foy, *Influenza virus infections in Seattle families*, 1975–1979, Amer. J. of Epidemiology, 116 (1982), pp. 212–227.
[18] W. H. Hethcote and S. A. Levin, *Periodicity in epidemiological modeling*, in Applied Mathematical Ecology, Biomathematics 18, Springer-Verlag, Berlin, 1989, pp. 193-211.
[19] H. W. Hethcote, *The mathematics of infectious diseases*, SIAM Rev., 42 (2000), pp. 599–653.
[20] W. H. Hethcote, M. Zhien, and L. Shengbing, *Effects of quarantine in six endemic models for infectious diseases*, Math. Biosci., 180 (2002), pp. 141–160.
[21] M. K. Holmes, *Introduction to Pertubation Methods*, Springer-Verlag, New York, 1995.
[22] R. E. Hope-Simpson, *Epidemic mechanisms of Type A influenza*, J. Hyg. Camb., 83 (1979), pp. 11–26.
[23] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1976.
[24] R. M. Krug, *The Influenza Viruses*, Plenum Press, New York, 1989.
[25] S. A. Levin and C. Castillo-Chavez, *Topics in evolutionary biology*, in Mathematical and Statistical Developments of Evolutionary Theory, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 299, S. Lessard, ed., Kluwer Academic Publishers, Dordrecht, Boston, London, 1990, pp. 327–358.
[26] J. Lin, V. Andreasen, and S. A. Levin, *Dynamics of influenza A drift: The linear three-strain model*, J. Math. Biosci., 162 (1999), pp. 33–51.
[27] L. H. Taber, A. Paredes, W. P. Glezen, and R. B. Couch, *Infection with influenza A/Victoria virus in Houston families*, 1976, J. Hyg. Cam., 86 (1981), pp. 303–313.
[28] S. B. Thacker, *The persistence of influenza in human populations*, Epidemi. Rev., 8 (1986), pp. 129–142.
[29] R. G. Webster, W. J. Beam, O. T. Gorman, T. M. Chambers, and Y. Kawaoka, *Evolution and ecology of influenza A viruses*, Micro. Biol. Rev., 56 (1992), pp. 152–179.

SIAM J. APPL. MATH.
Vol. 65, No. 3, pp. 983–1005

HOMOGENIZATION OF INTERFACES BETWEEN RAPIDLY OSCILLATING FINE ELASTIC STRUCTURES AND FLUIDS*

K.-H. HOFFMANN†, N. D. BOTKIN†, AND V. N. STAROVOITOV‡

Abstract. The paper studies the interaction of a periodic solid bristle structure with a fluid. Such problems arise, for example, when modelling biotechnological devices operating in liquids or when simulating epithelium surfaces of blood vessels. The fluid is described by the linearized Navier–Stokes equation whereas the solid part is governed by equations of linear elasticity. The interface conditions are accounted. A homogenized model of the structure is derived by employing the two-scale convergence technique. The model describes a new material which possesses some interesting properties.

Key words. homogenization, fluid-solid interface, biosensor, multi-layered structure

AMS subject classifications. 35B27, 74F10, 74Q10

DOI. 10.1137/S0036139903421572

1. Introduction. We study a mechanical system consisting of a fluid and a rapidly oscillating elastic fine structure interacting with the fluid. The goal is to obtain averaged equations which effectively describe the behavior of the system.

This investigation is motivated by modelling a surface acoustic wave sensor based on the generation and detection of horizontally polarized shear waves (see [3]). Acoustic shear waves are excited through an alternate voltage applied to electrodes deposited on a quartz crystal substrate. The waves are transmitted into a thin isotropic guiding layer covered by a thin gold film that contacts a liquid containing a protein to be detected. The protein adheres to a specific receptor (aptamer) placed on the surface of the gold film. The arising mass loading causes a phase shift in the electric signal to be measured by an electronic circuit.

One can impress the aptamer-protein layer as a periodic bristle or pin structure on the top of the gold film contacting with the liquid (see Figure 1). The thickness of the aptamer-protein layer is about 4 nm, and the number of bristles per surface unit is enormous large. Therefore, the direct numerical modelling of such a structure using fluid-solid interface conditions is impossible. Proper models can be derived using the homogenization technique from [12], [11], [1], [7], [8], and [5] along with the strict treatment of the solid-fluid interface (see, e.g., [6]).

Problems that are close to ours were studied in [13] and [2]. L. Baffico and C. Conca [2] considered the same geometry but the equations differ from ours. J. Sanchez-Hubert [13] investigated almost the same problem. She used techniques based on the Laplace transformation whereas we apply another approach which makes it possible to obtain an explicit representation of solutions to the cell equation, which allows us to investigate the limiting equations and to develop numerical algorithms.

2. Mathematical model. The coupled mechanical system under consideration is shown in Figure 1. The solid part consists of a substrate and pins located on its top.

*Received by the editors January 20, 2003; accepted for publication (in revised form) April 2, 2004; published electronically March 31, 2005.
http://www.siam.org/journals/siap/65-3/42157.html
†Center of Advanced European Studies and Research, Ludwig-Erhard-Allee 2, 53175 Bonn, Germany (phoffman@caesar.de, botkin@caesar.de).
‡Lavrentyev Institute of Hydrodynamics, 630090 Novosibirsk, Russia, Center of Advanced European Studies and Research, Ludwig-Erhard-Allee 2, 53175 Bonn, Germany (starovoitov@caesar.de).

The pin structure is assumed to be periodic in the plane (x_1, x_2) and independent of x_3. The domain of the coupled system is denoted by $\Omega \subset \mathbb{R}^3$. For simplicity, we suppose that Ω is the cube $\{\boldsymbol{x} \in \mathbb{R}^3 \,|\, x_k \in (-1;+1),\ k = 1,2,3\}$. The domains occupied by the fluid and elastic continua are denoted by $\Omega_{_F}$ and $\Omega_{_S}$, respectively; the boundary separating the continua by Γ. Thus, $\Omega = \Omega_{_F} \cup \Gamma \cup \Omega_{_S}$. Let $(\partial\Omega)_{_F} = \partial\Omega \cap \overline{\Omega}_{_F}$ and $(\partial\Omega)_{_S} = \partial\Omega \cap \overline{\Omega}_{_S}$. Then the sets $\Gamma \cup (\partial\Omega)_{_F}$ and $\Gamma \cup (\partial\Omega)_{_S}$ are the boundaries of the domains $\Omega_{_F}$ and $\Omega_{_S}$, respectively.

2.1. Governing equations. We assume that the fluid is weakly compressible, which is physically correct because the operation frequency of the coupled structure lies in the acoustic range and the displacements of the fluid particles are small. This is a typical acoustic approximation which additionally utilizes linearized Navier–Stokes equations (see [9]).

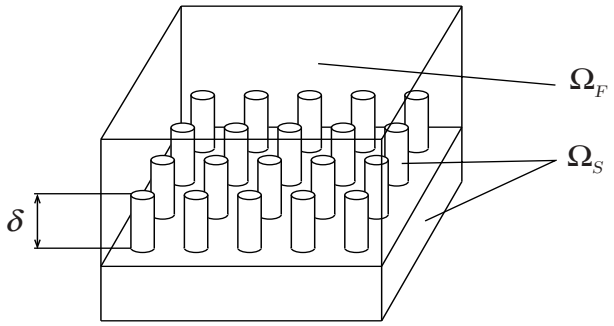


FIG. 1. *Coupled system:* $\Omega = \Omega_{_F} \cup \Gamma \cup \Omega_{_S}$

The solid part of the system will be described using the linear elasticity approach. This linear setting is supplemented by the assumption that the domains $\Omega_{_F}$ and $\Omega_{_S}$ remain unchangeable. Therefore, the coupled mechanical system is described by the following equations:

$$\rho_{_F}\boldsymbol{u}_t = -\nabla p + \operatorname{div} P\boldsymbol{u}_{\boldsymbol{x}} + \rho_{_F}\boldsymbol{f} \quad \text{in} \quad \Omega_{_F}, \tag{2.1}$$

$$\gamma p_t = -\operatorname{div}\boldsymbol{u} \quad \text{in} \quad \Omega_{_F}, \tag{2.2}$$

$$\rho_{_S}\boldsymbol{v}_{tt} = \operatorname{div} G\boldsymbol{v}_{\boldsymbol{x}} + \rho_{_S}\boldsymbol{f} \quad \text{in} \quad \Omega_{_S}. \tag{2.3}$$

Let $\boldsymbol{n}$ be the normal vector to the fluid-solid interface Γ. The no-slip and stress equilibrium conditions on Γ read

$$\boldsymbol{v}_t = \boldsymbol{u} \quad \text{on} \quad \Gamma, \tag{2.4}$$

$$G\boldsymbol{v}_{\boldsymbol{x}} \cdot \boldsymbol{n} = \big(-p\mathcal{I} + P\boldsymbol{u}_{\boldsymbol{x}} \big) \cdot \boldsymbol{n} \quad \text{on} \quad \Gamma. \tag{2.5}$$

The boundary and initial conditions are prescribed:

$$\boldsymbol{u} = 0 \quad \text{on} \quad (\partial\Omega)_{_F}, \tag{2.6}$$

$$\boldsymbol{v} = 0 \quad \text{on} \quad (\partial\Omega)_{_S}, \tag{2.7}$$

$$\boldsymbol{u}|_{t=0} = \boldsymbol{u}^0,\ p|_{t=0} = p^0 \quad \text{in} \quad \Omega_{_F}, \tag{2.8}$$

$$\boldsymbol{v}|_{t=0} = \boldsymbol{v}^0,\ \boldsymbol{v}_t|_{t=0} = \boldsymbol{v}'^0 \quad \text{in} \quad \Omega_{_S}. \tag{2.9}$$

Here, $\rho_{_F}$ and $\rho_{_S}$ are the constant densities of the fluid and of the solid parts, respectively; $\boldsymbol{u}$ is the velocity field of the fluid, p is the pressure in the fluid, $\boldsymbol{v}$ is the displacement field of the solid part, and $\boldsymbol{f}$ is an external force like the gravity. The coefficient γ characterizes the compressibility of the fluid. The fourth-rank tensor $P = \{P_{ijkl}\}$ is defined through the relation

$$P\boldsymbol{u}_{\boldsymbol{x}} = \lambda\,\mathcal{I}\,\mathrm{div}\,\boldsymbol{u} + 2\mu\,\mathcal{D}(\boldsymbol{u}). \tag{2.10}$$

The unit tensor $\mathcal{I}$ has the components $\mathcal{I}_{ij} = \delta_{ij}$, where δ_{ij} is the Kronecker symbol. The strain velocity tensor $\mathcal{D}(\boldsymbol{u})$ has, as is usual, the components $\mathcal{D}_{ij}(\boldsymbol{u}) = 1/2\,(\partial u_i/\partial x_j + \partial u_j/\partial x_i)$. The symbols λ and μ denote positive balk and dynamic viscosity coefficients of the fluid, respectively. As is usual, the summation over repeating indices is assumed. The components G_{ijkl} of the elastic stiffness tensor G can be arbitrary up to base restrictions so that arbitrary *anisotropic solids* can be considered.

The model (2.1)–(2.9) was investigated in [10] where it was supposed to use the velocity instead of the displacement in (2.3). Following this approach, we introduce the integral operator

$$\mathcal{J}_t\boldsymbol{w} = \int_0^t \boldsymbol{w}(s)ds$$

that enables us to rewrite (2.3) in the form

$$\rho_{_S}\boldsymbol{u}_t = \mathrm{div}\big(G\mathcal{J}_t\boldsymbol{u}_{\boldsymbol{x}}\big) + \mathrm{div}\,\mathcal{G}^0 + \rho_{_S}\boldsymbol{f}, \tag{2.11}$$

where $\boldsymbol{u} = \boldsymbol{v}_t$, $\mathcal{G}^0 = G\boldsymbol{v}^0_{\boldsymbol{x}}$ in $\Omega_{_S}$. Similarly, the pressure p can be expressed from (2.3) as follows:

$$p = -\gamma^{-1}\mathrm{div}\,\mathcal{J}_t\boldsymbol{u} + p^0 \quad \text{in} \quad \Omega_{_F}. \tag{2.12}$$

Let χ be the characteristic function of the domain $\Omega_{_F}$. Then (2.1), (2.2), and (2.3) can be written in the whole domain Ω as one equation with discontinuous coefficients

$$\rho\,\boldsymbol{u}_t = \mathrm{div}\big(\boldsymbol{M}^t\boldsymbol{u}_x\big) + \mathrm{div}\,\mathcal{N}^0 + \rho\boldsymbol{f}, \tag{2.13}$$

where

$$\boldsymbol{M}^t = \chi P + \big(\chi\gamma^{-1}\mathcal{I}\otimes\mathcal{I} + (1-\chi)G\big)\,\mathcal{J}_t,$$

$$\rho = \rho_{_F}\chi + \rho_{_S}(1-\chi), \quad \mathcal{N}^0 = -\chi p^0\mathcal{I} + (1-\chi)\mathcal{G}^0.$$

The interface condition (2.4) is equivalent to the "continuity" of $\boldsymbol{u}$ on Γ but the condition (2.5) now assumes the form

$$\big(G\mathcal{J}_t\boldsymbol{u}_{\boldsymbol{x}} + \mathcal{G}^0\big)\cdot\boldsymbol{n} = \big(\gamma^{-1}\mathrm{div}\mathcal{J}_t\,\boldsymbol{u}I - p^0I + P\boldsymbol{u}_{\boldsymbol{x}}\big)\cdot\boldsymbol{n} \quad \text{on} \quad \Gamma \tag{2.14}$$

accounting (2.12). The boundary and initial data are

$$\boldsymbol{u} = 0 \qquad \text{on} \quad (\partial\Omega)_{_F}, \tag{2.15}$$

$$\boldsymbol{u}|_{t=0} = \boldsymbol{u}^0 \qquad \text{in} \quad \Omega, \tag{2.16}$$

where the fluid initial condition $\boldsymbol{u}^0$ is extended to Ω_S by setting $\boldsymbol{u}^0(\boldsymbol{x}) = \boldsymbol{v}'^0(\boldsymbol{x})$ for $\boldsymbol{x} \in \Omega_S$.

Remark 2.1. One can forget the initial distribution $\boldsymbol{v}^0$ of the displacement when considering (2.13). It is sufficient to prescribe the initial velocity field $\boldsymbol{u}^0$ in Ω, the initial stress $\mathcal{G}^0$ in Ω_S (this replaces the information about $\boldsymbol{v}^0$), and initial pressure p^0 in Ω_F. The functions $\mathcal{G}^0$ and p^0 yield the function $\mathcal{N}^0$ involved in (2.13).

Remark 2.2. For mechanical reasons, the tensors P_{ijkl} and G_{ijkl} have the following properties:

$$Z_{ijkl} = Z_{ijlk} = Z_{klij} = Z_{jikl}, \quad Z_{ijkl}\mathcal{V}_{ij}\mathcal{V}_{kl} \geqslant 0,$$

$$Z_{ijkl}\mathcal{V}_{ij}\mathcal{V}_{kl} = 0 \quad \text{if and only if} \quad \mathcal{V}_{kl} + \mathcal{V}_{lk} = 0 \quad \text{for all} \quad k, l = 1, 2, 3.$$

Here, Z_{ijkl} stands for P_{ijkl} or G_{ijkl}.

2.2. Refinement of the structure. Let us define the structure of the regions Ω, Ω_F, and Ω_S more precisely. The pin structure (see Figure 1) is supposed to be (x_1, x_2)-periodic. Without loss of generality, we assume that the periodicity cell is a square with the side length equal to ε, where ε is a positive number. After scaling with the factor $1/\varepsilon$, the cell becomes the unit square $\Sigma = [0, 1] \times [0, 1]$. Let Σ_S be the $1/\varepsilon$-scaled projection of a solid pin to the (x_1, x_2)-plane. It is assumed to be a smooth, simply connected domain in Σ such that its boundary $\partial\Sigma_S$ does not meet $\partial\Sigma$. Denote by Σ_F the domain $\Sigma \setminus \overline{\Sigma}_S$ (see Figure 2).

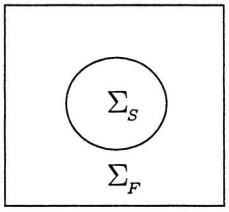


FIG. 2. *Structural cell* $\Sigma = [0, 1] \times [0, 1]$.

Let $\hat{\boldsymbol{x}} = (x_1, x_2)$ and $\hat{\chi}(\hat{\boldsymbol{x}})$ be the Σ-periodic extension of the characteristic function of the domain Σ_F to all $\mathbb{R}^2$. Then the function χ introduced in the previous subsection can be represented as follows:

$$\chi(\boldsymbol{x}) = \chi(\hat{\boldsymbol{x}}, x_3) = \begin{cases} 1, & x_3 > \delta, \\ \hat{\chi}(\frac{\hat{\boldsymbol{x}}}{\varepsilon}), & 0 \leqslant x_3 \leqslant \delta, \\ 0, & x_3 < 0. \end{cases} \tag{2.17}$$

Remember that δ is the thickness of the pin layer. If $\varepsilon \to 0$, the pin structure becomes finer in the (x_1, x_2)-plane, whereas its height remains constant. Thus, the problem (2.13)–(2.16) depends in fact on ε. For this reason, we call it *Problem* S_ε.

DEFINITION 2.3. *A function* $\boldsymbol{u}$ *is called a weak solution to Problem* S_ε *if*

$$\boldsymbol{u} \in L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega_F)), \quad \mathcal{J}_t\boldsymbol{u} \in L^\infty(0, T; H_0^1(\Omega)) \tag{2.18}$$

and the integral identity

$$\int_0^T \int_\Omega \Big(-\rho \boldsymbol{u}\cdot\boldsymbol{\varphi}_t + \boldsymbol{M}^t \boldsymbol{u}_x : \boldsymbol{\varphi}_x + \mathcal{N}^0 : \boldsymbol{\varphi}_x - \rho \boldsymbol{f}\cdot\boldsymbol{\varphi}\Big) d\boldsymbol{x}dt = \int_\Omega \rho \boldsymbol{u}^0\cdot\boldsymbol{\varphi}^0 d\boldsymbol{x} \tag{2.19}$$

holds for every smooth function $\boldsymbol{\varphi}$ *such that* $\boldsymbol{\varphi}|_{t=T} = \boldsymbol{\varphi}|_{\partial\Omega} = 0$.

In this definition and further, T is an arbitrary positive number; the colon denotes the convolution of tensors so that $\mathcal{U} : \mathcal{V} = \mathcal{U}_{ij}\mathcal{V}_{ij}$ for all second-rank tensors $\mathcal{U}$ and $\mathcal{V}$; and the notation f^0 means $f|_{t=0}$. Remark that the second inclusion of (2.18) prevents jumps of $\boldsymbol{u}$ on Γ.

2.3. Solvability of Problem S_ε. It is not difficult to prove existence of a weak solution to Problem S_ε. This question was investigated in [10, section 9.1], and the following result was established.

THEOREM 2.4. *Let* $\boldsymbol{u}^0 \in L^2(\Omega)$, $\mathcal{N}^0 \in L^2(\Omega)$, *and* $\boldsymbol{f} \in L^2([0,T]\times\Omega)$. *Then there exists a unique weak solution to Problem* S_ε, *and the following energy estimate holds:*

$$\operatorname{ess\,sup}_{t\in(0,T)} \Big(\|\boldsymbol{u}(t)\|^2_{L^2(\Omega)} + \|\mathcal{D}\big(\mathcal{J}_t\boldsymbol{u}\big)\|^2_{L^2(\Omega_S)} \Big) + \int_0^T \|\mathcal{D}(\boldsymbol{u}(t))\|^2_{L^2(\Omega_F)}\, dt \leqslant C, \tag{2.20}$$

where C *is a constant which depends on* $\|\boldsymbol{u}^0\|_{L^2(\Omega)}$, $\|\mathcal{N}^0\|_{L^2(\Omega)}$, *and* $\|\boldsymbol{f}\|_{L^2([0,T]\times\Omega)}$ *but does not depend on* ε.

COROLLARY 2.5. *Under the conditions of Theorem* 2.4, *there exists an independent of* ε *constant* C *such that*

$$\operatorname{ess\,sup}_{t\in(0,T)} \|\mathcal{J}_t\boldsymbol{u}(t)\|_{H^1(\Omega)} \leqslant C. \tag{2.21}$$

Generally speaking, the estimates (2.20) and (2.21) are sufficient to fulfill the homogenization of Problem S_ε due to Proposition 3.9 which will be given below. However, some technical difficulties must be overcome in this case. To avoid that, a stronger estimate for $\boldsymbol{u}$ will be obtained under some compatibility conditions. The next theorem states such a result.

THEOREM 2.6. *Let* $\boldsymbol{u}^0 \in H^1(\Omega)$, $\mathcal{N}^0 \in L^2(\Omega)$, $\boldsymbol{f}, \boldsymbol{f}_t \in L^2([0,T]\times\Omega)$, *and*

$$\operatorname{div}\big(\chi P\boldsymbol{u}^0_{\boldsymbol{x}} + \mathcal{N}^0\big) \in L^2(\Omega). \tag{2.22}$$

Then the weak solution to Problem S_ε *satisfies the estimate*

$$\operatorname{ess\,sup}_{t\in(0,T)} \big(\|\boldsymbol{u}_t(t)\|_{L^2(\Omega)} + \|\boldsymbol{u}_x(t)\|_{L^2(\Omega)} \big) \leqslant C, \tag{2.23}$$

where C *is an independent of* ε *constant.*

Proof. Let us introduce a function $\boldsymbol{w}$ as a solution of the problem

$$\begin{aligned} &\rho\,\boldsymbol{w}_t = \operatorname{div}\big(\boldsymbol{M}^t\boldsymbol{w}_x\big) + \operatorname{div}\big(\chi\gamma^{-1}I\operatorname{div}\boldsymbol{u}^0 + (1-\chi)G\boldsymbol{u}^0_{\boldsymbol{x}}\big) + \rho\boldsymbol{f}_t,\\ &\rho\,\boldsymbol{w}\big|_{t=0} = \rho\,\boldsymbol{w}_0 = \operatorname{div}\big(Pu^0_{\boldsymbol{x}} + \mathcal{N}^0\big) + \rho\boldsymbol{f}^0,\\ &\boldsymbol{w}\big|_{\partial\Omega} = 0. \end{aligned}$$

The energy estimate for this problem appears as follows:

$$\operatorname{ess\,sup}_{t\in(0,T)} \Big(\|\boldsymbol{w}(t)\|^2_{L^2(\Omega)} + \|\mathcal{D}\big(\mathcal{J}_t\boldsymbol{w}\big)\|^2_{L^2(\Omega_S)} \Big) + \int_0^T \|\mathcal{D}(\boldsymbol{w}(t))\|^2_{L^2(\Omega_F)}\, dt \leqslant C, \tag{2.24}$$

which yields

$$\operatorname{ess\,sup}_{t\in(0,T)}\|\mathcal{J}_t\boldsymbol{w}\|_{H^1(\Omega)} \leqslant C.$$

The assertion of the theorem is an immediate consequence of the last estimates because the function defined as

$$\boldsymbol{u}(\boldsymbol{x},t) = \int_0^t \boldsymbol{w}(\boldsymbol{x},s)\,ds + \boldsymbol{u}^0(\boldsymbol{x}) = \mathcal{J}_t\boldsymbol{w}(\boldsymbol{x},t) + \boldsymbol{u}^0(\boldsymbol{x})$$

is the solution of Problem S_ε, and $\boldsymbol{u}_t = \boldsymbol{w}$. ∎

According to the definition of $\boldsymbol{u}^0$, the requirement $\boldsymbol{u}^0 \in H^1(\Omega)$ expresses the no-slip condition on Γ at the initial time instant $t = 0$. The requirement (2.22) expresses the stress equilibrium condition on Γ at $t = 0$. From the mechanical point of view, such conditions hold for any time instant including the initial one. Therefore, the requirements of the theorem are feasible.

3. Homogenization of the structure.

3.1. Two-scale convergence. Let us denote by $\boldsymbol{u}_\varepsilon$ the solution of Problem S_ε. In order to emphasize the dependence of χ on ε, we denote it by χ^ε. Our goal is to perform the passage to the limit in Problem S_ε as $\varepsilon \to 0$. To do this, we use the two-scale convergence method introduced by G. Nguetseng and developed by other mathematicians (see [12], [11], [1], [7]). Let us formulate the main results of this approach adapted to our situation.

THEOREM 3.7. *Let $\boldsymbol{w}_\varepsilon$ be a bounded sequence in $L^2\big([0,T]\times\Omega\big)$. There exists a subsequence, still denoted by $\boldsymbol{w}_\varepsilon$, and a function $\overline{\boldsymbol{w}}(t,\boldsymbol{x},\hat{\boldsymbol{\xi}}) \in L^2\big([0,T]\times\Omega\times\Sigma\big)$ such that*

$$\lim_{\varepsilon\to 0}\int_0^T\int_\Omega \boldsymbol{w}_\varepsilon(t,\boldsymbol{x})\boldsymbol{\phi}\left(t,\boldsymbol{x},\frac{\hat{\boldsymbol{x}}}{\varepsilon}\right)d\boldsymbol{x} = \int_0^T\int_\Omega\int_\Sigma \overline{\boldsymbol{w}}(t,\boldsymbol{x},\hat{\boldsymbol{\xi}})\boldsymbol{\phi}(t,\boldsymbol{x},\hat{\boldsymbol{\xi}})\,d\hat{\boldsymbol{\xi}}d\boldsymbol{x}dt$$

for every smooth function $\boldsymbol{\phi}(t,\boldsymbol{x},\hat{\boldsymbol{\xi}})$ which is Σ-periodic in $\hat{\boldsymbol{\xi}}$. Such a sequence $\boldsymbol{w}_\varepsilon$ is said to be two-scale convergent to $\overline{\boldsymbol{w}}(t,\boldsymbol{x},\hat{\boldsymbol{\xi}})$.

Recall the notation $\hat{\boldsymbol{x}} = (x_1, x_2)$ and $\hat{\boldsymbol{\xi}} = (\xi_1, \xi_2)$.

THEOREM 3.8. *Let a sequence $\boldsymbol{w}_\varepsilon$ converge weakly to $\boldsymbol{w}$ in $L^2\big(0,T;H^1(\Omega)\big)$. Then $\boldsymbol{w}_\varepsilon$ two-scale converges to $\boldsymbol{w}$ and there exists a function $\overline{\boldsymbol{w}}(t,\boldsymbol{x},\hat{\boldsymbol{\xi}})$ in $L^2\big([0,T]\times\Omega;H^1_\#(\Sigma)/\mathbb{R}\big)$ such that $\nabla\boldsymbol{w}_\varepsilon$ two-scale converges to $\nabla_x\boldsymbol{w}(t,\boldsymbol{x}) + \nabla_\xi\overline{\boldsymbol{w}}(t,\boldsymbol{x},\hat{\boldsymbol{\xi}})$ up to a subsequence.*

Here $H^1_\#(\Sigma)$ is the space of Σ–periodic functions which belong to the space $H^1(\Sigma)$. Since all functions under consideration do not depend on ξ_3, the notation $\nabla_\xi = (\partial_{\xi_1}, \partial_{\xi_2}, 0)^\top$ is used below.

As a simple application of the theorems stated above, we formulate (without proof) the following result concerning the convergence of solutions of Problem S_ε.

PROPOSITION 3.9. *Let $\boldsymbol{u}_\varepsilon$ be the sequence of solutions to Problem S_ε. Then there exist a subsequence (still denoted by $\boldsymbol{u}_\varepsilon$) and a function $\boldsymbol{u}(t,\boldsymbol{x})$ such that*

1. *$\boldsymbol{u}_\varepsilon$ two-scale converges to $\boldsymbol{u}$, and $\boldsymbol{u}_\varepsilon \to \boldsymbol{u}$ weakly in $L^2\big([0,T]\times\Omega\big)$;*
2. *$\mathcal{J}_t\boldsymbol{u}_\varepsilon$ two-scale converges to $\mathcal{J}_t\boldsymbol{u}$, and $\mathcal{J}_t\boldsymbol{u}_\varepsilon \to \mathcal{J}_t\boldsymbol{u}$ in $L^2\big([0,T]\times\Omega\big)$;*
3. *$\nabla\mathcal{J}_t\boldsymbol{u}_\varepsilon$ two-scale converges to $\nabla_x\mathcal{J}_t\boldsymbol{u} + \nabla_\xi\boldsymbol{\zeta}$, where $\boldsymbol{\zeta}(t,\boldsymbol{x},\hat{\boldsymbol{\xi}})$ is a function from $L^2\big([0,T]\times\Omega;H^1_\#(\Sigma)/\mathbb{R}\big)$.*

3.2. Passage to the limit in Problem S_ε. Let the initial data of Problem S_ε satisfy the conditions of Theorem 2.6. A solution $\boldsymbol{u}_\varepsilon$ of Problem S_ε satisfies the following integral identity,

(3.1)
$$\int_0^T \int_\Omega \Big(-\rho^\varepsilon \boldsymbol{u}_\varepsilon \cdot \boldsymbol{\varphi}_t + \boldsymbol{M}^{\varepsilon t} \boldsymbol{u}_{\varepsilon x} : \boldsymbol{\varphi}_x + \mathcal{N}^{\varepsilon 0} : \boldsymbol{\varphi}_x - \rho^\varepsilon \boldsymbol{f} \cdot \boldsymbol{\varphi} \Big) d\boldsymbol{x} dt = \int_\Omega \rho^\varepsilon \boldsymbol{u}^0 \cdot \boldsymbol{\varphi}^0 d\boldsymbol{x},$$

where ρ^ε, $\boldsymbol{M}^{\varepsilon t}$, and $\mathcal{N}^{\varepsilon 0}$ are defined as in (2.13) but with χ replaced by χ^ε. Let us take

$$\boldsymbol{\varphi}(t, \boldsymbol{x}) = \boldsymbol{\phi}(t, \boldsymbol{x}) + \varepsilon \, \overline{\boldsymbol{\phi}}\Big(t, \boldsymbol{x}, \frac{\hat{\boldsymbol{x}}}{\varepsilon}\Big),$$

where $\boldsymbol{\phi}$ and $\overline{\boldsymbol{\phi}}$ are arbitrary functions that vanish for $\boldsymbol{x} \in \partial\Omega$ and at $t = T$. Theorem 3.8 enables the passage to the limit in (3.1) as $\varepsilon \to 0$. The limiting equations look as follows:

(3.2)
$$\int_0^T \int_\Omega \int_\Sigma \Big(-\rho \boldsymbol{u} \cdot \boldsymbol{\phi}_t + \boldsymbol{M}^t (\boldsymbol{u}_x + \overline{\boldsymbol{u}}_\xi) : \boldsymbol{\phi}_x + \mathcal{N}^0 : \boldsymbol{\phi}_x - \rho \boldsymbol{f} \cdot \boldsymbol{\phi} \Big) \, d\hat{\boldsymbol{\xi}} \, d\boldsymbol{x} dt = \int_\Omega \int_\Sigma \rho \boldsymbol{u}^0 \cdot \boldsymbol{\phi}^0 \, d\hat{\boldsymbol{\xi}} \, d\boldsymbol{x},$$

(3.3)
$$\int_\Sigma \Big(\boldsymbol{M}^t (\boldsymbol{u}_x + \overline{\boldsymbol{u}}_\xi) : \overline{\boldsymbol{\phi}}_\xi + \mathcal{N}^0 : \overline{\boldsymbol{\phi}}_\xi \Big) \, d\hat{\boldsymbol{\xi}} = 0 \quad \text{in} \quad L^2\big([0,T] \times \Omega\big).$$

These equations hold for all functions $\boldsymbol{\phi} \in H^1([0,T] \times \Omega)$ and $\overline{\boldsymbol{\phi}} \in H^1_\#(\Sigma)$ such that $\boldsymbol{\phi}$ vanish on $\partial\Omega$ and at $t = T$. The coefficients ρ, $\boldsymbol{M}^t$, and $\mathcal{N}^0$ are defined as in (2.13) with $\chi(\boldsymbol{x})$ replaced by $\chi(\boldsymbol{x}, \hat{\boldsymbol{\xi}})$. The function $\chi(\boldsymbol{x}, \hat{\boldsymbol{\xi}})$ is defined as in subsection 2.2:

$$\chi(\boldsymbol{x}, \hat{\boldsymbol{\xi}}) = \begin{cases} 1, & x_3 > \delta, \\ \hat{\chi}(\hat{\boldsymbol{\xi}}), & 0 \leqslant x_3 \leqslant \delta, \\ 0, & x_3 < 0. \end{cases}$$

Equation (3.3) is called a *cell equation*.

Equations (3.2) and (3.3) are coupled through the auxiliary function $\overline{\boldsymbol{u}}$. The next step consists of finding $\overline{\boldsymbol{u}}$ from the cell equation (3.3) and substituting the obtained expression into (3.2).

4. Explicit solving of the cell equation.

4.1. Operator form of the cell equation in a Hilbert space. It is appropriate to rewrite (3.3) as an equation in the Hilbert space $H = H^1_\#(\Sigma)/\mathbb{R}$ with the inner product

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \int_\Sigma \frac{\partial u_i}{\partial \xi_j} \frac{\partial v_i}{\partial \xi_j} \, d\hat{\boldsymbol{\xi}}.$$

The norm in H is denoted by $\| \cdot \|$. Let us define operators $\mathcal{A}$ and $\mathcal{B}$ as follows:

$$\langle \mathcal{A}\boldsymbol{u}, \boldsymbol{v} \rangle = \int_\Sigma \chi P_{ijkl} \frac{\partial u_k}{\partial \xi_l} \frac{\partial v_i}{\partial \xi_j} \, d\hat{\boldsymbol{\xi}}, \quad \langle \mathcal{B}\boldsymbol{u}, \boldsymbol{v} \rangle = \int_\Sigma \Big(\chi \gamma^{-1} \delta_{ij} \delta_{kl} + (1 - \chi) \, G_{ijkl} \Big) \frac{\partial u_k}{\partial \xi_l} \frac{\partial v_i}{\partial \xi_j} \, d\hat{\boldsymbol{\xi}}$$

for all functions $\boldsymbol{u}, \boldsymbol{v} \in H$. Due to the Riesz representation theorem, there exist $\boldsymbol{n}_0$, $\boldsymbol{a}_{kl}$, and $\boldsymbol{b}_{kl}$, $k, l = 1, 2, 3$, such that

$$\langle \boldsymbol{n}_0, \boldsymbol{v} \rangle = \int_\Sigma \mathcal{N}^0 : \boldsymbol{v}_\xi \, d\hat{\boldsymbol{\xi}}, \qquad \langle \boldsymbol{a}_{kl}, \boldsymbol{v} \rangle = \int_\Sigma \chi P_{ijkl} \frac{\partial v_i}{\partial \xi_j} \, d\hat{\boldsymbol{\xi}},$$

$$\langle \boldsymbol{b}_{kl}, \boldsymbol{v} \rangle = \int_\Sigma \left(\chi \gamma^{-1} \delta_{ij} \delta_{kl} + (1 - \chi) \, G_{ijkl} \right) \frac{\partial v_i}{\partial \xi_j} \, d\hat{\boldsymbol{\xi}}$$

for all $\boldsymbol{v} \in H$. Remark that $\mathcal{A}$, $\mathcal{B}$, $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{n}_0$ do not depend on t and depend on the variable $\boldsymbol{x}$ just in the same way as the function $\chi(\boldsymbol{x}, \hat{\boldsymbol{\xi}})$. So we can consider $\boldsymbol{x}$ and t in (3.3) as parameters.

Now, the problem (3.3) transforms to the following equation in the space H:

$$\mathcal{A}\overline{\boldsymbol{u}} + \mathcal{B}\mathcal{J}_t \overline{\boldsymbol{u}} = \boldsymbol{g}, \tag{4.1}$$

where

$$\boldsymbol{g} = -\big(\boldsymbol{a}_{kl} + \boldsymbol{b}_{kl} \mathcal{J}_t \big) \frac{\partial u_k}{\partial x_l} - \boldsymbol{n}_0$$

and $\boldsymbol{u}(\boldsymbol{x}, t)$ is from (3.2) and (3.3).

Since the operators $\mathcal{A}$ and $\mathcal{B}$ are trivial whenever $x_3 \notin [0, \delta]$, we consider (4.1) for $x_3 \in [0, \delta]$, which corresponds to the treatment of the pin layer. In this case, the operators $\mathcal{A}$ and $\mathcal{B}$ are degenerated. Therefore, some difficulties appear when solving (4.1).

The next section is devoted to the study of the data of (4.1) to prepare tools for its explicit solving.

4.2. Properties of $\mathcal{A}$, $\mathcal{B}$, and $\boldsymbol{g}$.

PROPOSITION 4.10. *The operator $\mathcal{A}$ has the following properties:*

1. *$\mathcal{A}$ is a bounded self-adjoint operator on H.*
2. *$\langle \mathcal{A}\boldsymbol{u}, \boldsymbol{u} \rangle \geqslant 0$ for all $\boldsymbol{u} \in H$.*
3. *The null-space $N(\mathcal{A}) = \{ \boldsymbol{u} \in H : \ \boldsymbol{u}$ is constant in $\Sigma_F \}$, and $N(\mathcal{A})^\perp \subset \{ \boldsymbol{u} \in H : \ \Delta \boldsymbol{u} = 0$ in $\Sigma_S \}$.*
4. *There exist positive constants c and C such that*

$$c \, \|\boldsymbol{u}\|^2 \leqslant \langle \mathcal{A}\boldsymbol{u}, \boldsymbol{u} \rangle \leqslant C \, \|\boldsymbol{u}\|^2 \tag{4.2}$$

for all $\boldsymbol{u} \in N(\mathcal{A})^\perp$.

5. *The range $R(\mathcal{A})$ is closed in H, $R(\mathcal{A}) = N(\mathcal{A})^\perp$, and $\mathcal{A}^{-1}$ is defined and bounded as an operator on $R(\mathcal{A})$.*

Proof. Assertions 1 and 2 are obvious (see Remark 2.2). The third assertion consists of two parts. In order to prove the first one we have only to establish that

$$N(\mathcal{A}) \subset \{ \boldsymbol{u} \in H : \ \boldsymbol{u} \text{ is constant on } \Sigma_F \}$$

because the opposite inclusion is clearly true. Due to the positiveness of the operator $\mathcal{A}$, its null-space consists of functions $\boldsymbol{u}$ which satisfy the condition $\langle \mathcal{A}\boldsymbol{u}, \boldsymbol{u} \rangle = 0$. Thus, $\boldsymbol{u} \in N(\mathcal{A})$ implies

$$\langle \mathcal{A}\boldsymbol{u}, \boldsymbol{u} \rangle = \int_\Sigma \chi P_{ijkl} \frac{\partial u_k}{\partial \xi_l} \frac{\partial u_i}{\partial \xi_j} \, d\hat{\boldsymbol{\xi}} = 0.$$

Consequently, $\mathcal{D}(\boldsymbol{u}) = 0$ in Σ_F, and, hence, $\boldsymbol{u}$ is constant in Σ_F because of its periodicity.

Let $\boldsymbol{u} \in N(\mathcal{A})^\perp$. By definition, this means that

$$\int_\Sigma \frac{\partial u_k}{\partial \xi_l}\frac{\partial v_k}{\partial \xi_l}\,d\hat{\boldsymbol{\xi}} = \int_{\Sigma_S} \frac{\partial u_k}{\partial \xi_l}\frac{\partial v_k}{\partial \xi_l}\,d\hat{\boldsymbol{\xi}} = 0$$

for any function $\boldsymbol{v} \in C^\infty(\Sigma)$ such that $\boldsymbol{v}$ is constant on $\overline{\Sigma}_F$. Consequently, $\boldsymbol{u}$ is harmonic in Σ_S, which proves the third assertion.

To validate assertion 3, we need only to prove the left inequality since the right one is obvious. Due to the Korn inequality (see, e.g., [14]), there exists a positive constant c_1 such that

$$\int_{\Sigma_F} |\boldsymbol{u}_\xi|^2\,d\hat{\boldsymbol{\xi}} \leqslant c_1\langle \mathcal{A}\boldsymbol{u}, \boldsymbol{u}\rangle$$

for every $\boldsymbol{u} \in H$. If $\boldsymbol{u} \in N(\mathcal{A})^\perp$, then $\boldsymbol{u}$ is harmonic in Σ_S and there exist positive constants c_2 and c_3 such that

$$c_2\int_{\Sigma_S} |\boldsymbol{u}_\xi|^2\,d\hat{\boldsymbol{\xi}} \leqslant \|\boldsymbol{u}\|_{H^{1/2}(\partial\Sigma_S)} \leqslant c_3\int_{\Sigma_F} |\boldsymbol{u}_\xi|^2\,d\hat{\boldsymbol{\xi}}.$$

That is, $\langle \mathcal{A}\boldsymbol{u}, \boldsymbol{u}\rangle \geqslant c\,\|\boldsymbol{u}\|^2$ for some constant c.

When proving assertion 5, denote by $\mathcal{A}_R$ the restriction of $\mathcal{A}$ to $N(\mathcal{A})^\perp$. Due to the estimate (4.2), $R(\mathcal{A}_R)$ is closed in H. Since $R(\mathcal{A}) = R(\mathcal{A}_R)$, we conclude that $R(\mathcal{A})$ is also a closed subspace of H. This implies that $N(\mathcal{A})^\perp = \overline{R(\mathcal{A})} = R(\mathcal{A})$, and (4.2) is true for $\boldsymbol{u} \in R(\mathcal{A})$. Thus, $\mathcal{A}^{-1}$ exists and is bounded if $\mathcal{A}$ is considered being restricted to $R(\mathcal{A})$. The proposition is proved. □

PROPOSITION 4.11. *The operator $\mathcal{B}$ has the following properties:*

1. *$\mathcal{B}$ is a bounded self-adjoint operator on H.*

2. *$\langle \mathcal{B}\boldsymbol{u}, \boldsymbol{u}\rangle \geqslant 0$ for all $\boldsymbol{u} \in H$.*

3. *The null-space $N(\mathcal{B}) = \{\boldsymbol{u} \in H :\ \mathcal{D}(\boldsymbol{u}) = 0$ in Σ_S and $\operatorname{div}\boldsymbol{u} = 0$ in $\Sigma_F\}$, and $N(\mathcal{B})^\perp \subset \{\boldsymbol{u} \in H :\ \Delta\boldsymbol{u} = \nabla q$ in Σ_F for some $q \in L^2(\Sigma)\}$.*

4. *There exist positive constants c and C such that*

$$c\,\|\boldsymbol{u}\|^2 \leqslant \langle \mathcal{B}\boldsymbol{u}, \boldsymbol{u}\rangle \leqslant C\,\|\boldsymbol{u}\|^2 \tag{4.3}$$

for all $\boldsymbol{u} \in N(\mathcal{B})^\perp$.

5. *The range $R(\mathcal{B})$ is closed in H, $R(\mathcal{B}) = N(\mathcal{B})^\perp$, and $\mathcal{B}^{-1}$ is defined and bounded as an operator on $R(\mathcal{B})$.*

Proof. The first two assertions are obvious. To prove the third one, note that

$$\begin{aligned}\langle \mathcal{B}\boldsymbol{u}, \boldsymbol{u}\rangle &= \int_\Sigma \left(\chi\gamma^{-1}\delta_{ij}\delta_{kl} + (1-\chi)G_{ijkl}\right)\frac{\partial u_i}{\partial \xi_j}\frac{\partial u_k}{\partial \xi_l}\,d\hat{\boldsymbol{\xi}}\\ &= \gamma^{-1}\int_{\Sigma_F} (\operatorname{div}\boldsymbol{u})^2\,d\hat{\boldsymbol{\xi}} + \int_{\Sigma_S} G_{ijkl}\frac{\partial u_i}{\partial \xi_j}\frac{\partial u_k}{\partial \xi_l}\,d\hat{\boldsymbol{\xi}}\end{aligned}$$

for every $\boldsymbol{u} \in H$. Therefore, $\langle \mathcal{B}\boldsymbol{u}, \boldsymbol{u}\rangle = 0$ if and only if $\operatorname{div}\boldsymbol{u} = 0$ in Σ_F and $\mathcal{D}(\boldsymbol{u}) = 0$ in Σ_S.

If $\boldsymbol{u} \in N(\mathcal{B})^{\perp}$, then the equalities

$$0 = \langle \boldsymbol{u}, \boldsymbol{v} \rangle = \int_{\Sigma} \boldsymbol{u}_{\xi} \boldsymbol{v}_{\xi} \, d\hat{\boldsymbol{\xi}} = \int_{\Sigma} \mathcal{D}(\boldsymbol{u}) : \mathcal{D}(\boldsymbol{v}) \, d\hat{\boldsymbol{\xi}} = \int_{\Sigma_F} \mathcal{D}(\boldsymbol{u}) : \mathcal{D}(\boldsymbol{v}) \, d\hat{\boldsymbol{\xi}} \tag{4.4}$$

hold for every $\boldsymbol{v} \in N(\mathcal{B})$. Let $\boldsymbol{u}^k \in N(\mathcal{B})^{\perp}$ be a sequence of smooth functions that converges to $\boldsymbol{u}$ in H. Such a sequence exists because $C^{\infty}(\Sigma)$ is dense in $N(\mathcal{B})^{\perp}$. Relation (4.4) is also valid for all $\boldsymbol{u}^k$. If $\boldsymbol{v}$ is an arbitrary smooth function such that $\operatorname{div} \boldsymbol{v} = 0$ and $\operatorname{supp} \boldsymbol{v} \subset \Sigma_F$, then $\boldsymbol{v} \in N(\mathcal{B})$, and

$$0 = \int_{\Sigma_F} \mathcal{D}(\boldsymbol{u}^k) : \mathcal{D}(\boldsymbol{v}) \, d\hat{\boldsymbol{\xi}} = - \int_{\Sigma_F} \operatorname{div}\big(\mathcal{D}(\boldsymbol{u}^k)\big) \cdot \boldsymbol{v} \, d\hat{\boldsymbol{\xi}}.$$

Consequently, there exist functions $\tilde{q}^k \in L^2(\Sigma)$ such that $\operatorname{div}\mathcal{D}(\boldsymbol{u}^k) = \nabla \tilde{q}^k$ for all k. Passing to the limit yields $\operatorname{div}\mathcal{D}(\boldsymbol{u}) = \nabla \tilde{q}$. That is, $\Delta \boldsymbol{u} = \nabla q$, where $q = \tilde{q} - \operatorname{div}\boldsymbol{u}$. This proves the third assertion.

The right inequality of the fourth assertion is obvious. Let us prove the left one. According to the classical theory of the Stokes equations (see [4, Chap. 4]), the following estimate holds for all $\boldsymbol{u} \in N(\mathcal{B})^{\perp}$:

$$\int_{\Sigma_F} |\boldsymbol{u}_{\xi}|^2 \, d\hat{\boldsymbol{\xi}} \leqslant c_1 \big(\|\operatorname{div} \boldsymbol{u}\|^2_{L_2(\Sigma_F)} + \|\boldsymbol{u}_{\Gamma}\|^2_{H^{1/2}(\partial\Sigma_S)/\mathbb{R}} \big),$$

where $\boldsymbol{u}_{\Gamma}$ is the trace of $\boldsymbol{u}$ on $\partial\Sigma_S$. On the other hand,

$$\|\boldsymbol{u}_{\Gamma}\|^2_{H^{1/2}(\partial\Sigma_S)/\mathbb{R}} \leqslant c_2 \int_{\Sigma_S} |\boldsymbol{u}_{\xi}|^2 \, d\hat{\boldsymbol{\xi}}.$$

Thus, there exists a positive constant c_3 such that

$$\|\boldsymbol{u}\|^2 \leqslant c_3 \left(\int_{\Sigma_S} |\boldsymbol{u}_{\xi}|^2 \, d\hat{\boldsymbol{\xi}} + \|\operatorname{div} \boldsymbol{u}\|^2_{L_2(\Sigma_F)} \right) \tag{4.5}$$

for every $\boldsymbol{u} \in N(\mathcal{B})^{\perp}$. In order to obtain (4.3), it is sufficient to prove that there exists a positive constant c_4 such that

$$\int_{\Sigma_S} |\boldsymbol{u}_{\xi}|^2 \, d\hat{\boldsymbol{\xi}} \leqslant c_4 \langle \mathcal{B}\boldsymbol{u}, \boldsymbol{u} \rangle \tag{4.6}$$

for $\boldsymbol{u} \in N(\mathcal{B})^{\perp}$. This can be done using standard contradiction arguments. Assume the converse, i.e., there exists a sequence $\boldsymbol{u}^n \in N(\mathcal{B})^{\perp}$, $n \in \mathbb{N}$, such that $\int_{\Sigma_S} |\boldsymbol{u}^n_{\xi}|^2 \, d\hat{\boldsymbol{\xi}} = 1$ and $\langle \mathcal{B}\boldsymbol{u}^n, \boldsymbol{u}^n \rangle \to 0$ as $n \to 0$. The estimate (4.5) implies that the sequence $\{\boldsymbol{u}^n\}$ is bounded in H too. Thus, there exists its subsequence (still denoted by $\{\boldsymbol{u}^n\}$) that converges weakly in H and $H^1(\Sigma_S)/\mathbb{R}$ but strongly in $L^2(\Sigma)$ to a function $\boldsymbol{u}$. Note that $\boldsymbol{u} \in N(\mathcal{B})^{\perp}$ since $N(\mathcal{B})^{\perp}$ is weakly closed in H. Using the Korn inequality yields

$$\int_{\Sigma_S} |\boldsymbol{u}^n_{\xi} - \boldsymbol{u}_{\xi}|^2 \, d\hat{\boldsymbol{\xi}} \leqslant C \left(\langle \mathcal{B}(\boldsymbol{u}^n - \boldsymbol{u}), \boldsymbol{u}^n - \boldsymbol{u} \rangle + \|\boldsymbol{u}^n - \boldsymbol{u}\|^2_{L^2(\Sigma)} \right).$$

The passage to the limit in this inequality implies that $\langle \mathcal{B}\boldsymbol{u}, \boldsymbol{u} \rangle = 0$ and $\boldsymbol{u}^n \to \boldsymbol{u}$ in H. This means that $\boldsymbol{u} \in N(\mathcal{B})^{\perp} \cap N(\mathcal{B})$ and $\boldsymbol{u} = 0$ in H. On the other

hand, $\int_{\Sigma_S} |\boldsymbol{u}_\xi|^2 \, d\hat{\boldsymbol{\xi}} = \lim_{n\to\infty} \int_{\Sigma_S} |\boldsymbol{u}_\xi^n|^2 \, d\hat{\boldsymbol{\xi}} = 1$. This contradiction proves (4.6) and, consequently, (4.3).

The proof of the fifth assertion is the same as for the operator $\mathcal{A}$ in Proposition 4.10. ∎

PROPOSITION 4.12. *The following is true:*

$$\boldsymbol{a}_{kl}, \boldsymbol{b}_{kl}, \boldsymbol{n}_0 \in R(\mathcal{A}) \cap R(\mathcal{B}), \quad k,l = 1,2,3.$$

Consequently, $\boldsymbol{g} \in R(\mathcal{A}) \cap R(\mathcal{B})$ *for almost all* t *and* $\boldsymbol{x}$*, where* $\boldsymbol{g}$ *is the right-hand side of the cell equation* (4.1).

Proof. Due to Propositions 4.10 and 4.11, $\boldsymbol{w} \in R(\mathcal{A}) \cap R(\mathcal{B})$ if and only if $\langle \boldsymbol{w}, \boldsymbol{v} \rangle = 0$ for all $\boldsymbol{v} \in N(\mathcal{A}) \cup N(\mathcal{B})$. Let us verify this condition for $\boldsymbol{a}_{kl}$. The functions $\boldsymbol{b}_{kl}$ and $\boldsymbol{n}_0$ can be treated in the same way. Let $\boldsymbol{v}$ be an arbitrary function from $N(\mathcal{A})$. That is, $\boldsymbol{v}$ is a constant in Σ_F because of Proposition 4.10. Thus,

$$\langle \boldsymbol{a}_{kl}, \boldsymbol{v} \rangle = \int_{\Sigma_F} P_{ijkl} \frac{\partial v_i}{\partial \xi_j} \, d\hat{\boldsymbol{\xi}} = 0.$$

If $\boldsymbol{v} \in N(\mathcal{B})$ then $\mathcal{D}(\boldsymbol{v}) = 0$ in Σ_S according to Proposition 4.11, and

$$\begin{aligned} \langle \boldsymbol{a}_{kl}, \boldsymbol{v} \rangle &= \int_{\Sigma_F} P_{ijkl} \frac{\partial v_i}{\partial \xi_j} \, d\hat{\boldsymbol{\xi}} = \int_{\Sigma} P_{ijkl} \frac{\partial v_i}{\partial \xi_j} \, d\hat{\boldsymbol{\xi}} - \int_{\Sigma_S} P_{ijkl} \frac{\partial v_i}{\partial \xi_j} \, d\hat{\boldsymbol{\xi}} \\ &= \int_{\Sigma_S} P_{ijkl} \mathcal{D}_{ij}(\boldsymbol{v}) \, d\hat{\boldsymbol{\xi}} = 0. \end{aligned}$$

Here, we used the periodicity of $\boldsymbol{v}$ in Σ and the symmetry of the tensor P (see Remark 2.2). This proves the proposition. ∎

PROPOSITION 4.13.

$$N(\mathcal{A}) \cap N(\mathcal{B}) = \{0\}.$$

Proof. If $\boldsymbol{u} \in N(\mathcal{A}) \cap N(\mathcal{B})$, then $\mathcal{D}(\boldsymbol{u}) = 0$ in Σ due to Propositions 4.10 and 4.11. That is, $\boldsymbol{u}$ is constant in Σ because of its periodicity. This means that $\boldsymbol{u} = 0$ in H. ∎

The result of Proposition 4.13 implies that the operator $\lambda\mathcal{A} + \mathcal{B}$ is invertible for every $\lambda > 0$. Besides that, it is not difficult to see that the operator $(\lambda\mathcal{A} + \mathcal{B})^{-1}$ is bounded in H: Let us introduce the following closed subspaces of H:

$$\begin{aligned} &E_A = (\lambda\mathcal{A} + \mathcal{B})^{-1} R(\mathcal{A}), \\ &E_B = (\lambda\mathcal{A} + \mathcal{B})^{-1} R(\mathcal{B}), \\ &E = E_A \cap E_B = (\lambda\mathcal{A} + \mathcal{B})^{-1} \big(R(\mathcal{A}) \cap R(\mathcal{B}) \big). \end{aligned}$$

Note that the spaces E, E_A, and E_B do not depend on λ. More precisely, if $E_A^\lambda = (\lambda\mathcal{A} + \mathcal{B})^{-1} R(\mathcal{A})$ then $E_A^\lambda = E_A^\mu$ for all $\lambda > 0$ and $\mu > 0$. This follows from simple arguments like those. If $\boldsymbol{x} \in E_A^\lambda$, then $(\lambda\mathcal{A} + \mathcal{B})\boldsymbol{x} \in R(\mathcal{A})$ and $\mathcal{B}\boldsymbol{x} \in R(\mathcal{A})$. Consequently, $(\mu\mathcal{A} + \mathcal{B})\boldsymbol{x} \in R(\mathcal{A})$ and $\boldsymbol{x} \in E_A^\mu$. That is, $E_A^\lambda \subset E_A^\mu$. In the same way we can obtain that $E_A^\mu \subset E_A^\lambda$.

LEMMA 4.14. *The operator* $\mathcal{A}$ *maps the space* E_B *into* $R(\mathcal{B})$*, and the operator* $\mathcal{B}$ *maps the space* E_A *into* $R(\mathcal{A})$.

Proof. The first part is true due to the following implications:

$$\boldsymbol{x} \in E_{\mathcal{B}} \Longrightarrow (\lambda\mathcal{A}+\mathcal{B})\boldsymbol{x} \in R(\mathcal{B}) \Longrightarrow \mathcal{A}\boldsymbol{x} \in R(\mathcal{B}).$$

The second part is being proved analogously. ∎

LEMMA 4.15. *If X is a closed subspace of H then $\mathcal{A}(X)$ and $\mathcal{B}(X)$ are closed in H.*

Proof. Let us verify this assertion for the operator $\mathcal{A}$ by taking an arbitrary sequence $\boldsymbol{u}_n \in \mathcal{A}(X)$ which converges to a function $\boldsymbol{u}$ in H. There exists a corresponding sequence $\boldsymbol{v}_n \in R(\mathcal{A}) \cap X$ such that $\boldsymbol{u}_n = \mathcal{A}(\boldsymbol{v}_n)$. Due to Proposition 4.10, the operator $\mathcal{A}^{-1}$ is bounded on $R(\mathcal{A})$. This implies that the sequence $\{\boldsymbol{v}_n\}$ converges in H to a function $\boldsymbol{v}$ which is in X because X is closed. In the limit, we have $\boldsymbol{u} = \mathcal{A}(\boldsymbol{v})$. That is, $\boldsymbol{u} \in \mathcal{A}(X)$, which proves the lemma. ∎

PROPOSITION 4.16.

$$\mathcal{B}E_{\mathcal{A}} = \mathcal{A}E_{\mathcal{B}} = R(\mathcal{A}) \cap R(\mathcal{B}).$$

That is, for every $\boldsymbol{\psi} \in R(\mathcal{A}) \cap R(\mathcal{B})$, there exist $\boldsymbol{\psi}_{\mathcal{B}} \in E_{\mathcal{B}}$ and $\boldsymbol{\psi}_{\mathcal{A}} \in E_{\mathcal{A}}$ such that $\boldsymbol{\psi} = \mathcal{A}\boldsymbol{\psi}_{\mathcal{B}} = \mathcal{B}\boldsymbol{\psi}_{\mathcal{A}}$.

Proof. Let us prove the first claim. Due to Lemma 4.14, $\mathcal{B}E_{\mathcal{A}} \subset R(\mathcal{A}) \cap R(\mathcal{B})$. Besides that, Lemma 4.15 implies that $\mathcal{B}E_{\mathcal{A}}$ is a closed subspace in H. Suppose that $\mathcal{B}E_{\mathcal{A}} \neq R(\mathcal{A}) \cap R(\mathcal{B})$. Then there exists $\boldsymbol{x} \in R(\mathcal{A}) \cap R(\mathcal{B})$ such that $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = 0$ for every $\boldsymbol{y} \in \mathcal{B}E_{\mathcal{A}}$. That is, $\langle \boldsymbol{x}, \mathcal{B}(\lambda\mathcal{A}+\mathcal{B})^{-1}\mathcal{A}\boldsymbol{z}\rangle = 0$ for all $\boldsymbol{z} \in H$, and

$$\langle \mathcal{A}(\lambda\mathcal{A}+\mathcal{B})^{-1}\mathcal{B}\boldsymbol{x}, \boldsymbol{z}\rangle = 0 \quad \text{for all} \quad \boldsymbol{z} \in H.$$

Consequently, $(\lambda\mathcal{A}+\mathcal{B})^{-1}\mathcal{B}\boldsymbol{x} \in N(\mathcal{A})$ and, hence, $\mathcal{B}\boldsymbol{x} \in (\lambda\mathcal{A}+\mathcal{B})N(\mathcal{A}) = \mathcal{B}N(\mathcal{A})$. That is, there exists $\boldsymbol{y} \in N(\mathcal{A})$ such that $\mathcal{B}\boldsymbol{x} = \mathcal{B}\boldsymbol{y}$ and, therefore, $\mathcal{B}(\boldsymbol{x}-\boldsymbol{y}) = 0$. This implies that $\boldsymbol{w} = \boldsymbol{x}-\boldsymbol{y} \in N(\mathcal{B})$. Thus, $\boldsymbol{x} = \boldsymbol{y}+\boldsymbol{w}$, where $\boldsymbol{y} \in N(\mathcal{A})$, and $\boldsymbol{w} \in N(\mathcal{B})$. That is, $\boldsymbol{x} \in N(\mathcal{A}) \oplus N(\mathcal{B})$. Consequently, $\boldsymbol{x} = 0$ because $\big(N(\mathcal{A}) \oplus N(\mathcal{B})\big) \cap \big(R(\mathcal{A}) \cap R(\mathcal{B})\big) = \{0\}$. The proposition is proved. ∎

Let us introduce the restrictions $\mathcal{A}_{\mathcal{E}}$ and $\mathcal{B}_{\mathcal{E}}$ of the operators $\mathcal{A}$ and $\mathcal{B}$ to the space E.

THEOREM 4.17.

1. *The operators $\mathcal{A}_{\mathcal{E}}$ and $\mathcal{B}_{\mathcal{E}}$ map E onto $R(\mathcal{A}) \cap R(\mathcal{B})$.*
2. *The operators $\mathcal{A}_{\mathcal{E}}$, $\mathcal{B}_{\mathcal{E}} : E \to R(\mathcal{A}) \cap R(\mathcal{B})$ are one to one.*
3. *There exist bounded operators $\mathcal{A}_{\mathcal{E}}^{-1}$, $\mathcal{B}_{\mathcal{E}}^{-1} : R(\mathcal{A}) \cap R(\mathcal{B}) \to E$.*

Proof. Let us prove these assertions for the operator $\mathcal{A}_{\mathcal{E}}$ only. The operator $\mathcal{B}_{\mathcal{E}}$ can be treated in the same way.

1. Since $E \subset E_{\mathcal{B}}$, Proposition 4.16 and Lemma 4.15 imply that $\mathcal{A}E \subset R(\mathcal{A}) \cap R(\mathcal{B})$, and $\mathcal{A}E$ is a closed subspace in H. Suppose that $\mathcal{A}E \neq R(\mathcal{A}) \cap R(\mathcal{B})$. This means that there exists $\boldsymbol{x} \in R(\mathcal{A}) \cap R(\mathcal{B})$ such that $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = 0$ for every $\boldsymbol{y} \in \mathcal{A}E$. That is,

$$\langle (\lambda\mathcal{A}+\mathcal{B})^{-1}\mathcal{A}\boldsymbol{x}, \boldsymbol{z}\rangle = \langle \boldsymbol{x}, \mathcal{A}(\lambda\mathcal{A}+\mathcal{B})^{-1}\boldsymbol{z}\rangle = 0$$

for all $\boldsymbol{z} \in R(\mathcal{A}) \cap R(\mathcal{B})$. Thus, due to Proposition 4.16,

$$\langle (\lambda\mathcal{A}+\mathcal{B})^{-1}\mathcal{A}\boldsymbol{x}, \mathcal{B}\boldsymbol{z}\rangle = 0 \quad \text{for all } \boldsymbol{z} \in E_{\mathcal{A}}. \tag{4.7}$$

Since $(\lambda\mathcal{A}+\mathcal{B})^{-1}\mathcal{A}\boldsymbol{x} \in E_{\mathcal{A}}$, we can take $\boldsymbol{z} = (\lambda\mathcal{A}+\mathcal{B})^{-1}\mathcal{A}\boldsymbol{x}$. Then the relation (4.7) implies that $(\lambda\mathcal{A}+\mathcal{B})^{-1}\mathcal{A}\boldsymbol{x} \in N(\mathcal{B})$, that is, $\mathcal{A}\boldsymbol{x} \in \mathcal{A}N(\mathcal{B})$. Consequently (see the

end of the proof of Proposition 4.16), $\boldsymbol{x} = 0$, which proves the first assertion of the theorem.

2. We have to prove that $N(\mathcal{A}) \cap E = \{0\}$. Let $\boldsymbol{x} \in E$ and $\mathcal{A}\boldsymbol{x} = 0$. Then $\mathcal{B}\boldsymbol{x} = (\lambda\mathcal{A} + \mathcal{B})\boldsymbol{x} \in R(\mathcal{A}) \cap R(\mathcal{B})$, that is, $\mathcal{B}\boldsymbol{x} \in R(\mathcal{A})$. But $\boldsymbol{x} \in N(\mathcal{A}) = R(\mathcal{A})^{\perp}$ and, consequently, $\langle \mathcal{B}\boldsymbol{x}, \boldsymbol{x}\rangle = 0$. Since $\mathcal{B}$ is a positive operator, the last relation implies that $\boldsymbol{x} \in N(\mathcal{B})$. Thus, $\boldsymbol{x} \in N(\mathcal{A}) \cap N(\mathcal{B}) = \{0\}$, which proves the second assertion of the theorem.

3. This assertion is the consequence of parts 1 and 2. The theorem is proved. □

4.3. Solving the cell equation. Now we are in position to find an explicit representation of solutions to the cell equation (4.1). With a new unknown function $\overline{\boldsymbol{\zeta}} = \mathcal{J}_t\overline{\boldsymbol{u}}$, the problem (4.1) assumes the form

$$\mathcal{A}\overline{\boldsymbol{\zeta}}_t + \mathcal{B}\overline{\boldsymbol{\zeta}} = \boldsymbol{g}, \qquad \overline{\boldsymbol{\zeta}}(0) = 0. \tag{4.8}$$

As it follows from Theorem 4.17, the operator $\mathcal{A}_{_E}$ ($\mathcal{A}$ restricted to E) is invertible, the operator $\mathcal{A}_{_E}^{-1}\mathcal{B}_{_E}$ bounded, and $\mathcal{A}_{_E}^{-1}\boldsymbol{g} \in E$. Therefore, the problem

$$\overline{\boldsymbol{\zeta}}_t + \mathcal{A}_{_E}^{-1}\mathcal{B}_{_E}\overline{\boldsymbol{\zeta}} = \mathcal{A}_{_E}^{-1}\boldsymbol{g}, \qquad \overline{\boldsymbol{\zeta}}(0) = 0 \tag{4.9}$$

is uniquely solvable on the subspace E, and the solution is of the form

$$\overline{\boldsymbol{\zeta}}(t) = \int_0^t e^{-(t-s)\mathcal{A}_{_E}^{-1}\mathcal{B}_{_E}}\mathcal{A}_{_E}^{-1}\boldsymbol{g}(s)\,ds. \tag{4.10}$$

THEOREM 4.18. *Equations* (4.8) *and* (4.9) *are equivalent.*

Proof. Obviously, if $\overline{\boldsymbol{\zeta}}$ is a solution to (4.9), then $\overline{\boldsymbol{\zeta}}$ satisfies (4.8). If $\overline{\boldsymbol{\zeta}}$ is a solution to (4.8), then the function $\overline{\boldsymbol{\eta}} = e^{-\lambda t}\overline{\boldsymbol{\zeta}}$ solves the problem

$$\mathcal{A}\overline{\boldsymbol{\eta}}_t + (\lambda\mathcal{A} + \mathcal{B})\overline{\boldsymbol{\eta}} = e^{-\lambda t}\boldsymbol{g}, \qquad \overline{\boldsymbol{\eta}}(0) = 0. \tag{4.11}$$

Since the operator $\lambda\mathcal{A} + \mathcal{B}$ is nondegenerate for any $\lambda > 0$, we can rewrite (4.11) as follows:

$$(\lambda\mathcal{A} + \mathcal{B})^{-1}\mathcal{A}\overline{\boldsymbol{\eta}}_t + \overline{\boldsymbol{\eta}} = e^{-\lambda t}(\lambda\mathcal{A} + \mathcal{B})^{-1}\boldsymbol{g}, \qquad \overline{\boldsymbol{\eta}}(0) = 0. \tag{4.12}$$

Due to Proposition 4.12, $\boldsymbol{g} \in R(\mathcal{A})$, and, hence $\overline{\boldsymbol{\eta}}(t)$ must belong to $E_{_A}$ for all t. Therefore, $\overline{\boldsymbol{\zeta}}(t) \in E_{_A}$ for all t. On the other hand, (4.8) can be rewritten as follows:

$$(\lambda\mathcal{A} + \mathcal{B})\overline{\boldsymbol{\zeta}}_t - \mathcal{B}\overline{\boldsymbol{\zeta}}_t + \lambda\mathcal{B}\overline{\boldsymbol{\zeta}} = \lambda\boldsymbol{g}.$$

That is,

$$\overline{\boldsymbol{\zeta}}_t = (\lambda\mathcal{A} + \mathcal{B})^{-1}\mathcal{B}(\overline{\boldsymbol{\zeta}}_t - \lambda\overline{\boldsymbol{\zeta}}) + \lambda(\lambda\mathcal{A} + \mathcal{B})^{-1}\boldsymbol{g}.$$

Taking into account that $\overline{\boldsymbol{\zeta}}(t)$ and $\overline{\boldsymbol{\zeta}}_t(t) \in E_{_A}$ for all t, we establish, using Proposition 4.16, that $\overline{\boldsymbol{\zeta}}_t(t) \in E$ for all t. Since $\overline{\boldsymbol{\zeta}}(0) = 0$, we conclude that $\overline{\boldsymbol{\zeta}}(t) \in E$ for all t. Therefore, $\overline{\boldsymbol{\zeta}}$ is a solution of (4.9). The theorem is proved. □

Thus, the unique solution of the problem (4.8) is given by (4.10) and the unique solution $\overline{\boldsymbol{u}}$ of the problem (4.1) reads as

$$\overline{\boldsymbol{u}}(t) = \overline{\boldsymbol{\zeta}}_t(t) = \mathcal{A}_{_E}^{-1}\boldsymbol{g}(t) - \mathcal{A}_{_E}^{-1}\mathcal{B}_{_E}\int_0^t e^{-(t-s)\mathcal{A}_{_E}^{-1}\mathcal{B}_{_E}}\mathcal{A}_{_E}^{-1}\boldsymbol{g}(s)\,ds. \tag{4.13}$$

5. Homogenized structure.

5.1. Limiting equations. Substitution of the expression for $\boldsymbol{g}$ into (4.13) gives

(5.1)
$$\overline{\boldsymbol{u}}(t) = -e^{-t\mathcal{A}_{E}^{-1}\mathcal{B}_{E}}\mathcal{A}_{E}^{-1}\boldsymbol{n}_0 - \mathcal{A}_{E}^{-1}\boldsymbol{a}_{kl}\,\frac{\partial u_k(t)}{\partial x_l} - \int_0^t \boldsymbol{m}_{kl}(t-s)\,\frac{\partial u_k(s)}{\partial x_l}\,ds,$$

(5.2)
$$\mathcal{J}_t\overline{\boldsymbol{u}}(t) = \big(e^{-t\mathcal{A}_{E}^{-1}\mathcal{B}_{E}} - \mathcal{I}\big)\mathcal{B}_{E}^{-1}\boldsymbol{n}_0 - \mathcal{B}_{E}^{-1}\boldsymbol{b}_{kl}\int_0^t \frac{\partial u_k(s)}{\partial x_l}\,ds - \int_0^t \widetilde{\boldsymbol{m}}_{kl}(t-s)\frac{\partial u_k(s)}{\partial x_l}\,ds,$$

where

$$\boldsymbol{m}_{kl}(t) = -\mathcal{A}_{E}^{-1}\mathcal{B}_{E}e^{-t\mathcal{A}_{E}^{-1}\mathcal{B}_{E}}\big(\mathcal{A}_{E}^{-1}\boldsymbol{a}_{kl} - \mathcal{B}_{E}^{-1}\boldsymbol{b}_{kl}\big) \in E,$$

$$\widetilde{\boldsymbol{m}}_{kl}(t) = e^{-t\mathcal{A}_{E}^{-1}\mathcal{B}_{E}}\big(\mathcal{A}_{E}^{-1}\boldsymbol{a}_{kl} - \mathcal{B}_{E}^{-1}\boldsymbol{b}_{kl}\big) \in E.$$

The integration by parts and the formula

$$\frac{d}{ds}e^{-(t-s)\mathcal{A}_{E}^{-1}\mathcal{B}_{E}} = \mathcal{A}_{E}^{-1}\mathcal{B}_{E}e^{-(t-s)\mathcal{A}_{E}^{-1}\mathcal{B}_{E}}$$

are applied when deriving (5.1) and (5.2). Now we are in position to compute the principal term

$$\int_\Sigma M_{ijkl}^t\frac{\partial\overline{u}_k}{\partial\xi_l}\,d\hat{\boldsymbol{\xi}} = \langle\boldsymbol{a}_{ij},\overline{\boldsymbol{u}}\rangle + \langle\boldsymbol{b}_{ij},\mathcal{J}_t\overline{\boldsymbol{u}}\rangle$$

appearing in the limiting (homogenized) equation (3.2). Utilizing (5.1) and (5.2) and computing other terms in (3.2), we obtain the following limiting equation:

(5.3)
$$\begin{aligned}&\int_0^T\int_\Omega\Big(-\rho_\theta u_i\,\frac{\partial\phi_i}{\partial t} + \big(\theta P_{ijkl} - \alpha_{ijkl}\big)\frac{\partial u_k}{\partial x_l}\,\frac{\partial\phi_i}{\partial x_j}\\ &\quad+\int_0^t\Big(\theta\gamma^{-1}\delta_{ij}\delta_{kl} + (1-\theta)G_{ijkl} - \beta_{ijkl} + \omega_{ijkl}(t-s)\Big)\frac{\partial u_k}{\partial x_l}\,ds\,\frac{\partial\phi_i}{\partial x_j}\Big)\,d\boldsymbol{x}dt\\ &\quad=\int_0^T\int_\Omega\Big(\rho_\theta f_i\phi_i - \big(\nu_{ij} - \theta p^0\delta_{ij} + (1-\theta)\mathcal{G}_{ij}^0\big)\,\frac{\partial\phi_i}{\partial x_j}\Big)\,d\boldsymbol{x}dt + \int_\Omega\rho_\theta\boldsymbol{u}^0\cdot\boldsymbol{\phi}^0\,d\boldsymbol{x},\end{aligned}$$

where

$$\theta(\boldsymbol{x}) = \int_\Sigma\chi\,d\hat{\boldsymbol{\xi}}, \qquad \rho_\theta = \theta\rho_{{}_F} + (1-\theta)\,\rho_{{}_S},$$

$$\begin{aligned}&\nu_{ij} = -\langle\boldsymbol{a}_{ij},e^{-t\mathcal{A}_{E}^{-1}\mathcal{B}_{E}}\mathcal{A}_{E}^{-1}\boldsymbol{n}_0\rangle + \langle\boldsymbol{b}_{ij},\big(e^{-t\mathcal{A}_{E}^{-1}\mathcal{B}_{E}} - \mathcal{I}\big)\mathcal{B}_{E}^{-1}\boldsymbol{n}_0\rangle,\\ &\alpha_{ijkl} = \langle\boldsymbol{a}_{ij},\mathcal{A}_{E}^{-1}\boldsymbol{a}_{kl}\rangle,\\ &\beta_{ijkl} = \langle\boldsymbol{b}_{ij},\mathcal{B}_{E}^{-1}\boldsymbol{b}_{kl}\rangle,\\ &\omega_{ijkl}(t) = -\langle\boldsymbol{a}_{ij},\boldsymbol{m}_{kl}\rangle - \langle\boldsymbol{b}_{ij},\widetilde{\boldsymbol{m}}_{kl}\rangle.\end{aligned}$$

Let us denote by $\overline{P}$, $\overline{G}$, and $\mathcal{S}^0$ the tensors with components

$$\overline{P}_{ijkl} = \theta P_{ijkl} - \alpha_{ijkl}, \quad \overline{G}_{ijkl} = \theta\gamma^{-1}\delta_{ij}\delta_{kl} + (1-\theta)G_{ijkl} - \beta_{ijkl},$$

$$\mathcal{S}^0_{ij} = \nu_{ij} - \theta p^0\delta_{ij} + (1-\theta)\mathcal{G}^0_{ij}.$$

Let us divide the domain Ω into three parts:

$$\Omega^f = \{\boldsymbol{x} \in \Omega \,|\, x_3 > \delta\}, \quad \Omega^s = \{\boldsymbol{x} \in \Omega \,|\, x_3 < 0\}, \quad \Omega^h = \{\boldsymbol{x} \in \Omega \,|\, 0 < x_3 < \delta\}.$$

Let Γ^+_δ be the boundary between Ω^f and Ω^h, Γ^-_δ the boundary between Ω^s and Ω^h. That is, $\Omega = \Omega^f \cup \Gamma^+_\delta \cup \Omega^h \cup \Gamma^-_\delta \cup \Omega^s$. Note that $\theta(\boldsymbol{x}) = 1$ if $\boldsymbol{x} \in \Omega^f$, $\theta(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \in \Omega^s$, and θ is a constant from the interval $(0,1)$ for $\boldsymbol{x} \in \Omega^h$. As for α_{ijkl}, β_{ijkl}, ν_{ij}, and ω_{ijkl}, they are constants for $\boldsymbol{x} \in \Omega^h$ and equal to zero if $\boldsymbol{x} \in \Omega^f \cup \Omega^s$, so that the integral identity (5.3) delivers the following equations which should be understood in the distributional sense:

(5.4)
$$\rho_{_F}\boldsymbol{u}_t - \operatorname{div} P\boldsymbol{u}_x - \gamma^{-1}\nabla\operatorname{div}\mathcal{J}_t\boldsymbol{u} = -\nabla p^0 + \rho_{_F}\boldsymbol{f}, \qquad \boldsymbol{x} \in \Omega^f,$$

(5.5)
$$\rho_{_S}\boldsymbol{u}_t - \operatorname{div}\mathcal{J}_t G\boldsymbol{u}_x = \operatorname{div}\mathcal{G}^0 + \rho_{_S}\boldsymbol{f}, \qquad \boldsymbol{x} \in \Omega^s,$$

(5.6)
$$\rho_\theta\boldsymbol{u}_t - \operatorname{div}\overline{P}\boldsymbol{u}_x - \operatorname{div}\mathcal{J}_t\overline{G}\boldsymbol{u}_x - \operatorname{div}\int_0^t \omega(t-s)\boldsymbol{u}_x(s)\,ds + \operatorname{div}\mathcal{S}^0 = \rho_\theta\boldsymbol{f}, \quad \boldsymbol{x} \in \Omega^h.$$

The natural interfacial boundary conditions at Γ^+_δ and Γ^-_δ can be derived from the integral identity (5.3). Equations (5.4) and (5.5) coincide with (2.1) and (2.11), respectively. That is, the governing equations for the pure fractions do not change after the homogenization, which have been expected. What is new is an integral-differential equation (5.6) which cannot be reduced to a pure differential equation by differentiating or by a substitution like $\boldsymbol{w} = \mathcal{J}_t\boldsymbol{u}$. The operators involved in the equation have to be investigated to confirm the parabolic type of its principal part.

5.2. Investigation of $\overline{P}$ and $\overline{G}$. It is not difficult to verify that the tensors $\overline{P}$ and $\overline{G}$ have the symmetry properties mentioned in Remark 2.2. Therefore, $\overline{P}_{ijkl}\mathcal{Z}_{ij}\mathcal{Z}_{kl} = 0$ and $\overline{G}_{ijkl}\mathcal{Z}_{ij}\mathcal{Z}_{kl} = 0$ for every skew symmetric matrix $\mathcal{Z}$. The main objective of this subsection is to prove the strong positiveness of the tensor $\overline{P}$ and the nonnegativeness of $\overline{G}$ on the space of symmetric matrices. The null-space of $\overline{G}$ will be also described.

PROPOSITION 5.19. *For every second-rank tensor $\mathcal{Z}$, the following is valid:*

$$\overline{P}_{ijkl}\mathcal{Z}_{ij}\mathcal{Z}_{kl} \geqslant 0, \qquad \overline{G}_{ijkl}\mathcal{Z}_{ij}\mathcal{Z}_{kl} \geqslant 0.$$

Proof. Let us prove the assertion for $\overline{P}$. Denote $\boldsymbol{z} = \boldsymbol{a}_{ij}\mathcal{Z}_{ij}$. Due to Proposition 4.12, $\boldsymbol{z} \in R(\mathcal{A}) \cap R(\mathcal{B})$ and, as it follows from Theorem 4.17, there exists a unique $\boldsymbol{y} \in E$ such that $\mathcal{A}_{_E}\boldsymbol{y} = \boldsymbol{z}$. This means that

$$\int_\Sigma \chi P_{ijkl}\frac{\partial y_i}{\partial\xi_j}\frac{\partial v_k}{\partial\xi_l}\,d\hat{\boldsymbol{\xi}} = \langle\boldsymbol{z},\boldsymbol{v}\rangle = \int_\Sigma \chi P_{ijkl}\mathcal{Z}_{ij}\frac{\partial v_k}{\partial\xi_l}\,d\hat{\boldsymbol{\xi}} \tag{5.7}$$

for all $\boldsymbol{v} \in H$. On the other hand, the definition yields

$$\alpha_{ijkl}\mathcal{Z}_{ij}\mathcal{Z}_{kl} = \langle\boldsymbol{a}_{ij}\mathcal{Z}_{ij}, \mathcal{A}_{_E}^{-1}\boldsymbol{a}_{kl}\mathcal{Z}_{kl}\rangle = \langle\boldsymbol{z}, \mathcal{A}_{_E}^{-1}\boldsymbol{z}\rangle = \langle\mathcal{A}_{_E}\boldsymbol{y},\boldsymbol{y}\rangle.$$

From the last relation and (5.7) with $\boldsymbol{v} = \boldsymbol{y}$, we obtain

(5.8)
$$\overline{P}_{ijkl}\mathcal{Z}_{ij}\mathcal{Z}_{kl} = \theta\, P_{ijkl}\mathcal{Z}_{ij}\mathcal{Z}_{kl} - \langle \mathcal{A}_{E}\boldsymbol{y}, \boldsymbol{y}\rangle = \int_{\Sigma}\Big(\chi P_{ijkl}\mathcal{Z}_{ij}\mathcal{Z}_{kl} - \chi P_{ijkl}\frac{\partial y_i}{\partial \xi_j}\frac{\partial y_k}{\partial \xi_l}\Big)\, d\hat{\boldsymbol{\xi}}$$
$$= \int_{\Sigma}\chi P_{ijkl}\Big(\mathcal{Z}_{ij} - \frac{\partial y_i}{\partial \xi_j}\Big)\Big(\mathcal{Z}_{kl} - \frac{\partial y_k}{\partial \xi_l}\Big)\, d\hat{\boldsymbol{\xi}}.$$

The right-hand side of the last relation is clearly positive and the required assertion is proved for the tensor $\overline{P}$. Positiveness of the tensor $\overline{G}$ can be verified in the same way. ∎

The next theorem states the strong positiveness of the tensor $\overline{P}$.

THEOREM 5.20. *There exists a positive constant C such that*

$$\overline{P}_{ijkl}\mathcal{Z}_{ij}\mathcal{Z}_{kl} \geqslant C\,|\mathcal{Z}|^2$$

for every symmetric second-rank tensor $\mathcal{Z}$. Here, $|\mathcal{Z}|^2 = \mathcal{Z}_{ij}\mathcal{Z}_{ij}$.

Proof. Assume that the assertion of the theorem is false. Then there exists a sequence $\{\mathcal{Z}^n\}$ such that $|\mathcal{Z}^n| = 1$ and $\overline{P}_{ijkl}\mathcal{Z}^n_{ij}\mathcal{Z}^n_{kl} \to 0$ as $n \to \infty$. The sequence $\{\mathcal{Z}^n\}$ is compact in $\mathbb{R}^3 \times \mathbb{R}^3$ and, therefore, it has a subsequence denoted again by $\{\mathcal{Z}^n\}$, which converges to a matrix $\mathcal{Z}^0$ such that $|\mathcal{Z}^0| = 1$. This means that the corresponding sequences $\boldsymbol{z}^n$ and $\boldsymbol{y}^n$, defined as $\boldsymbol{z}^n = \boldsymbol{a}_{ij}\mathcal{Z}^n_{ij}$ and $\boldsymbol{y}^n = \mathcal{A}_{E}^{-1}\boldsymbol{z}^n$, converge in H to $\boldsymbol{z}^0$ and $\boldsymbol{y}^0$, respectively. We use here the notations introduced in the proof of the previous proposition. Thus, the relation

$$\overline{P}_{ijkl}\mathcal{Z}^0_{ij}\mathcal{Z}^0_{kl} = \theta\, P_{ijkl}\mathcal{Z}^0_{ij}\mathcal{Z}^0_{kl} - \langle \mathcal{A}_{E}\boldsymbol{y}^0, \boldsymbol{y}^0\rangle = \int_{\Sigma}\chi P_{ijkl}\Big(\mathcal{Z}^0_{ij} - \frac{\partial y^0_i}{\partial \xi_j}\Big)\Big(\mathcal{Z}^0_{kl} - \frac{\partial y^0_k}{\partial \xi_l}\Big)\, d\hat{\boldsymbol{\xi}} = 0.$$

holds due to (5.8). That is,

$$\chi P_{ijkl}\Big(\mathcal{Z}^0_{ij} - \frac{\partial y^0_i}{\partial \xi_j}\Big)\Big(\mathcal{Z}^0_{kl} - \frac{\partial y^0_k}{\partial \xi_l}\Big) = 0 \qquad \text{in} \quad \Sigma,$$

and, consequently, $\mathcal{D}(\boldsymbol{y}^0) = \mathcal{Z}^0$ in Σ_{F}. This implies that $\mathcal{D}(\boldsymbol{y}^0 - \mathcal{Z}^0\boldsymbol{\xi}) = 0$ in Σ_{F}. Therefore, $\boldsymbol{y}^0(\boldsymbol{\xi})$ is a linear function of $\boldsymbol{\xi}$ for $\boldsymbol{\xi} \in \Sigma_{F}$. The only linear function satisfying the periodicity boundary conditions on $\partial\Sigma$ is a constant, which implies that $\mathcal{Z}^0 = 0$. This is impossible because $|\mathcal{Z}^0| = 1$. This contradiction proves the theorem. ∎

Remark that the arguments like those in the proof of Theorem 5.20 do not lead to a contradiction in the case of the tensor $\overline{G}$. The next theorem shows that the tensor $\overline{G}$ is degenerated and describes its null-space.

THEOREM 5.21. *The tensor $\overline{G}$ is degenerate, and $\overline{G}_{ijkl}\mathcal{Z}_{ij}\mathcal{Z}_{kl} = 0$ for a symmetric matrix $\mathcal{Z}$ if and only if $\mathcal{Z}_{11} + \mathcal{Z}_{22} = 0$ and $\mathcal{Z}_{33} = 0$.*

Proof. Let us denote $\boldsymbol{z} = \boldsymbol{b}_{ij}\mathcal{Z}_{ij}$. Due to Proposition 4.12 and Theorem 4.17, $\boldsymbol{z} \in R(\mathcal{A}) \cap R(\mathcal{B})$, and there exist unique elements $\boldsymbol{y}^{E} \in E$ and $\boldsymbol{y}^{R} \in R(\mathcal{B})$ such that

$$\mathcal{B}\boldsymbol{y}^{E} = \boldsymbol{z}, \qquad \mathcal{B}\boldsymbol{y}^{R} = \boldsymbol{z}. \tag{5.9}$$

It follows that $\boldsymbol{y}^{N} = \boldsymbol{y}^{E} - \boldsymbol{y}^{R} \in N(\mathcal{B})$. Besides that, $\mathcal{B}_{E}^{-1}\mathcal{B}\boldsymbol{y}^{E} = \boldsymbol{y}^{E}$. Therefore,

$$\langle \boldsymbol{z}, \mathcal{B}_{E}^{-1}\boldsymbol{z}\rangle = \langle \mathcal{B}\boldsymbol{y}^{E}, \boldsymbol{y}^{E}\rangle = \langle \mathcal{B}\boldsymbol{y}^{R}, \boldsymbol{y}^{R} + \boldsymbol{y}^{N}\rangle = \langle \mathcal{B}\boldsymbol{y}^{R}, \boldsymbol{y}^{R}\rangle.$$

The second equation in (5.9) implies that

$$\int_\Sigma K_{ijkl}(\chi)\frac{\partial y_i^R}{\partial \xi_j}\frac{\partial v_k}{\partial \xi_l}\,d\hat{\boldsymbol{\xi}} = \int_\Sigma K_{ijkl}(\chi)\mathcal{Z}_{ij}\frac{\partial v_k}{\partial \xi_l}\,d\hat{\boldsymbol{\xi}}, \tag{5.10}$$

for all $\boldsymbol{v} \in H$, where

$$K_{ijkl}(\chi) = \chi\gamma^{-1}\delta_{ij}\delta_{kl} + (1-\chi)G_{ijkl}.$$

As a consequence of this equation, we find

$$\begin{aligned}\overline{G}_{ijkl}\mathcal{Z}_{ij}\mathcal{Z}_{kl} &= K_{ijkl}(\theta)\mathcal{Z}_{ij}\mathcal{Z}_{kl} - \langle \boldsymbol{z}, \mathcal{B}_E^{-1}\boldsymbol{z}\rangle = K_{ijkl}(\theta)\mathcal{Z}_{ij}\mathcal{Z}_{kl} - \langle \mathcal{B}\boldsymbol{y}^R, \boldsymbol{y}^R\rangle \\ &= \int_\Sigma K_{ijkl}(\chi)\Big(\mathcal{Z}_{ij} - \frac{\partial y_i^R}{\partial \xi_j}\Big)\Big(\mathcal{Z}_{kl} - \frac{\partial y_k^R}{\partial \xi_l}\Big)\,d\hat{\boldsymbol{\xi}} \\ &= \int_{\Sigma_F}\gamma^{-1}(\operatorname{tr}\mathcal{Z} - \operatorname{div}\boldsymbol{y}^R)^2\,d\hat{\boldsymbol{\xi}} + \int_{\Sigma_S} G_{ijkl}\big(\mathcal{Z}_{ij} - \mathcal{D}_{ij}(\boldsymbol{y}^R)\big)\big(\mathcal{Z}_{kl} - \mathcal{D}_{kl}(\boldsymbol{y}^R)\big)\,d\hat{\boldsymbol{\xi}}.\end{aligned}$$

Notice that (5.10) is the Euler–Lagrange equation for the functional

$$F_z(\boldsymbol{y}) = \int_\Sigma K_{ijkl}(\chi)\Big(\mathcal{Z}_{ij} - \frac{\partial y_i}{\partial \xi_j}\Big)\Big(\mathcal{Z}_{kl} - \frac{\partial y_k}{\partial \xi_l}\Big)\,d\hat{\boldsymbol{\xi}}.$$

Due to Proposition 4.11 (assertion 4), this functional is strictly convex on $R(\mathcal{B})$ and $\boldsymbol{y}^R$ is its unique minimizer there. That is,

$$\overline{G}_{ijkl}\mathcal{Z}_{ij}\mathcal{Z}_{kl} = F_z(\boldsymbol{y}^R) = \min_{\boldsymbol{y}\in R(\mathcal{B})} F_z(\boldsymbol{y}).$$

Thus, $\overline{G}_{ijkl}\mathcal{Z}_{ij}\mathcal{Z}_{kl} = 0$ if and only if there exists $\boldsymbol{y}^R \in R(\mathcal{B})$ such that $F_z(\boldsymbol{y}^R) = 0$. It is not difficult to see that $F_z(\boldsymbol{y}) = F_z(\boldsymbol{y}+\boldsymbol{w})$ for every $\boldsymbol{w} \in N(\mathcal{B})$. Since $R(\mathcal{B}) \oplus N(\mathcal{B}) = H$, the existence of $\boldsymbol{y}^R \in R(\mathcal{B})$ with $F_z(\boldsymbol{y}^R) = 0$ is equivalent to the existence of a function $\boldsymbol{y} \in H$ which satisfies the condition $F_z(\boldsymbol{y}) = 0$. Due to the positiveness of the functional F_z, we can conclude that $\overline{G}_{ijkl}\mathcal{Z}_{ij}\mathcal{Z}_{kl} = 0$ if and only if there exists $\boldsymbol{y} \in H$ such that

$$\operatorname{div}\boldsymbol{y} = \operatorname{tr}\mathcal{Z} \quad \text{as} \quad \hat{\boldsymbol{\xi}} \in \Sigma_F, \tag{5.11}$$

$$\mathcal{D}(\boldsymbol{y}) = \mathcal{Z} \qquad \text{as} \quad \hat{\boldsymbol{\xi}} \in \Sigma_S. \tag{5.12}$$

Suppose that both of the last conditions are satisfied. Since functions from H do not depend on ξ_3, (5.12) implies that $\mathcal{Z}_{33} = 0$. Moreover, due to (5.12), $\operatorname{div}\boldsymbol{y} = \operatorname{tr}\mathcal{Z}$ in Σ_S. That is, $\operatorname{div}\boldsymbol{y} = \operatorname{tr}\mathcal{Z}$ in Σ. Integrating this equality over Σ we find that $\operatorname{tr}\mathcal{Z} = 0$ because $\boldsymbol{y}$ is periodic. Thus, we have proved the assertion of the theorem in one direction (the necessity).

Let us suppose that $\mathcal{Z}_{11} + \mathcal{Z}_{22} = 0$ and $\mathcal{Z}_{33} = 0$. In order to complete the proof of the theorem, we have to prove that there exists a function $\boldsymbol{y} \in H$ satisfying (5.11) and (5.12). Equation (5.12) is easy to solve. Namely, its solution appears as follows:

$$\boldsymbol{y}(\boldsymbol{\xi}) = \mathcal{Z}\boldsymbol{\xi} + \mathcal{Q}\boldsymbol{\xi} + \boldsymbol{y}_0, \qquad \hat{\boldsymbol{\xi}} \in \Sigma_S,$$

where $\mathcal{Q}$ is a skew-symmetric matrix and $\boldsymbol{y}_0$ is a constant which can be dropped because functions from the space H are defined up to a constant. Let us denote $\mathcal{T} =$

$\mathcal{Z} + \mathcal{Q}$. Since functions from H do not depend on ξ_3, we find that $\mathcal{T}_{i3} = 0\,(i = 1, 2, 3)$ and $y_3 = \mathcal{T}_{31}\xi_1 + \mathcal{T}_{32}\xi_2$ for $\hat{\boldsymbol{\xi}} \in \Sigma_{\scriptscriptstyle S}$. We extend y_3 to the whole domain Σ in such a way that it would be a periodic function (assuming equal values on the opposite edges of Σ).

In order to determine y_1 and y_2 in $\Sigma_{\scriptscriptstyle F}$, we have to solve the problem

$$\begin{aligned} &\frac{\partial y_1}{\partial \xi_1} + \frac{\partial y_2}{\partial \xi_2} = 0, && \hat{\boldsymbol{\xi}} \in \Sigma_{\scriptscriptstyle F}, \\ &\boldsymbol{y}(\hat{\boldsymbol{\xi}}) = \mathcal{T}\boldsymbol{\xi}, && \hat{\boldsymbol{\xi}} \in \partial\Sigma_{\scriptscriptstyle S}, \\ &y_1 \quad \text{and} \quad y_2 \quad \text{are} \quad \text{periodic} \quad \text{in} \quad \Sigma. \end{aligned}$$

This problem is clearly solvable, and the theorem is completely proved. □

As one can see from (5.6), the tensor $\overline{G}$ describes elastic stresses in the homogenized continuum. Theorem 5.21 says that the homogenized material has rather strange properties. Namely, it does not resist to the deformation, if the first invariant and the component (3,3) of the corresponding strain tensor are equal to zero. In other words, such deformations do not produce any stresses. The described class of deformations is sufficiently large. It contains all deformations which do not change volume. The following assertion is a simple consequence of Theorem 5.21.

COROLLARY 5.22. *If $i \neq j$ and $k \neq l$, then $\overline{G}_{ijkl} = 0$.*

This property of the tensor $\overline{G}$ yields an interesting conclusion about the passage to the limit as $\theta \to 0$. If we set $\theta = 0$ formally, the elastic structure will occupy the whole layer Ω^h. Therefore, it can seem that the limiting material must be the same as the original elastic one so that $\lim_{\theta\to 0} \overline{G} = G$. Nevertheless, it is wrong in general because the properties of the tensor $\overline{G}$ stated in Theorem 5.21 and in Corollary 5.22 do not depend on θ. Thus, if, for instance, the tensor G is not degenerate or $G_{1212} \neq 0$, then $\lim_{\theta\to 0} \overline{G} \neq G$. The physical reason is that the elastic structure consists of separate bristles for each $\theta > 0$, which differs from the bulk material corresponding to $\theta = 0$.

6. Numerical procedures. The formulas for the coefficients $\overline{P}$, $\overline{G}$, and ω contain the functions $\boldsymbol{a}_{kl}$, $\boldsymbol{b}_{kl}$, $\boldsymbol{n}_0$, the operators $\mathcal{A}_{\scriptscriptstyle E}$, $\mathcal{B}_{\scriptscriptstyle E}$, and their inverse defined in $H = H^1_{\#}(\Sigma)/\mathbb{R}$. From the mathematical point of view, all these functions and operators are well defined and completely described. However, numerical implementation of these formulas requires some effort. The computation of the functions $\boldsymbol{a}_{kl}$, $\boldsymbol{b}_{kl}$, and $\boldsymbol{n}_0$ is not difficult if one uses the finite element method. The situation with the operators $\mathcal{A}$, $\mathcal{B}$ is not so trivial, because they must be restricted to the subspace E, which creates additional problems when using finite elements. Below, we propose numerical procedures that can be implemented using conventional finite element software.

6.1. Calculation of $\boldsymbol{a_{kl}}$, $\boldsymbol{b_{kl}}$, and $\boldsymbol{n_0}$. Let us introduce functions $\sigma_k \in H$, $k = 1, 2, 3$, as solutions of the following problems:

$$\int_\Sigma \frac{\partial \sigma_k}{\partial \xi_i} \frac{\partial v}{\partial \xi_i}\, d\hat{\boldsymbol{\xi}} = \int_\Sigma \chi\, \frac{\partial v}{\partial \xi_k}\, d\hat{\boldsymbol{\xi}} \qquad \text{for all} \quad v \in H.$$

These problems can be easily solved applying the finite element method. Note that $\sigma_3 = 0$ because functions from the space H do not depend on ξ_3. It is not difficult to see that

$$\begin{aligned} a^i_{kl} &= P_{kli1}\sigma_1 + P_{kli2}\sigma_2, \\ b^i_{kl} &= \gamma^{-1}\delta_{kl}\sigma_i - G_{kli1}\sigma_1 - G_{kli2}\sigma_2, \\ n^i_0 &= -p^0\sigma_i - \mathcal{G}^0_{i1}\sigma_1 - \mathcal{G}^0_{i2}\sigma_2 \end{aligned}$$

for $i,k,l \in \{1,2,3\}$. Here, the superscript i denotes the components of the vectors $\boldsymbol{a}_{kl}$, $\boldsymbol{b}_{kl}$, and $\boldsymbol{n}_0$.

6.2. Calculation of $\mathcal{A}_{E}^{-1}$ and $\mathcal{B}_{E}^{-1}$. The problem can be formulated as follows: *for every* $\boldsymbol{w} \in R(\mathcal{A}) \cap R(\mathcal{B})$, *find* $\boldsymbol{u}, \boldsymbol{v} \in E$ *such that* $\mathcal{A}\boldsymbol{u} = \boldsymbol{w}$ *and* $\mathcal{B}\boldsymbol{v} = \boldsymbol{w}$. It is enough to solve this problem for the operator $\mathcal{A}$. The operator $\mathcal{B}$ can be treated similarly. Let us consider the equation

$$(\mathcal{A} + \varepsilon\mathcal{B})\,\boldsymbol{u}_\varepsilon = \boldsymbol{w}. \tag{6.1}$$

As follows from Proposition 4.13, the operator $\mathcal{A} + \varepsilon\mathcal{B}$ is invertible in H for every $\varepsilon > 0$. Thus, there exists a unique $\boldsymbol{u}_\varepsilon \in H$ that satisfies (6.1). Moreover, $\boldsymbol{u}_\varepsilon \in E$ for every $\varepsilon > 0$ by definition of the subspace E. Equation (6.1) can be easily solved numerically with finite elements. Let us show that $\boldsymbol{u}_\varepsilon$ is an approximation of a function $\boldsymbol{u} \in E$ that satisfies the equation $\mathcal{A}\boldsymbol{u} = \boldsymbol{w}$. Since $\boldsymbol{u}_\varepsilon \in E$, we can rewrite (6.1) as $(\mathcal{A}_{E} + \varepsilon\mathcal{B}_{E})\,\boldsymbol{u}_\varepsilon = \boldsymbol{w}$. Consequently,

$$\boldsymbol{u}_\varepsilon = \mathcal{A}_{E}^{-1}(\boldsymbol{w} - \varepsilon\mathcal{B}_{E}\boldsymbol{u}_\varepsilon). \tag{6.2}$$

Due to Theorem 4.17, the operator $\mathcal{A}_{E}^{-1} : R(\mathcal{A}) \cap R(\mathcal{B}) \to E$ is bounded. Therefore, there exists an independent of ε constant C such that

$$\|\boldsymbol{u}_\varepsilon\| \leqslant C(1 + \varepsilon\|\boldsymbol{u}_\varepsilon\|).$$

This means that the sequence $\{\boldsymbol{u}_\varepsilon\}_\varepsilon$ is weakly compact in H. That is, there exist $\boldsymbol{u} \in H$ and a subsequence $\{\boldsymbol{u}_\varepsilon\}_\varepsilon$ such that $\boldsymbol{u}_\varepsilon \to \boldsymbol{u}$ weakly in H as $\varepsilon \to 0$. Since E is weakly closed, $\boldsymbol{u} \in E$. The passage to the limit in (6.1) yields the desired relation $\mathcal{A}\boldsymbol{u} = \boldsymbol{w}$. Moreover, the whole sequence $\{\boldsymbol{u}_\varepsilon\}_\varepsilon$ converges to $\boldsymbol{u}$ in H. In reality, $\boldsymbol{u}_\varepsilon - \boldsymbol{u} = -\varepsilon\mathcal{A}_{E}^{-1}\mathcal{B}_{E}\boldsymbol{u}_\varepsilon$, which implies

$$\|\boldsymbol{u}_\varepsilon - \boldsymbol{u}\| \leqslant C\varepsilon. \tag{6.3}$$

Thus, the order of the approximation is obtained.

6.3. Calculation of $e^{-t\mathcal{A}_{E}^{-1}\mathcal{B}_{E}}\mathcal{A}_{E}^{-1}\mathcal{B}_{E}$. Problem: *For every* $\boldsymbol{w} \in E$ *and all* $t > 0$, *find* $\boldsymbol{u}(t) = e^{-t\mathcal{A}_{E}^{-1}\mathcal{B}_{E}}\mathcal{A}_{E}^{-1}\mathcal{B}_{E}\boldsymbol{w}$. Let us consider the following equation:

$$\boldsymbol{u}_t + \mathcal{A}_{E}^{-1}\mathcal{B}_{E}\boldsymbol{u} = 0, \qquad \boldsymbol{u}(0) = \mathcal{A}_{E}^{-1}\mathcal{B}_{E}\boldsymbol{w}. \tag{6.4}$$

It is obvious that $\boldsymbol{u}(t) = e^{-t\mathcal{A}_{E}^{-1}\mathcal{B}_{E}}\mathcal{A}_{E}^{-1}\mathcal{B}_{E}\boldsymbol{w} \in E$ is a unique solution of this equation (see section 4.3). Let us construct an approximate solution to problem (6.4) using the semidiscretization method. Fix t and introduce $\tau = t/N$, $N \in \mathbb{N}$. Define functions $\boldsymbol{u}_n$, $n = 1, 2, \ldots, N$, as solutions of the following problem:

$$(\mathcal{A} + \tau\,\mathcal{B})\boldsymbol{u}_n = \mathcal{A}\boldsymbol{u}_{n-1}, \qquad \boldsymbol{u}_0 = \mathcal{A}_{E}^{-1}\mathcal{B}_{E}\boldsymbol{w}. \tag{6.5}$$

That is,

$$\boldsymbol{u}_n = (\mathcal{A} + \tau\,\mathcal{B})^{-1}\mathcal{A}\boldsymbol{u}_{n-1}.$$

Note that $\boldsymbol{u}_n \in E$ for all $n \geq 0$ because $\boldsymbol{u}_0 \in E$ and $\boldsymbol{u}_{n-1} \in E$ implies $\boldsymbol{u}_n \in E$ for any $n > 0$. Therefore, the operators $\mathcal{A}$ and $\mathcal{B}$ can be replaced by $\mathcal{A}_{E}$ and $\mathcal{B}_{E}$ in (6.5).

To estimate $\boldsymbol{u}_N - \boldsymbol{u}(t)$ we prove first that there exists a constant C such that $\|\boldsymbol{u}_n\| \leqslant C$ for all n. Indeed,

$$\boldsymbol{u}_n = \boldsymbol{u}_0 - \tau \mathcal{A}_{E}^{-1} \mathcal{B}_{E} \sum_{k=1}^{n} \boldsymbol{u}_k,$$

which implies due to the boundness of $\mathcal{A}_{E}^{-1}$ and $\mathcal{B}_{E}^{-1}$ (see Theorem 4.17), the existence of a constant C such that

$$\|\boldsymbol{u}_n\| \leqslant \|\boldsymbol{u}_0\| + C\,\tau \sum_{k=1}^{n} \|\boldsymbol{u}_k\|.$$

The Gronwall inequality now implies the required estimate. The boundness of $\|\boldsymbol{u}_n\|$ and (6.5) provide the following estimate:

$$\|\boldsymbol{u}_n - \boldsymbol{u}_{n-1}\| \leqslant C\tau\, \|\mathcal{A}_{E}^{-1}\mathcal{B}_{E}\| \|\boldsymbol{u}_n\| \leqslant C\tau. \tag{6.6}$$

Let us introduce two time interpolations of $\{u_n\}$,

$$\hat{\boldsymbol{u}}^{\tau}(s) = \boldsymbol{u}_n\Big(1 - n + \frac{s}{\tau}\Big) + \boldsymbol{u}_{n-1}\Big(n - \frac{s}{\tau}\Big) \qquad \text{as} \quad s \in [(n-1)\tau, n\tau],$$
$$\bar{\boldsymbol{u}}^{\tau}(s) = \boldsymbol{u}_n \qquad \text{as} \quad s \in ((n-1)\tau, n\tau].$$

Due to (6.6),

$$\int_0^t \|\hat{\boldsymbol{u}}^{\tau}(s) - \bar{\boldsymbol{u}}^{\tau}(s)\|\, ds = \sum_{n=1}^{N} \|\boldsymbol{u}_n - \boldsymbol{u}_{n-1}\| \int_{(n-1)\tau}^{n\tau} \Big(n - \frac{s}{\tau}\Big)\, ds$$
$$= \frac{\tau}{2} \sum_{n=1}^{N} \|\boldsymbol{u}_n - \boldsymbol{u}_{n-1}\| \leqslant C\tau.$$

Moreover, we have

$$\frac{\partial \hat{\boldsymbol{u}}^{\tau}}{\partial t} + \mathcal{A}_{E}^{-1}\mathcal{B}_{E}\bar{\boldsymbol{u}}^{\tau} = 0.$$

Therefore,

$$\|\boldsymbol{u}_N - \boldsymbol{u}(t)\| = \|\hat{\boldsymbol{u}}^{\tau}(t) - \boldsymbol{u}(t)\| \leqslant \|\mathcal{A}_{E}^{-1}\mathcal{B}_{E}\| \int_0^t \|\bar{\boldsymbol{u}}^{\tau}(s) - \boldsymbol{u}(s)\|\, ds$$
$$\leqslant C \int_0^t \big(\|\hat{\boldsymbol{u}}^{\tau}(s) - \boldsymbol{u}(s)\| + \|\hat{\boldsymbol{u}}^{\tau}(s) - \bar{\boldsymbol{u}}^{\tau}(s)\|\big)\, ds \leqslant C\Big(\tau + \int_0^t \|\hat{\boldsymbol{u}}^{\tau}(s) - \boldsymbol{u}(s)\|\, ds\Big).$$

Finally, the Gronwall inequality yields

$$\|\boldsymbol{u}_N - \boldsymbol{u}(t)\| \leqslant C\tau.$$

6.4. Numerics. In this section we give some examples which demonstrate properties of the homogenized continuum for various values of θ. We consider the system consisting of the water and an isotropic elastic material (polymer) with the following properties:

$$P\boldsymbol{u}_{\boldsymbol{x}} = \lambda_f\, \mathcal{I}\, \mathrm{div}\, \boldsymbol{u} + 2\,\mu_f\, \mathcal{D}(\boldsymbol{u}), \qquad G\boldsymbol{u}_{\boldsymbol{x}} = \lambda_s\, \mathcal{I}\, \mathrm{div}\, \boldsymbol{u} + 2\,\mu_s\, \mathcal{D}(\boldsymbol{u}),$$
$$\lambda_f = 1.\mathsf{e}{-3}, \quad \mu_f = 1.\mathsf{e}{-3},$$
$$\lambda_s = 2.777778\mathsf{e}+9, \quad \mu_s = 4.166667\mathsf{e}+9.$$

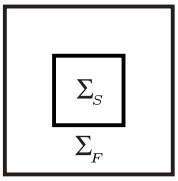


FIG. 3. *Structural cell* $\Sigma = [0,1] \times [0,1]$.

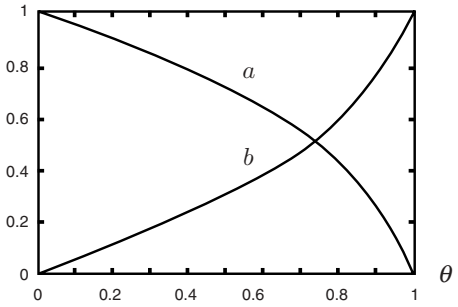


FIG. 4. *Dependence of* $|P - \overline{P}|/|P|$ *(curve a) and* $|\overline{P}|/|P|$ *(curve b) on* θ.

The constant which characterizes compressibility of the water is $\gamma = 4.597696\mathrm{e} - 10$. We take the structural cell of the form shown in Figure 3.

First, we investigate properties of the tensor $\overline{P}$. The graphics in Figure 4 present the dependence of $|P - \overline{P}|/|P|$ and $|\overline{P}|/|P|$ on θ, where $|P| = \left(\sum_{ijkl} P_{ijkl} P_{ijkl}\right)^{1/2}$. As one can see, $\lim_{\theta \to 1} \overline{P} = P$ and $\lim_{\theta \to 0} \overline{P} = 0$.

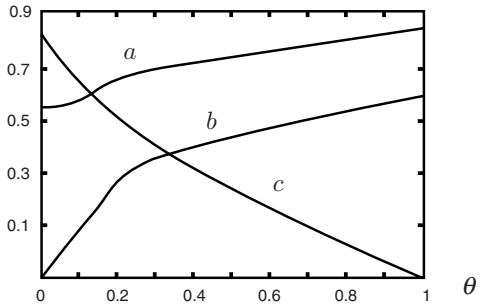


FIG. 5. *Dependence of* $|\overline{G} - G|/|G|$ *(curve a),* $|\overline{G} - Q|/|G|$ *(curve b), and* $|\overline{G} - \theta R|/|G|$ *(curve c) on* θ.

The dependence of the tensor $\overline{G}$ on θ is more complex. Let us introduce two tensors R and Q having the following components:

$$R_{ijkl} = \gamma^{-1} \delta_{ij} \delta_{kl}, \qquad Q_{ijkl} = \begin{cases} G_{ijkl} & \text{if} \quad i = j \quad \text{and} \quad k = l, \\ 0 & \text{otherwise.} \end{cases}$$

The curves in Figure 5 show the dependence of $|\overline{G} - G|/|G|$, $|\overline{G} - \theta R|/|G|$ and $|\overline{G} - Q|/|G|$ on θ. One can see that $\overline{G}$ does not tend to G as $\theta \to 0$ (see curve a). This fact was already noted in section 5.2 after Corollary 5.22. It is not surprising that

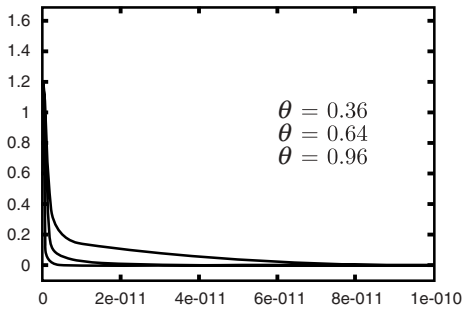


FIG. 6. *Graphics of the function* $|\omega(t)|/|\overline{G}|$ *for various values of* θ. *The time unit is* 1 *second.*

$\lim_{\theta\to 1}\overline{G}\neq 0$. It can be explained by the presence of the tensor R in the definition of $\overline{G}$. Really, curve c in Figure 5 shows that $\lim_{\theta\to 1}(\overline{G}-R)=0$. Curve b shows that $\lim_{\theta\to 0}\overline{G}=Q$. This means that, in the limit case as $\theta\to 0$ (the elastic continuum occupies the whole domain Ω^h), the tensor $\overline{G}$ can be obtained from the tensor G by vanishing all of the "nondiagonal" components.

Thus, the limits of (5.6) as $\theta\to 0$ and $\theta\to 1$ look as follows:

$$\rho_{\scriptscriptstyle S}\boldsymbol{u}_t-\operatorname{div}\mathcal{J}_tQ\boldsymbol{u}_x-\operatorname{div}\int_0^t\omega(t-s)\boldsymbol{u}_x(s)\,ds=\rho_{\scriptscriptstyle S}\boldsymbol{f},\qquad(\theta\to 0),\tag{6.7}$$

$$\rho_{\scriptscriptstyle F}\boldsymbol{u}_t-\operatorname{div}P\boldsymbol{u}_x-\gamma^{-1}\nabla\operatorname{div}\boldsymbol{u}-\operatorname{div}\int_0^t\omega(t-s)\boldsymbol{u}_x(s)\,ds=\rho_{\scriptscriptstyle F}\boldsymbol{f},\quad(\theta\to 1).\tag{6.8}$$

We take here the initial data being equal to zero.

In Figure 6, the graphs of the function $|\omega(t)|/|\overline{G}|$ are presented for several values of θ. The function $|\omega(t)|$ decreases very rapidly. In fact, $|\omega(t)|/|\overline{G}|$ vanishes practically at the time $t\sim 10^{-10}\,s$. Thus, the memory term in (5.6) is very small and can be dropped in applications, if high frequency oscillations are not present.

7. Conclusions. Homogenization of a fine elastic structure immersed in a viscous weakly compressible fluid yields a continuum that possesses very interesting and rather unexpected properties. Equation (5.6) describing the behavior of the resulting continuum includes three basic terms. Two of them containing the tensors $\overline{P}$ and $\overline{G}$ are related to stresses. The third integral-term represents a memory effect that is responsible for viscoelastic properties of the resulting material. The presence of such a memory is not surprising because similar results were already obtained by other authors (see, for instance, [13]). More interesting from the mathematical and mechanical viewpoints is the investigation of the above mentioned stress terms. The term containing the tensor $\overline{P}$ describes a viscous damping and originates from the fluid part of the structure. Theorem 5.20 states the strict positiveness of $\overline{P}$, which implies the ellipticity of the corresponding differential operator. The term containing the tensor $\overline{G}$ represents elastic stresses. The tensor $\overline{G}$ is degenerate, its kernel is described in Theorem 5.21. The theorem implies that volume conserving deformations (shear deformations in particular) do not produce elastic stresses.

All of the coefficients involved in the homogenized equation are found in an explicit form. Although expressions representing them are rather complex, the coefficients can be computed numerically using the algorithms given in section 6. The numerical treatment delivers another interesting properties of the homogenized model. It is stated

numerically that the memory effect is very weak. The system "forgets" the current history in a very short time. Therefore, the memory term can be dropped in most of the applications. Another interesting question is the dependence of properties of the homogenized continuum on the parameter θ which represents the volume fracture of the fluid so that the pure fluid corresponds to $\theta = 1$. As was expected, the homogenized equations coincide in the limit ($\theta \to 1$) with the ones being used for the description of the original fluid. In the opposite limiting case ($\theta \to 0$) the homogenized equations differ from the model of the original elastic continuum. In particular, the limiting elastic continuum can be nonisotropic even though the original material is isotropic. Thus, the limiting continuum inherits certain geometric properties of the fine elastic structure even if the fluid vanishes and the solid occupies the whole volume.

REFERENCES

[1] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[2] L. Baffico, C. Conca, *A mixing procedure of two viscous fluids using some homogenization tools*, Comput. Methods Appl. Mech. Engrg. 190 (2001), pp. 4245–4257.

[3] N. Botkin, M. Schlensog, M. Tewes, and V. Turova, *A mathematical model of a biosensor*, in Technical Proceedings of the Fourth International Conference on Modeling and Simulation of Microsystems, Hilton Head Island, SC, 2001, 231–234.

[4] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes equations*, Vol. 1. *Linearized Steady Problems*, Springer Tracts in Natural Philosophy 38, Springer-Verlag, New York, 1994.

[5] K.-H. Hoffmann and N. D. Botkin, *Homogenization of von Kármán plates excited by piezoelectric patches*, ZAMM Z. Angew. Math. Mech., 80 (2000), pp. 579–590.

[6] K.-H. Hoffmann and V. N. Starovoitov, *On a motion of a solid body in a viscous fluid. Two-dimensional case*, Adv. Math. Sci. Appl., 9 (1999), pp. 633–648.

[7] U. Hornung, *Homogenization and Porous Media*, Springer-Verlag, New York, 1997.

[8] W. Jäger and A. Mikelić, *On the roughness-induced effective boundary conditions for an incompressible viscous flow*, J. Differential Equations, 170 (2001), pp. 96–122.

[9] L. Landau and E. Lifshits., *Lehrbuch der theoretischen Physik*, Band VI: *Hydrodynamik*, 5th ed., Akademie-Verlag, Berlin, 1991.

[10] J.-L. Lions, *Quelques méthodes de résolution des problémes aux limites non linéaires*, Etudes mathematiques, Gauthier-Villars, Paris, 1969.

[11] D. Lukkassen, G. Nguetseng, and P. Wall, *Two–scale convergence*, Int. J. Pure Appl. Math., 2 (2002), pp. 35–86.

[12] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.

[13] J. Sanchez-Hubert, *Asymptotic study of the macroscopic behaviour of a solid-fluid mixture*, Math. Methods Appl. Sci., 2 (1980), pp. 1–11.

[14] R. Temam, *Problèmes mathématiques en plasticité*, Méthodes Mathématiques de l'Informatique 12, Gauthier-Villars, Montrouge, 1983.

SIAM J. APPL. MATH.
Vol. 65, No. 3, pp. 1006–1026

MODELING CALCIUM DYNAMICS IN DENDRITIC SPINES*

D. HOLCMAN†‡ AND Z. SCHUSS§

Abstract. Dendritic spines are microstructures located on dendrites of neurons, where calcium can be compartmentalized. They are usually the postsynaptic parts of synapses and may contain anywhere from a few up to thousands of calcium ions at a time. Initiated by an action potential, a back-propagating action potential, or a synaptic stimulation, calcium ions enter spines and are known to bring about their fast contractions (twitching), which in turn affect calcium dynamics. In this paper, we propose a coarse-grained reaction-diffusion (RD) model of a Langevin simulation of calcium dynamics with twitching and relate the biochemical changes induced by calcium to structural changes occurring at the spine level. The RD equations model the contraction of proteins as chemical events and serve to describe how changes in spine structure affect calcium signaling. Calcium ions induce contraction of actin-myosin-type proteins and produce a flow of the cytoplasmic fluid in the direction of the dendritic shaft, thus speeding up the time course of calcium dynamics in the spine, relative to pure diffusion. Experimental and simulation results reveal two time periods in spine calcium dynamics. Simulations [D. Holcman, Z. Schuss, and E. Korkotian, *Biophysical Journal*, 87 (2004), pp. 81–91] show that in the first period, calcium motion is mainly driven by the hydrodynamics, while in the second period it is diffusion. The coarse-grained RD model also gives this result, and the analysis reveals how the two time constants depend on spine geometry. The model's prediction, that there are not two time periods in the diffusion of inert molecules in the spine, has been verified experimentally.

Key words. modeling microstructures, reaction-diffusion equations, dendritic spines, calcium, stochastic dynamics

AMS subject classifications. 92C05, 92C17, 35K57

DOI. 10.1137/S003613990342894X

1. Dendritic spines and their function.

Dendritic spines are microstructures, about 1 μm across, made of a head and connected by a cylindrical neck to the dendrite. Although discovered more than 100 years ago by Ramón y Cajal [1] on dendrites of most neurons, including cortical pyramidal neurons and cerebellar Purkinje cells [2], their function is still unclear. The current consensus is that the main function of dendritic spines is to compartmentalize calcium [3]. Regulated by synaptic activity, spines are constantly moving and changing shape [4]. The 100,000 to 300,000 spines on a single spiny neuron drastically increase the active surface of a dendrite [5], [6], and more than 90% of excitatory synapses terminate on dendritic spines. Spines are considered to be basic units of dendritic integration [7], [8], though their role and function are still unclear. There is evidence that morphological changes in spines are associated with synaptic plasticity [4], that is, with the structural and biochemical changes in spines, dendrites, and neuronal synapses.

*Received by the editors June 4, 2003; accepted for publication (in revised form) June 15, 2004; published electronically March 31, 2005.
http://www.siam.org/journals/siap/65-3/42894.html
†Department of Mathematics, Weizmann Institute of Science, Rehovot 76100 Israel (holcman@wisdom.weizmann.ac.il). The research of this author was supported by the Sloan Foundation and the Swartz Foundation.
‡Keck Center, Department of Physiology, UCSF, 513 Parnassus Ave, San Francisco, CA 94143 (holcman@phy.ucsf.edu).
§Department of Mathematics, Tel-Aviv University, Tel-Aviv 69978, Israel (schuss@post.tau.ac.il). The research of this author was partially supported by grants from the US-Israel Bi-National Science Foundations, the Israel Science Foundations, and DARPA.

A debate is still raging about the specific function of dendritic spines. In particular, two main views prevail [9], [10]. The first maintains that a dendritic spine constitutes a privileged location for calcium restriction, and consequently, it is a place where synaptic plasticity can be induced. Calcium in dendritic spines triggers changes, such as long-term potentiation (LTP) and long-term depression (LTD) [11], which result in a permanent modification of the synaptic weight. Indeed, calcium dynamics, defined as the rise and duration of concentration inside a dendritic spine, is believed to be determinant for the nature of spine synaptic plasticity. These processes constitute the implementation of some of the memory in the brain at the cellular and subcellular levels. The second view maintains that by changing the shape of the spine, the electrical characteristics of the spine change, thereby modulating the voltage and the depolarization of the dendrite. This way the spines participate in the dendritic computation process.

Recently, it has been observed [12] that after calcium ions flow in, a dendritic spine can change shape in a few hundreds milliseconds. This fast change of shape decreases the volume of the spine head. Spine motility was proposed by Blomberg, Cohen, and Siekevitz [13] and the fast twitching movement of the spine was anticipated by Crick in [14], where questions were asked about the rules "governing the change of shape of the spine and, in particular, the neck of the spine" and also on "how these rules are implemented in molecular terms." Many models of calcium dynamics in dendritic spines have been proposed in the literature [5], [7], [15], [16], [17], [18], [19]; however, calcium dynamics was not considered in conjunction with Crick's questions and with the observations of [12].

When the spine shape is described by a spherical head connected to a cylindrical neck, several classes of shapes can be distinguished, according to three independent geometrical parameters (see Figure 1). According to this representation, the three parameters are the radius of the head (R), the length of the neck (l), and its diameter (d). There are at most eight possible classes of spines, according to the relative sizes (large or small) of the three parameters. It is not clear yet what are the rules, if any, of the distribution of the different classes in a given neuron. Spines may appear isolated or in clusters on a dendrite [4]. The number of spines and their distribution are regulated by neuronal activity, because increased activity tends to increase the production of spines, whereas light deprivation tends to reduce the number of spines. However, the details remain unclear.

Dendritic spines can change shape on various time scales. On the time scale of minutes, synaptic stimulation can generate new spines. LTP experiments in the dentate gyrus are correlated with a change in the diameter of the spine neck. A single spine can split into two, and transitions between filopodia (spines with no head) and the standard form have been observed experimentally [4]. Modulation of sensory inputs, such as monocular deprivation in specific periods of development, modulates spine motility [20], [21]. Spines are less motile in adult neurons than in neurons of juvenile animals. Changes of shape on the time scale of minutes are due to actin (de-)polymerization and can be induced by a variation in the concentration of a neurotransmitter, such as glutamate [12], [23]. Dendritic spines have been observed to move on a very short time scale. For example, vibrations along the spine axis, which are independent of the calcium concentration, occur on the time scale of tens of milliseconds [12].

Spine movements on the time scale of seconds have been observed directly by recent imaging techniques, such as confocal microscopy or two-photon microscopy. It was reported that spines are constantly changing shape [18], [24]. This motility is also

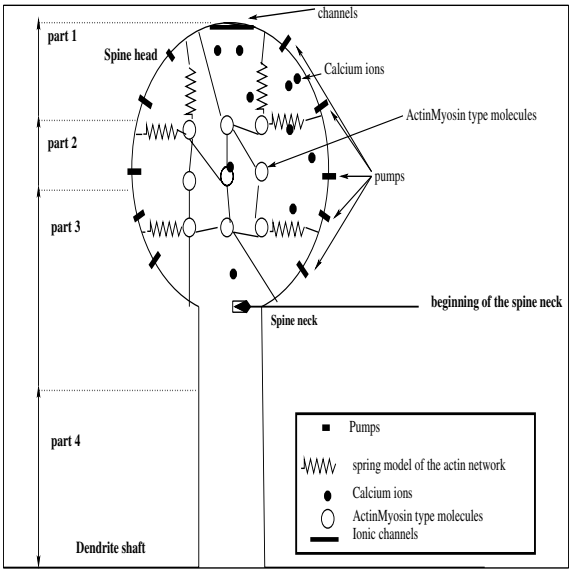


FIG. 1. *Schematic description of a dendritic spine. A dendritic spine is modelled as a spherical head connected to the dendrite by a cylindrical neck. The surface of the head contains a postsynaptic density, where various types of protein channels, such as the glutamate receptors NMDA and AMPA, are anchored and conduct ions into the spine when opened. A spine contains signalling molecules, such as calmodulin; cytoskeletal proteins, such as actin-myosin; and organelles, such as smooth endoplasmic reticulum. Pumps are located on the side and channels at the top of the spine head. Actin-myosin sites are represented schematically, attached to the actin network, so that a contraction of a single protein affects the entire spine. (Figure reprinted from* [D. Holcman, Z. Schuss, and E. Korkotion, *Calcium dynamics in dendritic spines and spine motility,* Biophysical Journal, 87 (2004), pp. 81–91] *with permission.*

an actin-dependent process. The postsynaptic current can be modified by affecting the spine geometry, thus modulating the synapse. Changes in spine shape can then affect the efficacy of calcium dynamics. Specifically, it was reported in [18] that changing the spine neck affects the time course of calcium dynamics: High calcium concentration is maintained for a shorter period of time when the neck is shorter. Thus dendritic spines with shorter necks are less efficient in compartmentalizing calcium. In summary, spines undergo a constant readjustment, which can be viewed as an intrinsic spine property [20], and motility possibly contributes to synaptic plasticity.

It has recently been observed [12] that a dendritic spine can change shape quickly, on the scale of a few hundreds of milliseconds, after calcium ions flow into the spine. Transient calcium causes the spine to twitch (see [12]). This quick change of shape consists in head contractions oriented on the average in the direction of the dendritic shaft. This contraction can be induced by agonists, such as a neurotransmitter, or by a back-propagating action potential.

Evidence of high concentration of actin and myosin in spines was reported in [25], where proteins were observed to form clusters inside the spine head, near the channels. These clusters are called the *postsynaptic density* (PSD). More uniform distributions of clusters of myosin molecules were also observed. As in muscle cells, high concentrations of actin molecules indicate that rapid movement can be ascribed to the contraction of these molecules, because blocking them prevents all shape fluctuations [12].

It is remarkable that the description of the diffusive motion of ions in spines can be considered in the intermediate regime between continuum and discrete. Due to its specific geometry, the dendritic spine can be studied as a separate unit from the remaining part of the dendrite. Chemical reactions in the spine involve only a small number of molecules (hundreds), which explains the relatively large fluctuations in the reactions. These may lead to synaptic plasticity. This fact also reinforces the idea that the spine has a major role in converting a random signal, carried by the motion of ions or secondary messengers, into a more deterministic, less fluctuating, and more stable variable, represented by the synaptic weight.

2. Introduction. Despite the rapid development of high-quality technology, today's biophysical analysis of calcium in dendritic spines is limited by the resolution of the instruments. Thus models become useful tools for the analysis and prediction of spine activity, based on the evidence of molecular chemical reactions.

2.1. Modeling the dendritic spine dynamics. We propose here an answer to Crick's question about the cause and effect of the twitching and its role in the functioning of the spine as a conductor of calcium. Specifically, we attribute the twitching motion to the contraction of actin-myosin-type proteins, denoted AM, when they bind calcium, and include its effect on the dynamics of the calcium ions in the spine. This is the first quantitative theoretical and mathematical treatment of the twitching and its role in calcium dynamics in the spine. In [26] we constructed a Langevin dynamics simulation of calcium dynamics in the spine, and here we propose a continuum model of the same.

The calmodulin proteins (CaM) can bind up to four calcium ions to form the complex $CaMCa_4$. This complex starts other important chemical reactions, involving, for example, calmodulin-protein kinase-II. This kinase plays a crucial role in LTP induction [16]. When a sufficiently large number of $CaMCa_4$ complexes are formed, it produces LTP changes and/or induces dephosphorylation and (de-)polymerizations. It can also affect certain biophysical properties of certain channels, such as N-methyl-D-aspartale (NMDA) receptors. More generally, this type of reaction is known to induce modifications in the spine shape and biophysical changes at various levels, such as synaptic modifications, and changes in the number of channels (see [17]): When channel subunits are modified, the selectivity and/or ionic conductivity is changed, affecting the number of ions that enter the spine. When the number of receptors increases, e.g., of AMPA receptors, the spine's depolarization increases, resulting in a higher probability of opening of NMDA receptors and thus increasing the total number of calcium ions entering the spine.

We model the spine as a machine powered by the calcium it conducts, and we describe here the induced movement. Proteins involved in the calcium conduction process are found inside the dendritic spine. Their spatial distribution was reported in [25]. As mentioned above, relevant proteins involved in spine motility include actin, which has been shown in [12], [20], [24] to be directly involved in the biophysical process underlying fast spine motility. We maintain here that AM sites are driving the motility events. It was shown in [25] that dendritic spines contain a network of myosin molecules. The spatial distribution of myosin molecule in the spine has been observed to be uniform and to be sparse inside the PSD.

From a biological point of view, it is of primary interest to answer two related questions about calcium dynamics in dendritic spines, after their channels open: (1) How much calcium is there inside the spine? (2) How long does a given quantity of calcium stay inside the spine? Obviously, the answers depend on the geometry of the

cell. In this context, the aim of our model is to reproduce the time course of events, such as calcium dynamics, which determine the transition between depression and facilitation, or long- and short-term depression (see [28]). We propose that calcium ions set the machine in motion by initiating the contraction of AM as they bind at active sites [26]. We elucidate the cause and effect of twitching in the functioning of the spine by adding up the local contractions of the separate calcium-saturated proteins to achieve a global contraction effect. The contraction of the spine head induces a flow field of the cytoplasmic fluid, which in turn pushes the ions, thus speeding up their movement in the spine.

2.2. Biological consequences. We reported and discussed the biological consequences of a Langevin simulation, designed at a molecular level, in our first paper [26]. The purpose of the simulation was to investigate at a molecular level biochemical events induced by calcium and thus to explain structural changes occurring at the spine level. The main biological conclusion of [26] concerns the quantification of the effect of the hydrodynamical push on calcium dynamics in the spine. In particular, we showed not only that the push effect is created by the calcium ions, but that the push targets the same calcium ions towards the dendrite and in the direction of the center of the spine, where the spine apparatus and other relevant proteins are located. The flow due to the push does not allow the calcium ions to stay inside the spine head and to return to the head once they are inside the neck. The drift increases the efficiency of calcium conduction from the synapse to the dendrite and speeds up the calcium clearance of the spine. The simulations of [26] show that in the absence of the drift effect, the proportion of calcium ions conducted to the dendrite is two to three times smaller than in its presence. This led to the prediction that there are not two time periods in the diffusion of inert dye molecules in the spine, as has been recently verified experimentally [27].

We propose here a coarse-grained description of the coupling between changes in spine structure and calcium dynamics. A set of nonlinear reaction-diffusion equations is derived from the Langevin description. The analysis of the model reveals the time scale of the hydrodynamical effect and leads to the calculation of the time constant of the first concentration decay period. Consequently the push effect offers a possible reinterpretation of the results of [19], about the double exponential decay of the calcium concentration inside the spine. The first decay period was reported in [19] to be the consequence of buffered calcium diffusion from spine to dendrite and the effect of calcium pumps in the spine head. According to [19], the second period starts when near equilibrium is achieved between the spine's and the dendrite's calcium concentrations. We show in this paper that the two time periods of calcium concentration decay are recovered, under specific conditions, when the hydrodynamical push effect is included. We observe that the decay, corresponding to a predominantly hydrodynamical effect, starts immediately after the ions enter the spine head. This decay is rapid and its duration is random. It ends when hardly any saturated contractive molecules are left. The ionic motion in the second period is mainly pure diffusion and pump extrusion. An analytic expression for the fast decay rate is derived from the model in terms of the average hydrodynamical flow velocity. The main biological result of our model is that the rapid spine movement produces fast clearance of calcium from the dendritic spine and directs calcium ions to a specific location between the neck and the dendritic shaft, preventing the pumping out of the majority of ions. As mentioned above, the model also predicts that there are not two time periods in the diffusion of inert dye molecules in the spine, which means that diffusion alone cannot

be responsible for the double exponential decay.

2.3. The need of a molecular approach. The models of calcium diffusion in dendritic spines used, e.g., in [5], [6], [7], [15], [16], [17], [18], [19], [22] are based on a phenomenological approach that uses the coupling between the diffusion equation and the ambient chemical reactions. They are based on compartmentalization of the spine into several subunits, where the calcium diffusion process is discretized, while ordinary differential equations describe the chemical bonding of calcium to buffer protein molecules.

In this paper, we present a mathematical model of calcium dynamics in dendritic spines, based on molecular-level considerations. Actually, we propose a unified approach to modeling calcium dynamics inside microstructures, including dendritic spines, that postulates Brownian motion of the calcium ions in the cell. The randomness of the ionic motion becomes significant when the number of ions in the cell is small. At the molecular level, all phenomena, beginning with the motion of a single ion and up to the dynamics of the entire ionic population, are stochastic processes. These include the random walk of an ion, forming or breaking bonds with proteins by an individual ion. On the entire calcium population level, they include the dynamics of the number of bound proteins, which depends on the trajectory of each ion, and the distribution of the protein molecules.

Our model begins with the description of the dynamics of individual calcium ions in terms of a system of Langevin equations. The collective effect inside the spine of the entire calcium population, due to the interaction between the calcium and the proteins, is captured in our model by a nonlinear drift term that couples the hydrodynamical flow field to the number of ions bound to certain proteins. This produces a new effect that has to be included in the diffusion equation. The distribution of proteins inside the cell becomes an important part of the model.

2.4. Biological simplifications of the model. We make several simplifications in constructing the model of the spine. Thus, we neglect other types of organelles that are also involved in calcium dynamics: the spine apparatus, mitochondria, and other types of proteins. We have included a low concentration (.5–1 μM) of binding molecules such as calcineurin. However, at this concentration these molecules cannot capture fully the role played by the buffer activity. The present model ignores the effect of a large buffer regulation, but we keep in mind that it can affect the calcium dynamics.

Furthermore, it is known that calcium stores in the spine release calcium ions when prompted by external calcium ions, under specific conditions. We neglect this effect here to avoid complicating our model. We also restrict the biochemical structure of the spine by singling out the CaM, AM, calcineurin, and one type of calcium pump. All these proteins constrain calcium flow in the dendritic spine by binding calcium ions for random periods of time. The technical assumption in our model is that the AM proteins contract at a fixed rate as long as they keep four calcium ions bound. Thus contraction begins and ends at random times. Since we are interested in the dynamics of calcium, when the ions are already inside the spine, we avoid the computation of the transient time starting from the action potential and the opening of the voltage-sensitive calcium channels.

The specific geometry of the spine needs to be considered in order to evaluate the time evolution of calcium concentration in the spine. In the present simplified model, the spine geometry has been described by three parameters: the length and diameter

of the spine neck and the radius of the spine head (see Figure 1), smoothing out the local irregularities of the boundary.

Another geometric feature is the distribution $S_0(\boldsymbol{x})$ of calcium-dependent molecules that contract when they bind enough calcium. Two extreme possible distributions of proteins have been considered in the simulation, reported in [26]: a uniform distribution inside the head or an accumulation at the PSD area and the simulations show that the calcium dynamics depend on the distributions of the proteins. In reality, a mixture of the two distributions is observed in [26], but we will ignore it in the derivation of the decay rate.

3. A simplified physical model of the spine. The two main components of the dendritic spine in our model are a spherical head and a cylindrical neck, which connects it to the dendritic shaft. On top of the spine head, opposite the neck, there are protein channels that conduct calcium into the spine head. These channels can be of two types: NMDA channels (opened by the glutamate neurotransmitter) and calcium channels, which are voltage sensitive. There are only 2–5 NMDA channels open at a time. For the purpose of this model, we use only the location of these channels as the initial positions for the ions. Our model concerns times after the calcium ions have entered the spine head. A schematic figure of the spine is presented in Figure 1. Active pumps are located on the lower half of the spine head. Their role is to conduct calcium out of the spine head. Pumping is an active process that requires energy, provided by the adenosine tri phosphate (ATP) molecules, whereas when calcium enters through the channels, no extra molecular energy is needed. We assume that there is only one ion at a time inside a pump and, due to the active structure, it requires a certain time to be pumped out. This time can be assumed random or deterministic. The latter case is valid when the exit time distribution is concentrated around the mean value. In a coarse-grained continuum model the pumping time is neglected, so the part of the boundary occupied by pumps becomes an absorbing boundary.

The many organelles inside the spine head do not affect the nature of the random motion of ions, mainly due to their large size relative to that of ions. They only restrict the volume available for free diffusion of calcium. Neglecting their presence effectively frees the interior of the spine head from obstacles to ionic movement. This can be compensated for by decreasing the radius of the head. The incompressible cytoplasmic fluid that fills out the spine and its flow are a part of our model.

3.1. A schematic model of spine twitching. Once calcium ions enter the spine they reach AM binding sites by diffusion and can bind there. When four calcium ions bind to a single AM protein, a local contraction of the protein occurs. All the local contractions at a given time produce a global contraction and induce a hydrodynamical movement of the cytoplasmic fluid. Calcium ions can reach the dendritic shaft through the spine neck and be totally absorbed there, or they can be pumped out of the spine by active pumps. Our model allows us to calculate the fraction of ions that are pumped out, relative to those that reach the dendrite. At a molecular level, in a phenomenological approach, a contraction produced by AM occurs with a characteristic time $t_c = 1$ ms, say, and the length fluctuates about $l_c = 0.02$ μm [2, p. 681], depending only on the type of protein. In a homogenization approximation of the spine head, the result of this local contraction produces a fixed average velocity of the cytoplasmic fluid of the order $v_Q = l_c/t_c = 0.02$ μm/ms.

Since it is known that there are only a few AM binding sites (less than a hundred [25]), each binding event can modify the dynamics significantly. It is important

therefore to keep track of the number of bound ions at any given time. Both the distribution of AM binding sites and the binding times are random. Consequently, the twitching of the spine head is also random. This, in turn, implies that the evolution of calcium concentration inside the spine is random. In a continuum description of this process, only average motion is observed, so the random realizations that can be observed in molecular simulations are smoothed out.

3.2. Final model simplifications. We simplify the model further by neglecting the long range ion-ion electrostatic interactions, as well as the ion-protein interactions. At a molecular level, when 500 calcium ions enter the dendritic spine, they create a difference of potential of about 16 mV (compared to -70 mV of the cell potential), so there are enough negative ions inside the spine to electrostatically neutralize the calcium ions. Specifically, the cooperative effect of the ions creates dipoles that screen the long-range interaction forces (r^{-2}) to short-range interactions. The shield around each ion is a basis for an approximation that neglects the electrostatic forces in order to study the dynamics of calcium ions inside the spine. In this approximation the trajectories of the calcium ions are independent. The ion-water interaction is simplified into hydrodynamical drag and a zero mean fluctuating force that describes the randomness of the water-ion collisions [29]. The ion-protein interaction near a binding site, where a high electrical field targets the ions toward the active center of a binding site, is represented by a short range parabolic potential well. This allows us to include the backward binding reaction constant in the model. The effect of the forward constant is discussed below.

Each time an ion nears an active neighborhood of a protein, we assume that the electrostatic forces direct the ion so that a bond is formed with a given probability, depending on the forward rate constant. The backward reaction rate is the reciprocal of the mean time an ion stays bound. A binding site that holds an ion cannot bind additional ions before the bound ion escapes. We say that a protein is saturated if each of the four binding sites contains a calcium ion at the same time.

The chemical kinetics of the binding and unbinding of calcium to and from the substrate proteins (CaM, AM, calcineurin) in the spine cannot be described by the usual Arrhenius kinetics because of the small number of the reactant particles, the large fluctuations in the number of bound ions, and the hydrodynamic effect on the binding and unbinding reactions. We describe the forward and backward reactions on a molecular level in section 5 and then coarse-grain the equations in section 6. We consider in our model only two classes of binding proteins: one that includes CaM and AM (that is, proteins that can bind 4 calcium ions) and a second that includes calcineurin, which can bind only one calcium ion at a time. The simplified model described above was used for a molecular simulation of calcium dynamics in a dendritic spine in [26].

4. The mathematical model. The mathematical model of the simplified physical model of the dendritic spine has several components. First, the domain Ω_t, available for the motion of an ion at time t, has quite a complicated geometry, due to the presence of many obstacles, as mentioned in section 3, and it may change in time. Actually, this change is one of the main phenomena captured by our model. Second, when Ω_t changes, a flow of the cytoplasmic fluid in the dendritic spine ensues, which in turn gives rise to an hydrodynamical drag force on the ions inside the dendritic spine. This drag is a frictional force proportional to the relative velocity between the ions and the fluid. This force is not neglected in our simplified model. Third, the mathematical expression of these assumptions is a Langevin model of the ionic

motion. That is, the motion is described by a system of identical uncoupled Langevin equations driven by independent Brownian motions.

4.1. Mathematical simplifications. To simplify the analysis and simulation of the spine, we make several drastic simplifications. The quality of the simplified model is evaluated by its ability to capture the main phenomenology observed in experiment and by its ability to predict the fluid flow and the time dependence of the measured calcium concentration inside the dendritic spine.

The first simplification is that we consider the ions to be point charges, that is, we neglect Lennard–Jones repulsion. The second simplification is that we neglect electrostatic ion-ion interactions. This means that we can neglect all ionic species except the calcium, whose concentration needs to be predicted. We replace electrostatic interactions by interactions with a fixed mean field (that is, with a field not computed from Poisson's equation). Thus, we assume that the calcium ions move in an effective electrostatic field created by their interactions with each other and with other ions and by the permanent charge distribution on the CaM, calcineurin proteins, and AM complex. The behavior of this potential is assumed, rather than computed. We also neglect the change in the shape of the potential when a calcium ion binds to a protein molecule. The third simplification is that we neglect the impenetrable obstacles to the ionic motion posed by the presence of the proteins. Thus, we assume that the ionic motion inside the dendritic spine is geometrically unrestricted. Therefore, the domain Ω_t is the interior of the dendritic spine.

4.2. The Langevin equations. For a dendritic spine containing N ions of different species (e.g., Ca^{++}, Na^{+}, Cl^{-}, and so on), $\boldsymbol{x}_i(t)$ is the displacement vector of the ith ion, m_i is its mass, and z_i is its valence. $\tilde{\boldsymbol{x}} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)$ is the coordinate of the N ions in configuration space. We assume that a flow field $\boldsymbol{V}(\boldsymbol{x}, t)$ is given (see description below) and that ions interact with a fixed potential of the charges on the proteins, $U_0(\boldsymbol{x})$, and with the variable potential of all other ions. The variable potential consists of both the electrostatic ion-ion interaction potential, $U_{\mathrm{ii}}(\tilde{\boldsymbol{x}})$, and the potential of Lennard–Jones-type repulsions, $U_{\mathrm{LJ}}(\tilde{\boldsymbol{x}})$ (that represents the finite size of the ions). The force per unit mass on the ith ion is

$$\boldsymbol{F}_i(\tilde{\boldsymbol{x}}) = -z_i e \nabla_{\boldsymbol{x}_i} \left[U_0(\boldsymbol{x}_i) + U_{\mathrm{ii}}(\tilde{\boldsymbol{x}}) \right] - \nabla_{\boldsymbol{x}_i} U_{\mathrm{LJ}}(\tilde{\boldsymbol{x}}).$$

The dynamics of the ith ion is given by the Langevin equation

$$\ddot{\boldsymbol{x}}_i + \gamma_i \left[\dot{\boldsymbol{x}}_i - \boldsymbol{V}(\boldsymbol{x}_i, t) \right] + \boldsymbol{F}_i(\tilde{\boldsymbol{x}}) = \sqrt{2\epsilon_i \gamma_i}\ \dot{\boldsymbol{w}}_i, \tag{4.1}$$

where e is the electronic charge. Here $\epsilon_i = k_B T / m_i$, T is the temperature, k_B is the Boltzmann constant, $\gamma_i = 6\pi a_i \eta_i$ is the dynamical viscosity (where η_i is the viscosity coefficient per unit mass), and a_i is the radius of the ion. The frictional drag force, $-\gamma \left[\dot{\boldsymbol{x}}_i - \boldsymbol{V}(\boldsymbol{x}_i, t) \right]$, is proportional to the relative velocity of the ion and the cytoplasmic fluid. The accelerations $\dot{\boldsymbol{w}}_i$ represent the thermal fluctuations of the fluid. The relation between the velocity diffusion constant and the friction coefficient,

$$D_i = \frac{k_B T}{m_i \gamma_i},$$

is Einstein's fluctuation-dissipation principle [30].

In the Smoluchowski limit of large damping [30], the Langevin equation (4.1) reduces to

$$\gamma_i \left[\dot{\boldsymbol{x}}_i - \boldsymbol{V}(\boldsymbol{x}_i, t) \right] + \boldsymbol{F}_i(\tilde{\boldsymbol{x}}) = \sqrt{2\epsilon_i \gamma_i}\ \dot{\boldsymbol{w}}_i. \tag{4.2}$$

In this paper, we neglect the ion-ion interactions; that is, we set $U_{\mathrm{LJ}}(\tilde{\boldsymbol{x}}) = U_{\mathrm{ii}}(\tilde{\boldsymbol{x}}) = 0$ so that (4.2) becomes

$$\gamma_i \left[\dot{\boldsymbol{x}}_i - \boldsymbol{V}(\boldsymbol{x}_i, t)\right] + \boldsymbol{F}(\boldsymbol{x}_i) = \sqrt{2\epsilon_i \gamma_i}\ \dot{\boldsymbol{w}}_i, \tag{4.3}$$

where

$$\boldsymbol{F}(\boldsymbol{x}_i) = -z_i e \nabla_{\boldsymbol{x}_i} U_0(\boldsymbol{x}_i).$$

Since we are interested in tracing only one species in the spine, namely, the concentration of calcium, we assume that $\gamma_i = \gamma_{_{Ca}++},\ m_i = m_{_{Ca}++},\ z_i = z = 2$.

Under these assumptions, equations (4.3) are independent and identical, so that their transition probability densities are identical. We denote the transition probability density function (pdf) of each ion by $p(\boldsymbol{x}, t \,|\, \boldsymbol{x}_0, t_0)$ so that the calcium concentration is

$$c(\boldsymbol{x}, t) = \int_{\Omega_t} p(\boldsymbol{x}, t \,|\, \boldsymbol{x}_0, t_0) c_0(\boldsymbol{x}_0)\, d\boldsymbol{x}_0,$$

where $c_0(\boldsymbol{x}_0)$ is the initial calcium density.

4.3. Reaction-diffusion description of the binding and unbinding reactions. We derive a reaction-diffusion system of equations in a slightly more general setting. We consider a single reactant M (e.g., calcium), whose density, $c_M(\boldsymbol{x}, t)$, satisfies the Nernst–Planck (or Smoluchowski) equation corresponding to the Langevin dynamics (4.2) [30],

$$\frac{\partial c_M(\boldsymbol{x}, t)}{\partial t} = -\nabla \cdot \boldsymbol{J}(\boldsymbol{x}, t), \tag{4.4}$$

where the flux $\boldsymbol{J}(\boldsymbol{x}, t)$ is defined as

$$\boldsymbol{J}(\boldsymbol{x}, t) = \left[\boldsymbol{V}(\boldsymbol{x}, t) - \frac{\boldsymbol{F}(\boldsymbol{x}, t)}{\gamma}\right] c_M(\boldsymbol{x}, t) - D \nabla c_M(\boldsymbol{x}, t). \tag{4.5}$$

The immobile substrate protein S is represented in this model by the potential $U_0(\boldsymbol{x}, t)$ of the electrostatic force $\boldsymbol{F}(\boldsymbol{x}, t)$. This force varies in time as reactant ions bind to or unbind from the substrate, thus changing the net electrostatic charge on the substrate. Instead of following the details of the binding and unbinding process and the fluctuations in the force $\boldsymbol{F}(\boldsymbol{x}, t)$, we coarse grain the Nernst–Planck equation (4.4) by replacing it with reaction-diffusion equations.

To formulate our problem in terms of reaction-diffusion equations, we partition the boundary of the domain Ω into three parts: the pumps and the bottom of the neck, denoted $\partial\Omega_a(t)$, which absorb calcium ions; the remaining surface of the head, denoted $\Sigma_H(t)$; and the surface of the neck, denoted $\partial\Omega_N$, where the normal flux equals the velocity of the boundary at each point. We introduce the variables $S^{(j)}(\boldsymbol{x}, t)$, $(0 \le j \le 4)$, that represent the number of proteins in a test volume about $\boldsymbol{x}$ that contains j bound M ions at time t. Then the number of occupied binding sites on these proteins is $jS^{(j)}(\boldsymbol{x}, t)$, and the number of free binding sites on these proteins is $(4-j)S^{(j)}(\boldsymbol{x}, t)$. Obviously, at all times

$$\sum_{j=0}^{4} S^{(j)}(\boldsymbol{x}, t) = S_0(\boldsymbol{x}),$$

where $S_0(\boldsymbol{x})$ is the number of proteins in the volume element.

We assume that the forward and backward reaction rates, k_1 and k_{-1}, respectively, are constant and independent of the densities (see discussion in section 8 below). It follows that the reaction-diffusion equations for the number of free calcium ions, $M(\boldsymbol{x},t)$, and $S^{(j)}(\boldsymbol{x},t)$ are

$$\frac{\partial M(\boldsymbol{x},t)}{\partial t} = -\nabla\cdot\boldsymbol{J}_M(\boldsymbol{x},t) - k_1 M(\boldsymbol{x},t)\sum_{j=0}^{4}(4-j)S^{(j)}(\boldsymbol{x},t)$$

$$+k_{-1}\sum_{j=0}^{4} jS^{(j)}(\boldsymbol{x},t), \tag{4.6}$$

$$\frac{\partial S^{(j)}(\boldsymbol{x},t)}{\partial t} = k_1 M(\boldsymbol{x},t)\left[(5-j)S^{(j-1)}(\boldsymbol{x},t) - (4-j)S^{(j)}(\boldsymbol{x},t)\right]$$

$$-k_{-1}\left[jS^{(j)}(\boldsymbol{x},t) - (j+1)S^{(j+1)}(\boldsymbol{x},t)\right], \tag{4.7}$$

where the flux is defined by

$$\boldsymbol{J}_M(\boldsymbol{x},t) = -D\nabla M(\boldsymbol{x},t) + \boldsymbol{V}(\boldsymbol{x},t)M(\boldsymbol{x},t), \tag{4.8}$$

and $S^{(-1)}(\boldsymbol{x},t) = S^{(5)}(\boldsymbol{x},t) = 0$. The initial conditions are

$$S^{(0)}(\boldsymbol{x},0) = S_0(\boldsymbol{x}), \quad S^{(j)}(\boldsymbol{x},0) = 0 \quad \text{for } 1\le j\le 4. \tag{4.9}$$

The system (4.6), (4.7) is a coarse-grained reaction-diffusion model of the transient chemical reaction in $\Omega(t)$. Renormalizing the numbers of the different species per unit test volume converts them into densities. Obviously, the forward rate constant k_1 has to be changed accordingly. The initial and boundary conditions for $M(\boldsymbol{x},t)$ are the initial reactant density, absorption at the absorbing boundary, and flux given by the motion of the reflective boundary,

$$M(\boldsymbol{x},0) = c_0(\boldsymbol{x}) \quad \text{for } \boldsymbol{x}\in\Omega(t),$$

$$M(\boldsymbol{x},t) = 0 \quad \text{for } \boldsymbol{x}\in\partial\Omega_a(t), \tag{4.10}$$

$$\boldsymbol{J}_M(\boldsymbol{x},t)\cdot\boldsymbol{\nu}(\boldsymbol{x}) = 0 \quad \text{for } \boldsymbol{x}\in\partial\Omega_N,$$

$$\frac{\partial M(\boldsymbol{x},t)}{\partial n(\boldsymbol{x})} = 0 \quad \text{for } \boldsymbol{x}\in\Sigma_H(t). \tag{4.11}$$

The boundary condition (4.11) means that the boundary flux $\boldsymbol{J}_M(\boldsymbol{x},t)\cdot\boldsymbol{\nu}(\boldsymbol{x})$ (see (4.8)) is actually the flux of the particles carried by the moving boundary. Note that $\boldsymbol{V}(\boldsymbol{x},t)\cdot\boldsymbol{\nu}(\boldsymbol{x}) = 0$ on $\partial\Omega_N$. The geometrical effect of substrate distribution is expressed in $S_0(\boldsymbol{x})$. There are no moving internal boundaries, because the support of $S^{(j)}(\boldsymbol{x},t)$ at all times is that of $S_0(\boldsymbol{x})$.

4.4. Specification of the hydrodynamical flow. The flow of the incompressible cytoplasmic fluid, as explained above, is due to the local contraction of the saturated AM complexes. We assume that the flow field is derived from a potential $\phi(\boldsymbol{x},t)$

(see, e.g., [31]),

$$\boldsymbol{V}(\boldsymbol{x},t) = \nabla\phi(\boldsymbol{x},t). \tag{4.12}$$

The incompressibility condition, $\nabla\cdot\boldsymbol{V}(\boldsymbol{x},t) = 0$, reduces to the Laplace equation in the head $\Omega_H(t)$ of the spine at time t. The surface of the head, $\Sigma(t)$, is partitioned into the surface $\Sigma_H(t)$ of the spine head that does include the surface common with the neck and the cap $\Sigma_N(t)$ of the surface of the head inside the neck, $\Sigma(t) = \Sigma_H(t) \cup \Sigma_N(t)$.

The Laplace equation in $\Omega_H(t)$ is

$$\Delta_{\boldsymbol{y}}\phi(\boldsymbol{y},t) = 0 \quad \text{for } \boldsymbol{y} \in \Omega_H(t),\ t > 0\ , \tag{4.13}$$

with the boundary conditions

$$\left.\frac{\partial\phi(\boldsymbol{y},t)}{\partial n}\right|_{\boldsymbol{y}\in\Sigma_H(t)} = -V(t), \quad \left.\frac{\partial\phi(\boldsymbol{y},t)}{\partial n}\right|_{\boldsymbol{y}\in\Sigma_N(t)} = F(V(t)), \tag{4.14}$$

where $V(t)$ is the average velocity induced by the deformation of the head (see (4.17) below and the appendix), due to the sum of all the local contractions, and $F(V(t))$ is the induced field velocity at the top of the neck $\Sigma_N(t)$. The function $F(V)$ is described in the appendix. The quantities $V(t)$ and $F(V(t))$ are stochastic processes that are proportional to the number of saturated proteins at any given time t.

The flow field can be expressed explicitly in terms of the functions $V(t)$ and $F(V(t))$ by Green's function for the Neumann problem for Poisson's equation in a sphere (or a disk) through Stokes's formula. Green's function $G(\boldsymbol{x},\boldsymbol{y},t)$ is the solution (defined up to a constant) of the equation

$$\begin{aligned} -\Delta_{\boldsymbol{y}}G(\boldsymbol{x},\boldsymbol{y},t) &= \delta(\boldsymbol{x}-\boldsymbol{y}) - \frac{1}{|\Omega_t|} \quad \text{for } \boldsymbol{x},\boldsymbol{y} \in \Omega_H(t), \\ \frac{\partial G(\boldsymbol{x},\boldsymbol{y},t)}{\partial\boldsymbol{\nu}(\boldsymbol{y})} &= 0 \quad \text{for } \boldsymbol{x} \in \Omega_H(t),\ \boldsymbol{y} \in \Sigma(t). \end{aligned} \tag{4.15}$$

Multiplying (4.13) by $G(\boldsymbol{x},\boldsymbol{y},t)$ and (4.15) by $\phi(\boldsymbol{y},t)$ and integrating with respect to $\boldsymbol{y}$ over the domain, using Stokes's theorem and the boundary condition (4.14), we get

$$\begin{aligned} \phi(\boldsymbol{x},t) &= \int_{\boldsymbol{y}\in\Sigma(t)} \frac{\partial\phi(\boldsymbol{y},t)}{\partial n} G(\boldsymbol{x},\boldsymbol{y},t)\, dS_{\boldsymbol{y}} - \int_{\boldsymbol{y}\in\Sigma(t)} \frac{\partial G(\boldsymbol{x},\boldsymbol{y},t)}{\partial n}\phi(\boldsymbol{y},t)\, dS_{\boldsymbol{y}} \\ &\quad + \frac{1}{V_H}\int_{\Omega_H(t)} \phi(\boldsymbol{y},t) d\boldsymbol{y} \\ &= \int_{\boldsymbol{y}\in\Sigma(t)} \frac{\partial\phi(\boldsymbol{y},t)}{\partial n} G(\boldsymbol{x},\boldsymbol{y},t)\, dS_{\boldsymbol{y}} + \frac{1}{V_H}\int_{\Omega_H(t)} \phi(\boldsymbol{y},t) d\boldsymbol{y} \\ &= -\int_{\Sigma_H(t)} V(t) G(\boldsymbol{x},\boldsymbol{y},t)\, dS_{\boldsymbol{y}} + \int_{\Sigma_N(t)} F(V(t)) G(\boldsymbol{x}\, y,t)\, S_{dy} \\ &\quad + \frac{1}{V_H}\int_{\Omega_H(t)} \phi(\boldsymbol{y},t) d\boldsymbol{y} \\ &= -V(t)\int_{\Sigma_H(t)} G(\boldsymbol{x},\boldsymbol{y},t)\, dS_{\boldsymbol{y}} + F(V(t))\int_{\Sigma_N(t)} G(\boldsymbol{x},\boldsymbol{y},t)\, dS_{\boldsymbol{y}} \\ &\quad + \frac{1}{V_H}\int_{\Omega_H(t)} \phi(\boldsymbol{y},t) d\boldsymbol{y}. \end{aligned}$$

The flow field is given by

$$\nabla\phi(\boldsymbol{x},t) = -V(t)\int_{\Sigma_H} \nabla_{\boldsymbol{x}} G(\boldsymbol{x},\boldsymbol{y})\, dS_{\boldsymbol{y}} + F(V(t))\int_{\Sigma_N} \nabla_{\boldsymbol{x}} G(\boldsymbol{x},\boldsymbol{y})\, dS_{\boldsymbol{y}}.$$

In the neck, due to the symmetries and the uniform initial conditions, we simplify the flow field by assuming its velocity is parallel to the axis of the neck. It is given by

$$\nabla\phi(\boldsymbol{x},t) = \boldsymbol{V}(\boldsymbol{x},t) = F(V(t))\boldsymbol{k},$$

where $\boldsymbol{k}$ is a unit vector along the axis of the neck.

In order to close the equations, we recall that the velocity of the boundary $V(t)$ is a function of the number of proteins

$$S^{(4)}(t) = \int_\Omega S^{(4)}(\boldsymbol{x},t)\, d\boldsymbol{x} \tag{4.16}$$

that are saturated at time t. Thus

$$\begin{aligned} V(t) &= v_Q S^{(4)}(t), \\ F(V(t)) &= K v_Q S^{(4)}(t), \end{aligned} \tag{4.17}$$

where v_Q is a constant velocity, depending on the nature of the contractile protein, and K is a constant that depends on the geometry of the spine and the dimension (here 2 or 3). We note that, according to (4.17), as the number of saturated proteins increases the hydrodynamical flow begins to dominate the diffusion.

5. The chemical kinetics of the binding and unbinding reactions. The forward binding reaction of M to S, $M + S \xrightarrow{k_1} MS$, is governed by a forward rate constant k_1, because the process of binding consists of an ion falling into a potential well. The survival probability of a single ion inside the spine head, in the presence of potential traps, decays exponentially fast, so that the rate constant for binding is the exponential decay rate. If the binding process involves many ions, the binding rate is the total absorption flux on the boundaries of the potential wells [32].

More precisely, the instantaneous binding rate is

$$k_1(t) = \oint_{\partial\Omega_{S(t)}} \boldsymbol{J}(\boldsymbol{x},t)\cdot\boldsymbol{\nu}(\boldsymbol{x})\, dS_{\boldsymbol{x}}, \tag{5.1}$$

where $\partial\Omega_{S(t)}$ is the boundary of the free binding sites on the substrate at time t. An approximation to $k_1(t)$ can be obtained by replacing the flux density $\boldsymbol{J}(\boldsymbol{x},t)\cdot\boldsymbol{\nu}(\boldsymbol{x})$ with its instantaneous average over the entire boundary $\partial\Omega_{S(t)}$. Then the local instantaneous binding rate of calcium near $\boldsymbol{x}$ is

$$k_1(t) = k_1\sum_{j=0}^{4}(4-j)S^{(j)}(\boldsymbol{x},t), \tag{5.2}$$

where k_1 is the forward binding rate per ion per protein and $S^{(j)}(\boldsymbol{x},t)$ is the number of proteins with j attached calcium ions. When the radius of a potential well with circular cross-section is R_p, the forward binding rate constant k_1 is given by Smoluchowski's formula

$$k_1 = 2\pi R_p D_M, \tag{5.3}$$

where D_M is the diffusion constant of M-ions [32], [33]. This determination is done for a separate reaction, not necessarily in the domain Ω. The forward rate constant k_1 is an input parameter into the model, e.g., from a molecular dynamics simulation or from direct measurement in a separate chemical reaction [26]. Note that the forward binding rate depends on the radius of the potential well, but not on its depth.

The backward binding rate, k_{-1}, is the rate at which ions escape the potential well. According to Kramers's theory [30], [34] such a dissociation is due to thermal activation of the ions inside the potential well, and its rate is given by the Arrhenius law with a given activation energy. We recall that in Kramers's theory of thermal activation over a smooth (parabolic) potential barrier, the dissociation rate is one-half the reciprocal of the mean first passage time (MFPT) of an ion, initially inside the well, to its boundary [30], [34], [35]. This constant is also an input parameter. Given k_1, k_{-1}, both the depth and the radius of a binding site can be selected by calibration according to (5.3) and Kramers's formulas. Explicit calculations are given in [36].

6. Simulation of calcium kinetics in dendritic spines. When channels open, the maximal number of calcium ions that flow into the dendritic spine is of the order of a few hundred [19], which is also the order of magnitude of the number of calmodulin or myosin molecules inside the spine head. A Brownian simulation gives a description of calcium dynamics over a wide range of parameters, starting with only one ion in the spine and up to a number, where a continuum approximation is valid. In such a simulation the number of bonds formed by each calcium ion can be monitored over time. The number of bound proteins at a given time is a random process, because the forward and backward binding processes occur at random times and at random places. Consequently, the twitching movement of the dendritic spine is a random process as well.

Simulations of our model give the probability that an ion forms a bond, given the protein distribution. They also demonstrate the role of the drift in modifying the recurrent bindings and unbindings of the Brownian particles to given proteins.

6.1. A Langevin (Brownian) dynamics simulation. The binding (unbinding) of Ca^{++} ions to (from) a fixed substrate S (e.g., CaM, AM, etc.) can be described in a Langevin simulation at various degrees of molecular resolution. The simplest way is to describe the binding sites as appropriately calibrated potential wells and count the number of occupied wells as a function of time. A trajectory that hits a free pump on the boundary of the spine head or reaches the dendritic shaft at the bottom of the spine neck is terminated there. The remaining part of the boundary is reflecting to trajectories. This is essentially the simplified molecular dynamics simulation described above. A coarse-grained description of the reaction of binding and unbinding of the diffusing ions with the immobile substrate is given by the reaction-diffusion equations (4.6)–(4.10). A numerical study of this system will be presented in a separate paper.

The results of a full Langevin simulation based on our simplified physical and mathematical model are summarized below (see [26] for details). These results can be used as benchmarks for the results of the coarse-grained model described above.

Figure 2 shows the results of a simulation with $N_{init} = 100$, $k_{-1} = K_{\text{back}}^{\text{AM}} = 1\,\text{kHz}$, $k_1 = K_{\text{back}}^{\text{cal}} = 5\,\text{Hz/M}$, $R = 0.5\ \mu\text{m}$, $l = 0.2\ \mu\text{m}$, $d/2 = \mu\text{m}$, $N_{\text{pumps}} = 10$, and the protein molecules (about 50 of AM type and 10 of the other type) distributed in the PSD.

Two types of decay can be discerned in the first graph of calcium concentration vs. time in Figure 2: quick decay, starting at the beginning of the simulation and ending at about 250 msec, is followed by a slower decay that continues to the end of

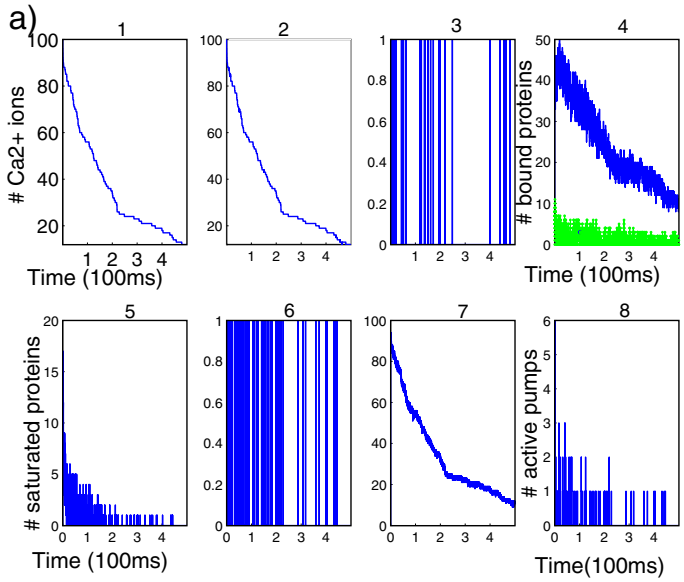


FIG. 2. *Dynamics of* 100 *calcium ions in dendritic spine. Time evolution of the concentration and binding. First row, concentration vs. time (in μsec), from left to right:* 1. $[Ca^{2+}]$ *in the total spine.* 2. $[Ca^{2+}]$ *in spine head.* 3. *Number of ions in the neck. Note that the neck contains only one ion at a time.* 4. *Number of bound proteins (type* 1*, blue; type* 2*, green). Note the stochastic nature of those curves. Second row, from left to right:* 5. *Number of saturated proteins of type* 1 *vs. time.* 6. *Arrival times of ions at active pumps: the ions leave one at a time.* 7. *Number of bound ions vs. time.* 8. *Number of active pumps vs. time. (Figure reprinted from* [D. Holcman, Z. Schuss, and E. Korkotion*, Calcium dynamics in dendritic spines and spine motility,* Biophysical Journal, 87 (2004), pp. 81–91] *with permission.*

the simulation. The first period is the decay curve of the saturation of type 1 proteins. When a simulation starts with 100 ions, only 10 proteins get saturated; that is, only about 40 ions are captured at the beginning and the number of saturated proteins continues to decay in time. To have a rough idea of the effect of the hydrodynamical push, we can approximate the push by its average of 2.5 proteins saturated for the first 250 msec, where each protein contributes to the speed a total of 50 nm/msec. The total speed of the push is $0.5\,\mu$m/ms. The push speeds up the arrival of ions at the lower part of the spine head, where the pumps are located, relative to arrivals by pure diffusion. Since the sojourn time of ions in the pumps is chosen to be short, the ions leave mainly through the head. At 500 msec into the simulation only about 20% of initial ions are still in the spine. In this simulation the effect of the push is not sufficiently strong to direct all the ions toward the neck. The 1:4 ratio of the efflux through the pumps, compared to that through the dendrite, may be due to the large number of fast pumps. These results are in agreement with the experimental results of [19].

In Figure 3 the results of simulations with and without the hydrodynamical push are shown: blue curves correspond to a simulation without the push effect, while magenta curves correspond to simulations with it. The characteristic parameters of the simulation are $N_{init} = 200$, $k_{-1} = K^{\rm AM}_{\rm back} = 10\,$kHz, $k_1 = K^{\rm cal}_{\rm back} = 0.5\,$kHz/M, $R = 1\,\mu$m, $l = 0.3\,\mu$m, and $d/2 = 0.3\,\mu$m. There are 4 pumps, 60 AM proteins, and 10 calcineurin proteins [15]. Results similar to those predicted in Figure 3 were

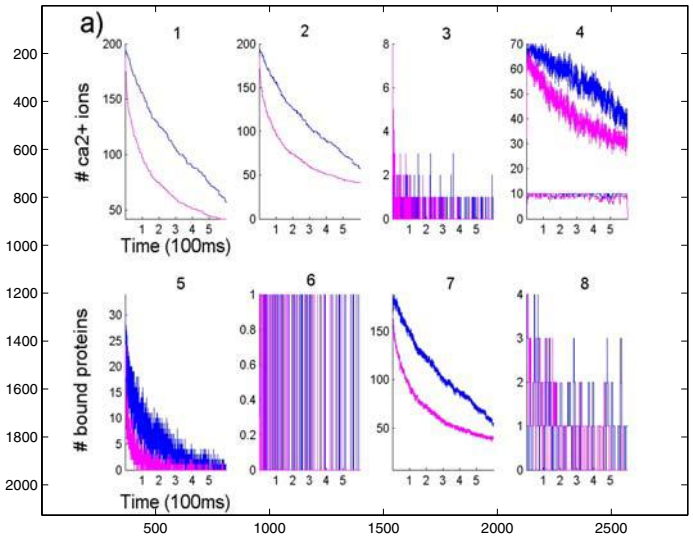


FIG. 3. *Comparison of the time evolution for a postsynaptic distribution with and without push. Blue curves correspond to a simulation without the push effect, while magenta curves correspond to simulations with it. First row, concentration vs. time (in μsec). From left to right:* 1. $[Ca^{2+}]$ *in the total spine.* 2. $[Ca^{2+}]$ *in spine head.* 3. *Number of ions in the neck. Note that the neck contains few ions at a time.* 4. *Number of bound proteins (type* 1*, blue; type* 2*, magenta). Note the stochastic nature of those curves. Second row, from left to right:* 5. *Number of saturated proteins of type* 1 *vs. time.* 6. *Arrival times of ions at active pumps: the ions leave one at a time.* 7. *Number of bound ions vs. time.* 8. *Number of active pumps vs. time. (Figure reprinted from* [D. Holcman, Z. Schuss, and E. Korkotion*, Calcium dynamics in dendritic spines and spine motility,* Biophysical Journal, 87 (2004), pp. 81–91] *with permission.*

obtained recently with calcium replaced by an inert dye that does not bind to proteins [27].

7. An estimate of a decay rate. In the absence of the flow field $\boldsymbol{V}(\boldsymbol{x},t)$, the decay of $M(\boldsymbol{x},t)$ is governed by the first eigenvalue of the Laplace operator in the head with the mixed reflecting and absorbing boundary conditions. In the presence of $\boldsymbol{V}(\boldsymbol{x},t)$, the decay rate can be estimated as follows (see also [37] for another estimate of the fast rate constant, using internal buffer kinetics).

Consider the dynamics

$$\dot{x} = v(t) + \sqrt{2D}\,\dot{w}_x,$$

$$\dot{y} = \sqrt{2D}\,\dot{w}_y, \quad \dot{z} = \sqrt{2D}\,\dot{w}_z$$

in the neck, where w_x, w_y, and w_z are independent Brownian motions. The solution is

$$x(t) = x_0 + \int_0^t v(s)\,ds + \sqrt{2D}\,w_x(t),$$

$$y = y_0 + \sqrt{2D}\,w_y(t), \quad z = z_0 + \sqrt{2D}\,w_z(t).$$

This means that the solution to the Nernst–Planck equation (4.4) in the neck,

$$c_t = D\Delta c - v(t)c_x,$$

is given by

$$c(x,y,z,t) = \int\int_{\text{neck}} \int G(y,z,t \,|\, y_0,z_0) \times \exp\left\{ -\frac{\left(x - x_0 - \int_0^t v(s)\,ds\right)^2}{4Dt} \right\} f(x_0,y_0,z_0)\,dx_0\,dy_0\,dz_0,$$

where $f(x_0,y_0,z_0)$ is the initial ionic density in an infinite neck and $G(y,z,t \,|\, y_0,z_0)$ is Green's function for the diffusion equation in the cross-section of the neck, with reflecting boundary conditions.

We consider the initial decay law, when the decay is due primarily to the hydrodynamical effect, because it is faster than that due to diffusion. Suppose that the ions are concentrated near a single point (x_0',y_0',z_0'); that is, $f(x_0,y_0,z_0) = \delta(x_0 - x_0', y_0 - y_0', z_0 - z_0')$. In the initial tenths of a second, the decay of the concentration in the neck is due to the large hydrodynamical effect. The velocity $v(t)$ is maximal when all proteins are saturated. We write

$$c(x,y,z,t) = G(y,z,t \,|\, y_0',z_0') \exp\left\{ -\frac{\left(x - x_0' - \int_0^t v(s)\,ds\right)^2}{2Dt} \right\}$$

and

$$\frac{\left(x - x_0' - \int_0^t v(s)\,ds\right)^2}{2Dt} = \frac{(x-x_0)^2}{2Dt} - \frac{(x-x_0)}{D}\bar{v}(t) + \frac{\bar{v}^2(t)}{2D}t, \tag{7.1}$$

where

$$\bar{v}(t) = \frac{1}{t}\int_0^t v(s)\,ds \approx \text{const} \equiv \bar{v}_0.$$

Set

$$u_d(x,y,z,t) = G(y,z,t \,|\, y_0',z_0') \exp\left\{ -\frac{(x-x_0)^2}{4Dt} \right\}$$

to represent the diffusion term. We can write in (7.1)

$$-\frac{(x-x_0)\bar{v}(t)}{2D} \approx -\frac{(x-x_0)\bar{v}_0}{2D},$$

so this term does not contribute to the time decay of the concentration $c(x,y,z,t)$ in the initial period. The last term in the exponent (7.1) is approximately $\frac{\bar{v}_0^2}{4D}t$, so in the limit of fast binding, which lasts a few hundreds of milliseconds, we can write

$$c(x,y,z,t) = C u_d(x,y,z,t) \exp\left\{ -\frac{\bar{v}_0^2}{4D}t \right\}.$$

This gives the decay time

$$\tau = \frac{4D}{\bar{v}_0^2}. \tag{7.2}$$

The initial average velocity $\bar{v}_0$ can be estimated, if we assume that all proteins are distributed along the surface of the head and are saturated at the same time. In this case the membrane shrinks on the time scale of a single protein contraction time and of length equal to the number of proteins, $N_p = \int_\Omega S_0(\boldsymbol{x})\, d\boldsymbol{x}$, times the contraction length of a single protein. We consider two models of saturation. First, if the proteins are located on the membrane, we can say that the lengths do not sum, but act in parallel to contract the head. This yields a contraction length of order l_c, the length of contraction of one protein, independently of the number of proteins distributed on the surface. Knowing the size of the myosin protein and the size of the head (e.g., radius of $1\,\mu$m), the maximal number of proteins packed on the membrane surface is $1\,\mathrm{mM}(= 600$ proteins).

Second, if there are different layers of contractile proteins, then all contractions add together, if they occur simultaneously. In that case the length of the contraction equals $N_p l_c$.

We have (see the appendix)

$$\bar{v}_0 = \bar{F}(V(0)) = \begin{cases} \dfrac{4\pi R_0^2 \bar{V}_0}{|\Sigma_N|} & \text{in dimension 3,} \\[2ex] \dfrac{2\pi R_0 \bar{V}_0}{|\Sigma_N|} & \text{in dimension 2,} \end{cases}$$

where $\bar{V}_0 = v_Q N_p$ is the average initial velocity of the surface of the spine head, as given by (4.17) with $S^{(4)}(0) = N_p$ (by assumption); R_0 is the initial radius of the spine head; and $|\Sigma_N|$ is the surface area of the cross-section of the neck.

The value $v_Q = 0.02\,\mu m/\mathrm{ms}$ is given in [2], so that $\bar{v}_0 = 0.1\,\mu m/\mathrm{ms}$; hence

$$\bar{v}_0 = \bar{F}(V(0)) = \begin{cases} \dfrac{0.4\pi R_0^2}{|\Sigma_N|} & \text{in dimension 3,} \\[2ex] \dfrac{0.2\pi R_0}{|\Sigma_N|} & \text{in dimension 2.} \end{cases}$$

Now, (7.2) gives $\tau = 160\,\mathrm{msec}$, which is comparable to the experimental result given in [19].

This result can be obtained also from the following calculations. First, using the assumption that each protein contributes additively to the total contraction in the simulation of [26], we see that during the hydrodynamical push period about 5 contractile proteins are saturated on the average, so the average velocity of the head is $\bar{V}_0 = 0.02 \times 5 = 0.1\,\mu/\mathrm{ms}$. Second, if in our model all proteins are distributed in a single layer and are instantaneously saturated, they produce a contraction of $0.02\,\mu\mathrm{m/sec}$. In this case, when the ratio of surface areas of the head and the neck $(|\Sigma_H| + |\Sigma_N|)/|\Sigma_N| = 5$, the average velocity of the head is $\bar{v}_0$.

8. Discussion and conclusions. We have introduced a Brownian dynamics simulation of calcium dynamics in a dendritic spine and its coarse-grained continuum

description by partial differential equations. The equations couple the hydrodynamical flow, caused by spine motility, to the chemical reaction between the diffusing calcium and the immobile substrate. We have identified the dominant molecular mechanism of the fast macroscopic twitching as the contraction of the calcium saturated AM proteins. This contraction produces a hydrodynamical flow, which causes the fast decay of calcium. The decay rate has been derived theoretically, and the reaction-diffusion equations described here are the coarse-grained version of the Langevin simulation of [26]. They provide a mathematical description of the molecular events during the fast motility of the spine.

The main goal of the coarse-graining is to capture the features of calcium dynamics with a small number of equations, ideally with a single ordinary differential equation, so that a comprehensive model of calcium dynamics in a spiny dendrite can be derived and the effect of hundreds of thousands of spines can be integrated and coupled to an action potential. In this context the dynamics of calcium can possibly be linked to the induction of synaptic plasticity.

An important feature of the Langevin simulation is that the number of bonds per protein, as a function of time, can be followed and compared to the initial calcium concentration. The coarse-grained reaction-diffusion description of the Langevin simulation should be helpful in relating the threshold of initiation of synaptic plasticity, such as LTP, to the initial calcium concentration.

One of the most significant results of this paper is the derivation of the decay rate from the fast motility of the spine. This result should be compared to the calcium extrusion rate in spines, as presented in [19]. The two very distinct decay rates suggest that the fast extrusion period can also be due to the spine fast motility. This phenomenon was ascribed in [19], [37] to the fast pumping of calcium ions into stores. Indeed, in the present paper we have neglected calcium stores and high concentration of buffers, which may impose a decay rate faster than diffusion. An improved model that includes a large number of buffers should reveal the precise contribution of buffers to the calcium fast decay rate, as compared to the rate imposed by the spine contraction. Such a model has to be based on the mechanism of interactions between diffusing calcium and the buffers.

We conclude, on the basis of the present model and of the Langevin simulation of [26], that one of the possible roles in calcium dynamics of the spine's fast motility is to increase significantly the fraction of ions that are directed toward the dendrite and the organelles, compared to the ions that are pumped out.

Appendix. The velocities $V(t)$ and $F(V(t))$. In this section, we compute the velocity $V(t)$ of the spine head surface, given by (4.17), and the velocity $F(V(t))$, used in the boundary condition (4.14). While $V(t)$ is given in (4.17), the velocity in the neck, $F(V(t))$, has to be calculated.

To calculate the cytoplasmic fluid velocity $F(V(t))$ at the surface of the sphere inside the neck, we make the following simplifying assumptions that lead to an explicit expression for the velocity of the efflux. The entire fluid displaced by the contraction of the spine head flows into the spine neck with a uniform velocity $v(t)$ in a direction normal to the sphere. We also assume that the neck is sufficiently narrow so that all normals to the spherical surface inside the neck, Σ_N, are parallel to the axis of the neck.

Under these assumptions the volume displaced per unit time is $4\pi R^2(t)V(t)$ in dimension 3 and $2\pi R(t)V(t)$, where $R(t)$ is the instantaneous radius of the head and

$\dot{R}(t) = -V(t)$. The flux through Σ_N is $|\Sigma_N|v(t)$; hence

$$v(t) = F(V(t)) = \begin{cases} \dfrac{4\pi R^2(t)V(t)}{|\Sigma_N|} & \text{in dimension 3,} \\ \dfrac{2\pi R(t)V(t)}{|\Sigma_N|} & \text{in dimension 2.} \end{cases}$$

Acknowledgments. We would like to thank E. Korkotian and N. Rouach for useful discussions of the biological part of the paper.

REFERENCES

[1] S. Ramón y Cajal, *Histologie du système nerveux de l'homme et des vertébrés*, L. Azouly, transl. Malaine. Paris, France 1909. "New ideas on the structure of the nervous system of man and vertebrates," translated by N. and N. L. Swanson from *Les nouvelles idées sur la structure du système nerveux chez l'homme et chez les vertébrés*, MIT Press, Cambridge, MA, 1991.

[2] E. R. Kandel, J. H. Schwartz, and T. M. Jessel, *Principles of Neural Science*, 4th ed., McGraw-Hill, New York, 2001.

[3] R. Yuste and W. Denk, *Dendritic spines as basic functional units of neuronal integration*, Nature, 375 (1995), pp. 682–684.

[4] T. Bonhoeffer and R. Yuste, *Spine motility: Phenomenology, mechanisms, and function*, Neuron, 35 (2002), pp. 1019–1027.

[5] C. Koch and I. Segev, eds., *Methods in Neuronal Modeling*, MIT Press, Cambridge, MA, 1998.

[6] C. Koch, *Biophysics of Computation*, Oxford University Press, New York, 1999.

[7] C. Koch and A. Zador, *The function of dendritic spines: Devices subserving biochemical rather than electrical compartmentalization*, J. Neurosci., 13 (1993), pp. 413–422.

[8] A. Zador, C. Koch, and T. H. Brown, *Biophysical model of a hebbian synapse*, Proc. Natl. Acad. Sci. USA, 87 (1990), pp. 6718–6722.

[9] G. M. Shepherd, *The dendritic spine: A multi-functional integrative unit*, J. Neurophysiol., 75 (1996), pp. 2197–2210.

[10] I. Segev and W. Rall, *Computational study of an excitable dendritic spine*, J. Neurophysiol., 60 (1988), pp. 499–523.

[11] R. S. Zucker and W. G. Regehr, *Short-term synaptic plasticity*, Ann. Rev. Physiol., 64 (2002), pp. 355–405.

[12] E. Korkotian and M. Segal, *Spike-associated fast twitches of dendritic spines in cultured hippocampal neurons*, Neuron, 30 (2001), pp. 751–758.

[13] F. Blomberg, R. S. Cohen, and P. Siekevitz, *The structure of postsynaptic densities isolated from dog cerebral cortex*. II. *Characterization and arrangement of some of the major proteins within the structure*, J. Cell Biol., 74 (1977), pp. 204–225.

[14] F. Crick, *Do dendritic spines twitch?*, Trends Neurosci., 5 (1982), pp. 44–46.

[15] N. Volfovsky, H. Parnas, M. Segal, and E. Korkotian, *Geometry of dendritic spines affects calcium dynamics in hippocampal neurons: Theory and experiments*, J. Neurophysiol., 82 (1999), pp. 450–454.

[16] J. Lisman, *The CaM kinase* II *hypothesis for the storage of synaptic memory*, Trends Neurosci., 10 (1994), pp. 406–412.

[17] J. Lisman, H. Schulman, and H. Cline, *The molecular basis of CaMK*II *function in synaptic and behavioural memory*, Nat. Rev. Neurosci., 3 (2002), pp. 175–190.

[18] A. Majewska, A. Tashiro, and R. Yuste, *Regulation of spine calcium dynamics by rapid spine motility*, J. Neurosci., 20 (2000), pp. 8262–8268.

[19] A. Majewska, E. Brown, J. Ross, and R. Yuste, *Mechanisms of calcium decay kinetics in hippocampal spines: Role of spine calcium pumps and calcium diffusion through the spine neck in biochemical compartmentalization*, J. Neurosci., 20 (2000), pp. 1722–1734.

[20] A. Dunaevsky, A. Tashiro, A. Majewska, C. Mason, and R. Yuste, *Developmental regulation of spine motility in the mammalian central nervous system*, Proc. Natl. Acad. Sci. USA, 96 (1999), pp. 13438–13443.

[21] E. A. Nimchinsky, B. L. Sabatini, and K. Svoboda, *Structure and function of dendritic spines*, Annu. Rev Physiol., 64 (2002), pp. 313–353.

[22] K. M. Franks and T. J. Sejnowski, *Complexity of calcium signaling in synaptic spines*, Bioessays, 24 (2002), pp. 1130–1144.
[23] M. Fischer, S. Kaech, U. Wagner, H. Brinkhaus, and A. Matus, *Glutamate receptors regulate actin-based plasticity in dendritic spines*, Nat. Neurosci., 3 (2000), pp. 887–894.
[24] M. Fischer, S. Kaech, D. Knutti, and A. Matus, *Rapid actin-based plasticity in dendritic spines*, Neuron, 20 (1998), pp. 847–854.
[25] M. Morales and E. Fifkova, *Distribution of MAP2 in dendritic spines and its colocalization with actin. An immunogold electron-microscope study*, Cell Tissue Res., 256 (1989), pp. 447–456.
[26] D. Holcman, Z. Schuss, and E. Korkotian, *Calcium dynamics in dendritic spines and spine motility*, Biophysical Journal, 87 (2004), pp. 81–91.
[27] M. Segal, *private communication.*
[28] R. C. Malenka and R. A. Nicoll, *Long-term potentiation—A decade of progress?*, Science, 285 (1999), pp. 1870–1874.
[29] B. J. Berne and R. Pecora, *Dynamic Light Scattering*, John Wiley & Sons, New York, 1978.
[30] Z. Schuss, *Theory and Applications of Stochastic Differential Equations*, Wiley Series in Probability and Statistics, John Wiley & Sons, New York, 1980.
[31] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics*, Pergamon Press, Elmsford, NY, 1975.
[32] B. Nadler, T. Naeh, and Z. Schuss, *The stationary arrival process of independent diffusers from a continuum to an absorbing boundary is Poissonian*, SIAM J. Appl. Math., 62 (2001), pp. 433–447.
[33] S. Chandrasekhar, *Noise and stochastic process in physics and astronomy*, Rev. Mod. Phys., 15 (1943), p. 1.
[34] P. Hänggi, P. Talkner, and M. Borkovec, *Reaction rate theory: Fifty years after Kramers*, Rev. Mod. Phys., 62 (2) (1990), pp. 251–341.
[35] B. J. Matkowsky and Z. Schuss, *The exit problem for randomly perturbed dynamical systems*, SIAM J. Appl. Math., 33 (1977), pp. 365–382.
[36] D. Holcman and Z. Schuss, *Kinetics of non-Arrhenius reactions*, preprint.
[37] B. L. Sabatini, M. Maravall, and K. Svoboda, *Ca^{2+} signalling in dendritic spines*, Curr. Opin. Neurobiol., 11 (2001), pp. 349–356.

SIAM J. APPL. MATH.
Vol. 65, No. 3, pp. 1027–1052

EXPLOITING SYMMETRY IN FAN BEAM CT: OVERCOMING THIRD GENERATION UNDERSAMPLING*

STEVEN H. IZEN†, DAVID P. ROHLER‡, AND SASTRY K.L.A.‡

Abstract. A new reconstruction algorithm is presented for tomographic reconstruction from fan beam data acquired with a quarter detector shift. This algorithm exploits the reflection property of the divergent beam transform by representing the sample and reflected points as a discrete, multichannel sample set. A doubling of the reconstruction resolution is achieved by increasing the number of source locations without changing the detector sample density. The algorithm can be used to reconstruct images from a third generation CT scanner at the maximum resolution consistent with the frequency characteristics of the detector.

Key words. tomography, divergent beam transform, Radon transform, multichannel sampling theorem

AMS subject classifications. 92C55, 44A12, 65R10, 94A20

DOI. 10.1137/S0036139902417001

1. Introduction.

A CT scanner provides two-dimensional images of a cross section of an object by measuring integrals of the object's x-ray attenuation through the cross section. The attenuation data can be modeled as giving the Radon transform of x-ray attenuation. The Radon transform of a function f in $\mathbb{R}^2$ is

$$Rf(\phi, p) = \int_{x\Phi=p} f(x)dx, \tag{1.1}$$

where $\Phi = (\cos\phi, \sin\phi)^T \in S^1$ is normal to the line of integration and $p\Phi$ locates the closest point to the origin on the line. The parameterization (ϕ, p) of lines in $\mathbb{R}^2$ is called the parallel beam geometry, as fixing ϕ results in a family of parallel lines.

The divergent beam (or fan beam) parameterization for lines in $\mathbb{R}^2$ is appropriate to use when modeling scanners in which the x-ray source moves on a circle (the source circle) of radius r exterior to a circle (scan circle) of radius ρ containing the object being imaged. The source angle β measures the position of the x-ray source with respect to x axis. The fan angle α measures the detector position with respect to the central ray of the fan, the ray from the source through the origin. By convention, α is positive if, viewed from the source, the detector position is to the left of the central ray. See Figure 1. The parallel beam and fan beam parameterizations are related by

$$p = r\sin\alpha, \qquad \phi = \beta + \alpha - \pi/2, \tag{1.2}$$

where r is the radius of the source circle. The reparameterization gives rise to the divergent beam transform

$$Df(\beta, \alpha) = Rf(\beta + \alpha - \pi/2, r\sin\alpha). \tag{1.3}$$

*Received by the editors November 1, 2002; accepted for publication (in revised form) July 13, 2004; published electronically March 31, 2005.
http://www.siam.org/journals/siap/65-3/41700.html
†Department of Mathematics, Case Western Reserve University, Cleveland, OH, 44106-7058 (shi@cwru.edu).
‡Plexar Associates, 3722 Meadowbrook Blvd., University Heights, OH, 44118 (dpr@plexar.com, sastryk@plexar.com).

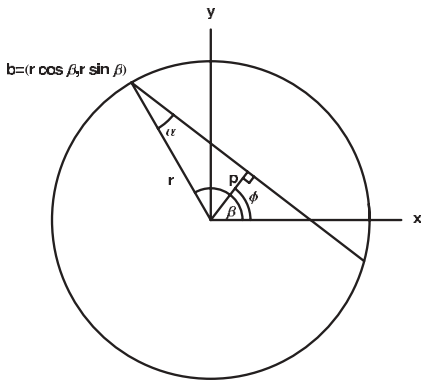


FIG. 1. *The parallel beam and divergent beam parameterizations of lines in* $\mathbb{R}^2$. *The angles* ϕ *and* β *are measured counterclockwise from the* x *axis, and* α *is measured counterclockwise from the line connecting* $(r\cos\beta, r\sin\beta)$ *to the origin.*

In a real measurement system, only a finite set of data is available. The goal in fan beam tomography is to reconstruct f within $|x| \leq \rho$ with as much resolution as possible from discrete samples of Df.

In a third generation CT scanner, there is a detector bank which is physically connected to an x-ray source located on the opposite side of the object being imaged. The entire source-detector assembly is rotated around the object. At a discrete set of positions, the detector apparatus reads out, and an entire fan of data is acquired. In order to attain maximum use of the available photons, it is normal to have each detector element abut its neighbors. While this geometry is efficient in collecting all the available photons, it effectively bandlimits the measured signal. Were the standard fan-beam sampling scheme and reconstruction algorithm (Algorithm 5.3 in [13]) to be used to attain the resolution corresponding to the detector bandwidth, the detector sample density needed would be twice that which is available. In other words, the measurement would be undersampled by a factor of two. This so-called third generation undersampling is explained in section 3.1.

In this paper, we present an algorithm which can be used with third generation CT scanners to attain reconstructions at a resolution limited only by the bandwidth of the detector elements.

The divergent beam transform has a twofold symmetry which can be exploited to compensate for the third generation undersampling. Placing a detector bank off center by one quarter of the width of a detector element breaks the acquisition symmetry, effectively doubling the sample density. A similar detector shift was suggested for the third generation problem in the parallel beam geometry by Cormack [2].

For the parallel beam geometry, the sample points obtained by introducing a detector shift can be reorganized as a lattice with twice the density along the detector direction. When a bandlimited function is sampled on a suitably dense lattice, its integral can be replaced by a discrete sum without error. Applying this to continuous inversion formulas is the foundation for both the standard and interlaced parallel beam reconstruction algorithms (Algorithms 5.1 and 5.2 in [13]).

For the divergent beam geometry, the fortuitous realignment of the original and reflected sample points into a single lattice does not occur, so the standard multidimensional sampling theory cannot be applied. Instead, multichannel sampling theory

[1, 16] in a lattice context [3, 6, 7, 8, 9] must be used. In this paper, it is shown that the original and reflected sample points can be reorganized as a union of identical rectangular lattices. The samples on each rectangular lattice become a channel of a multichannel lattice sample scheme. For appropriate scanner geometries, the Parseval-like result of [9] for sampling on unions of lattices can be used to justify the replacement of the continuous inversion integral by the discrete analog. This leads to a new fan beam algorithm which can reconstruct at the maximum resolution consistent with the detector-limited bandwidth. An alternate approach [6, 7] can also be applied. However, that method is more complex than our method, as it requires the multichannel lattice data to be interpolated to a sufficiently dense single lattice.

The remainder of this paper is organized as follows. In section 2.1, a brief review of lattice sampling theory and multichannel sampling for lattices is presented. In section 2.2, the sampling requirements for the divergent beam transform are reviewed. The third generation detector model is described in section 3.1. It is shown how to apply multichannel sampling theory in this context in section 3.2. The new divergent beam reconstruction algorithm appears in section 3.2.2. Finally, a numerical implementation which validates the algorithm is exhibited in section 4.

2. Sampling in tomography. In order to achieve a desired resolution in the reconstructed image, the divergent beam transform must be sampled at an appropriate density. A framework for studying reconstruction resolution in terms of spatial frequency content was presented by Natterer in [10]. A function f is Ω-bandlimited if its Fourier transform,

$$\hat{f}(\xi) = (2\pi)^{-n/2} \int_{\mathbb{R}^n} f(x) e^{-ix\xi} dx, \tag{2.1}$$

satisfies $\hat{f}(\xi) = 0$ for $|\xi| > \Omega$. Such functions can represent details no smaller than π/Ω. In practice, as the reconstruction region will have compact support, the functions to be recovered cannot be bandlimited. As a result, one works with essentially bandlimited functions. A function f on $\mathbb{R}^n$ is essentially Ω-bandlimited if $\hat{f}(\xi)$ is small for $\xi > \Omega$. More precisely, it is assumed that $\varepsilon_0(f, \Omega) = \int_{|\xi|>\Omega} |\hat{f}(\xi)| d\xi$ is negligible. Refer to [10] for a thorough treatment of this notion in the context of sampling theory.

The sample density needed to reconstruct a bandlimited function f can be determined by examining the bandregion $K = \operatorname{supp} \hat{f}$. Roughly speaking, a sampled function has as its spectrum sums of translates of the spectrum of the unsampled function. The translation distance is inversely proportional to the sample density. To reconstruct accurately, it is necessary to ensure that translates of K are mutually disjoint.

The Radon transform and divergent beam transforms possess symmetries in both the Fourier domain and the sample domain. The essential support of the divergent beam transform of a bandlimited function has the shape [11, 15] shown in Figure 2. Selecting nonoverlapping translates parallel to the coordinate axes lead to the standard fan beam reconstruction algorithm (Algorithm 5.3 in [13]). Much of Fourier space remains uncovered. A completely efficient tiling of Fourier space is possible [11]. The corresponding sampling scheme requires a complex dynamic detector shift, the periodicity of which depends on the ratio of the source and scan radii. This acquisition is difficult to implement in hardware.

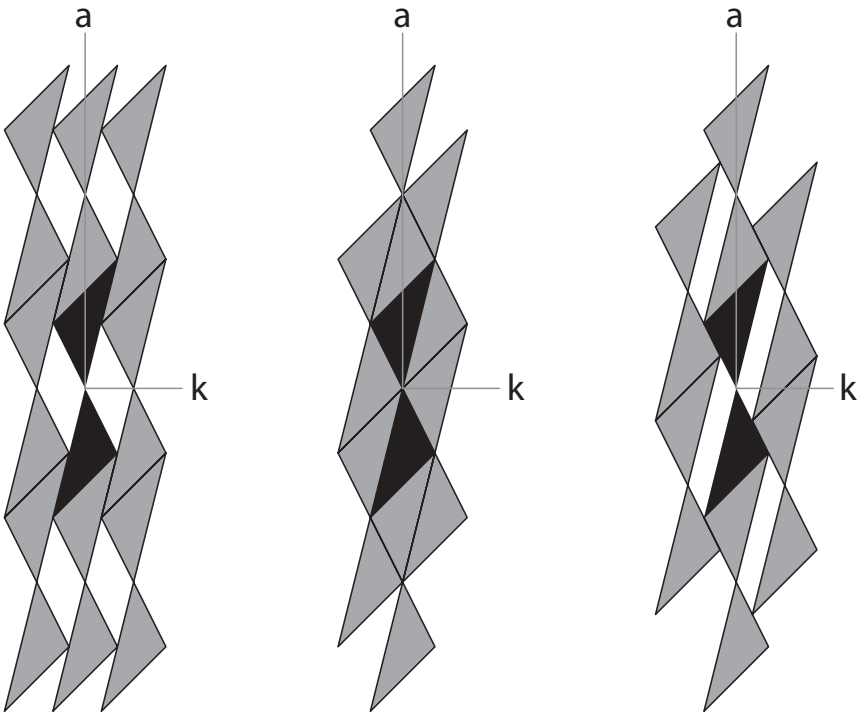


FIG. 2. *The solid region shows the essential support K when $\rho/r = 1/3$ of the Fourier transform of the divergent beam transform of an essentially Ω-bandlimited function. K scales linearly with Ω along each axis and is symmetric with respect to the origin. The coordinates of the upper and lower right corners of K are, respectively, $(r\Omega, (\rho + r)\Omega)$ and $(r\Omega, (\rho - r)\Omega)$. In the leftmost figure, the gray areas are translates of K parallel to the coordinate axes by the minimal distance required to avoid overlaps. The corresponding sampling scheme is the standard fan beam sampling. In the middle figure, the gray areas are translates of K which are optimally packed. The corresponding sampling scheme is the efficient fan beam sampling. To acquire data for the efficient sampling scheme requires a dynamic detector shift. In the rightmost figure, the gray areas are the translates of K which correspond to sampling the divergent beam transform at points which are reflections of the points in the standard sampling scheme. Since there is no overlap, this sample set can be used to reconstruct at the same resolution as the standard fan beam sampling.*

Alternatively, symmetry can be seen directly in the sample domain via the reflection property, which for the Radon transform reads

$$Rf(\phi, p) = Rf(\phi + \pi, -p), \tag{2.2}$$

and for the divergent beam transform reads

$$Df(\beta, \alpha) = Df(\beta + \pi + 2\alpha, -\alpha). \tag{2.3}$$

That is, any sample can be used twice, once at the actual sample point and again at the reflected point. By carefully choosing the sample points to avoid duplication as the source is rotated through a full 2π on the source circle, the sample density is effectively doubled, which, in principle, can compensate for undersampling. A shift of a quarter detector width in the placement of the origin of the detector bank avoids duplication and interlaces the reflected sample points as evenly as possible [12].

In order to directly apply the standard reconstruction algorithms, the original and reflected sample points must line up on a suitable rectangular grid. This scenario

occurs for the parallel beam geometry. In fact, it can be shown that the result of the standard parallel beam reconstruction algorithm (Algorithm 5.1 in [13]) applied to a 2π scan ($0 \le \phi \le 2\pi$) with a 1/4 detector shift is numerically identical to the result when the same algorithm is applied to a π scan ($0 \le \phi \le \pi$) with a 1/2 detector shift with samples spaced at a (nonphysical) half detector width.

For the divergent beam geometry, the reflected sample points do not line up in a rectangular grid. Although the sample positions will be regular in the coordinate tracking detector position, the reflection geometry induces a shift in the coordinate tracking the source position. This shift is periodic, with the periodicity depending on the choice of sampling density. In this paper, a new divergent beam reconstruction algorithm is constructed which properly accounts for these shifts. The key idea is that the sampling pattern resulting from the union of the original and reflected sample points can be viewed as a multichannel sample scheme for the divergent beam transform. This construction answers the question posed in [12] about how to effectively exploit the symmetry in the divergent beam transform.

2.1. Sampling on lattices. To analyze sampling in tomography [4, 6, 10, 11, 13], it is necessary to invoke the Petersen–Middleton sampling theory on lattices [17]. For the analysis of the divergent beam geometry, it is necessary to use an extension of lattice sampling theory to a multichannel setting. This can be done with the theory developed in [6, 7, 8], or, as in this paper, the theory developed in [9].

In this subsection, notation will be introduced, and the relevant results will be quoted.

Given n linearly independent vectors $w_i \in \mathbb{R}^n$, denote by W the invertible matrix with columns w_i. The lattice L_W is defined by

$$L_W = \{Wz \mid z \in \mathbb{Z}^n\} = W\mathbb{Z}^n. \tag{2.4}$$

Although the lattice L_W does not uniquely determine the generating matrix W, $|\det W|$ is independent of the choice of the generating matrix for the lattice.

The dual lattice $L_W^\perp$ is defined to be $L_{W^\perp}$, where $W^\perp = 2\pi W^{-T}$. Any nonsingular $n \times n$ matrix M with integer entries defines a sublattice $L_P \subseteq L_W$, where $P = WM$. If $L_P \subseteq L_W$, then $L_P^\perp \supseteq L_W^\perp$, and $|\frac{\det P}{\det W}| = |\frac{\det W^\perp}{\det P^\perp}| = |\det M|$.

A fundamental region of a lattice L_P is a subset $F_P \subseteq \mathbb{R}^n$ such that each $x' \in F_P$ uniquely defines an equivalence class $x' + L_P \in \mathbb{R}^n/L_P$, and every $x \in \mathbb{R}^n/L_P$ can be written as $x = x' + L_P$ for some $x' \in F_P$. A fundamental region of L_P can be identified with $\mathbb{R}^n/L_P$. With $[0,1)^n$ denoting the n-fold Cartesian product of $[0,1)$, the set $P[0,1)^n = \{Px \mid x \in [0,1)^n\}$ is the canonical example of a fundamental region for the lattice L_P. From [9, Theorem 1], we have the following.

THEOREM 2.1. *Let $L_P \subseteq L_W$. When $\mathbb{R}^n/L_W$ is identified with $W[0,1)^n$, the set $\mathbb{R}^n/L_P$ can be identified with the direct sum $\mathbb{R}^n/L_W \oplus L_W/L_P$.*

DEFINITION 2.2. *A function f on $\mathbb{R}^n$ is said to be W-bandlimited with bandregion K if $\hat{f} = 0$ outside some compact set $K \in \mathbb{R}^n$, and $(K^\circ + \xi) \cap (K^\circ + \xi') = \emptyset$ for $\xi, \xi' \in L_W^\perp$, and $\xi \neq \xi'$. K° denotes the interior of K.*

A square integrable, continuous, W-bandlimited function can be reconstructed from its samples on L_W. In particular, Parseval's formula holds.

THEOREM 2.3. *Let $f_1, f_2 \in L^2(\mathbb{R}^n)$ be W-bandlimited and continuous, both with the same bandregion K. Then*

$$\int_{\mathbb{R}^n} f_1(x)\overline{f_2(x)}dx = |\det W| \sum_{x \in L_W} f_1(x)\overline{f_2(x)}. \tag{2.5}$$

DEFINITION 2.4. *A function on $\mathbb{R}^n$ is periodic with respect to the lattice L_Q (or alternatively Q-periodic) if, for all $x \in L_Q$ and $y \in \mathbb{R}^n$, $f(y+x) = f(y)$.*

A Q-periodic function f can be sampled on a lattice L_W if and only if $L_Q \subseteq L_W$. The sample points can be treated as elements of L_W/L_Q. The Fourier and inverse Fourier transforms in the periodic setting are

$$\hat{f}(\eta) = \frac{1}{|\det Q|} \int_{\mathbb{R}^n/L_Q} f(x) e^{-ix\eta} dx, \quad \eta \in L_Q^\perp, \tag{2.6}$$

$$f(x) = \sum_{\eta \in L_Q^\perp} \hat{f}(\eta) e^{ix\eta}. \tag{2.7}$$

The definition of bandlimited functions in the periodic setting requires the use of the discrete topology. In this case, one cannot ignore the boundary of K when considering shifts.

DEFINITION 2.5. *Let $L_Q \subseteq L_W$, and $f \in L_2(\mathbb{R}^n/L_Q)$. f is W-bandlimited if $\hat{f} = 0$ outside some compact set $K \subseteq L_Q^\perp$, and $(K+\xi) \cap (K+\xi') = \emptyset$ for $\xi, \xi' \in L_W^\perp$, and $\xi \neq \xi'$.*

Parseval's formula in the periodic setting is the following.

THEOREM 2.6. *Let $f_1, f_2 \in L^2(\mathbb{R}^n/L_Q)$ be W-bandlimited and continuous, both with the same bandregion K. Then*

$$\int_{L^2(\mathbb{R}^n/L_Q)} f_1(x)\overline{f_2(x)} dx = |\det W| \sum_{x \in L_W/L_Q} f_1(x)\overline{f_2(x)}. \tag{2.8}$$

Let the lattice L_P satisfy $L_Q \subseteq L_P \subseteq L_W$, with $r = \frac{|\det P|}{|\det W|}$. Under suitable conditions, a Q-periodic, W-bandlimited function with bandregion K can be recovered from samples on L_P/L_Q of r filtered versions of f (channels) [7, 9]. Following [7], when each channel g_k is of the form $g_k = f * \delta(x - \gamma_k)$ for independent translations $\gamma_k \in \mathbb{R}^n/L_Q$, L_2 inner products of two such functions can be computed using Theorem 2.6 after upsampling by interpolation from L_P to L_W. Alternatively, following [9], if the lattice L_P undersamples along one generator of L_W, and if K is shift-convex with respect to $W^\perp$ and $P^\perp$, a Parseval-like result will hold. The definitions for shift-convexity and independent translations follow.

Let $\mathbf{i} \in \mathbb{Z}^n$ denote the multi-index $\mathbf{i} = (i_1, \ldots, i_n)^T$.

DEFINITION 2.7. *Let $L_Q \subseteq L_P \subseteq L_W$. Suppose that K is the bandregion for a W-bandlimited Q-periodic function, and that $P = WM$, where M is a diagonal matrix with positive integer entries, $M = \mathrm{diag}(r_1, \ldots, r_n)$. K is said to be shift-convex with respect to $W^\perp$ and $P^\perp$ if $K \cap (K + P^\perp \mathbf{i}) = \emptyset$ for all $\mathbf{i}$ such that $-(\mathbf{r}-1) \leq \mathbf{i} \leq \mathbf{r}-1$ does not hold.*

The idea of shift-convexity is to eliminate the pathological cases where a suitably sampled function becomes undersampled when the sample density increases. W-bandlimiting implies that shifts of K by nonzero elements of $L_W^\perp$ do not overlap K. Roughly speaking, any additional shifts of the bandregion K by elements of $L_P^\perp$ do not re-introduce overlap with K. For two examples, see section 3.2.3.

Now consider $L_Q \subseteq L_P \subseteq L_W$. Fix K, the bandregion for a W-bandlimited Q-periodic function. Since each coset $\xi \in L_Q^\perp/L_W^\perp$ intersects K in at most one point, when $K \cap \xi \neq \emptyset$, the point of intersection is a canonical representative $\xi' \in L_Q^\perp$ for $\xi = \xi' + L_W^\perp \in L_Q^\perp/L_W^\perp$. When $K \cap \xi$ is empty, the fixed representative ξ' can be chosen arbitrarily. As $L_P^\perp/L_W^\perp \subseteq L_Q^\perp/L_W^\perp$, the same construction provides a canonical representative $\zeta' \in L_P^\perp$ for each coset $\zeta = \zeta' + L_W^\perp \in L_P^\perp/L_W^\perp$.

DEFINITION 2.8. *Let ζ_j, $j = 0, \ldots, r-1$, be the r distinct equivalence classes in $L_P^\perp / L_W^\perp$. Let $\zeta_0 = L_W^\perp$, and ζ_j' be the canonical representative for ζ_j. The translations $\gamma_k \in \mathbb{R}^n / L_Q$, $k = 0, \ldots, r-1$, are independent if the matrix $\mathbf{H}$, with k, jth entry $\mathbf{H}_{k,j} = e^{-i\zeta_j' \gamma_k}$, is invertible. Let $\lambda_{k,j}$ denote the k, jth entry of $\mathbf{H}^{-T}$.*

Since $\zeta_j' \in L_P^\perp \subseteq L_Q^\perp$, $\mathbf{H}_{k,j}$ is independent of the representative chosen for $\gamma_k \in \mathbb{R}^n / L_Q$, and thus is well defined.

THEOREM 2.9. *Let $f_1, f_2 \in L_2(\mathbb{R}^n / L_Q)$ be W-bandlimited and continuous, both with bandregion K, and with $f_1 \overline{f_2}$ real. Let $L_Q \subseteq L_P \subseteq L_W$, with r equivalence classes in $L_P^\perp / L_W^\perp$. Moreover, suppose that K is shift-convex with respect to $W^\perp$ and $P^\perp$. Let the r filters h_r, $k = 0, \ldots, r-1$, implement independent translations $\gamma_k \in \mathbb{R}^n / L_Q$ such that λ_{k0}, $k = 0, \ldots, r-1$, are all real. Then,*

$$\int_{\mathbb{R}^n / L_Q} f_1(x) \overline{f_2}(x) dx = |\det P| \sum_{k=0}^{r-1} \lambda_{k0} \sum_{y \in L_P / L_Q} f_1(y - \gamma_k) \overline{f_2}(y - \gamma_k). \tag{2.9}$$

Proof. This is Theorem 10 in [9]. □

We remark that another method [6, 7] is available for computing the L_2 inner product of W-bandlimited functions from samples on L_P. That method does not require the restrictive hypotheses of shift-convexity, one generator subsampling, or the reality conditions of Theorem 2.9. In that approach, the missing samples of f_1 and f_2 on L_W needed for the application of Theorem 2.6 are obtained by interpolation. The interpolation is accomplished with a generalized sampling expansion. On the other hand, when Theorem 2.9 can be used, the computation is simpler and more direct, as no interpolation is required.

2.2. Divergent beam geometry. In [11], the sampling criteria for a standard fan beam reconstruction of an essentially Ω-bandlimited function were derived. The divergent beam transform is periodic in both arguments, so for $(k, a) \in \mathbb{Z} \times 2\mathbb{Z}$, the Fourier transform of Df is

$$\widehat{Df}(k, a) = \frac{1}{4\pi^2} \int_0^{2\pi} \int_0^{\pi} Df(\beta, \alpha) e^{-i(k\beta + a\alpha)} d\alpha d\beta. \tag{2.10}$$

THEOREM 2.10. *Let $f \in \mathcal{S}(\mathbb{R}^2)$ have support $|x| \leq \rho$ and be essentially Ω-bandlimited. Then the essential support of $\widehat{Df}(k, a)$ is the set*

$$K = \left\{ (k, a) \in \mathbb{Z} \times 2\mathbb{Z} \,\middle|\, |k - a| \leq \Omega r, |k| r \leq |k - a| \rho \right\}. \tag{2.11}$$

An essentially Ω-bandlimited reconstruction of f from fan beam data is achieved by converting the parallel beam reconstruction, equation (5.22) in [13], to fan beam coordinates:

$$V_\Omega * f(x) = r \int_0^{2\pi} \int_{-\pi/2}^{\pi/2} v_\Omega(x\Phi - r \sin\alpha) Df(\beta, \alpha) \cos\alpha d\alpha d\beta, \tag{2.12}$$

where Φ is the unit vector associated with the angle $\phi = \beta + \alpha - \pi/2$, v_Ω is an Ω-bandlimited approximation to the ramp filter, and V_Ω is the corresponding Ω-bandlimited approximate δ function. That is, for a filter $\hat{\psi}(\sigma)$ which is supported on $|\sigma| \leq 1$, and which is approximately unity on its support,

$$\begin{aligned} \hat{V}_\Omega(\xi) &= (2\pi)^{-1} \hat{\psi}(|\xi| / \Omega), \\ \hat{v}_\Omega(\sigma) &= \frac{(2\pi)^{-3/2}}{2} |\sigma| \hat{\psi}(\sigma / \Omega). \end{aligned} \tag{2.13}$$

In order to accurately compute (2.12) by a discrete sum, Df must be suitably sampled.

2.2.1. Standard divergent beam geometry. Sampling density criteria for the standard fan beam reconstruction (Algorithm 5.3 in [13]) result when translates of K parallel to the coordinate axes are packed as densely as possible. See Figure 2. For

$$W = \begin{pmatrix} \Delta\beta & 0 \\ 0 & \Delta\alpha \end{pmatrix}, \qquad \Delta\beta \leq \frac{r+\rho}{r}\frac{\pi}{\rho\Omega}, \qquad \Delta\alpha \leq \frac{\pi}{r\Omega}, \tag{2.14}$$

the translates of K by $L_W^{\perp}$ do not overlap [11, 15]. Since the lattice L_W must be a sublattice of $2\pi\mathbb{Z} \times \pi\mathbb{Z}$, when (2.14) holds with equality, both $r\Omega$ and $\frac{r+\rho}{2r\rho\Omega}$ are required to be integral. The sample points are (β_j, α_ℓ), with

$$\beta_j = j\Delta\beta, \qquad \alpha_\ell = (\ell + \delta)\Delta\alpha. \tag{2.15}$$

The parameter δ is the detector shift. This parameter indicates the alignment of the elements in the detector assembly with respect to the central ray. That is, the center of each detector element will be shifted by $\delta\Delta\alpha$ with respect to the lattice $\Delta\alpha\mathbb{Z}$. We remark that, as with the standard parallel beam sample lattice, the standard fan beam sample lattice is sufficiently dense along the detector direction to allow the inner integral in the reconstruction formula (2.12) to be replaced by a sum of discrete samples. Also, the translations of K do not completely fill $\mathbb{R}^2$ so this sampling scheme is not efficient.

2.2.2. Efficient geometry. An efficient sample scheme can be obtained by tiling $\mathbb{R}^2$ with translations of K. Translates by $L_W^{\perp}$ will tile the plane if the lattice L_W is generated by

$$W = \begin{pmatrix} \Delta\beta & \dfrac{r-\rho}{2r}\Delta\beta \\ 0 & \Delta\alpha \end{pmatrix}, \qquad \Delta\beta = \frac{2\pi}{\rho\Omega}, \qquad \Delta\alpha = \frac{\pi}{r\Omega}. \tag{2.16}$$

See Figure 2. Here, both $r\Omega$ and $\rho\Omega$ are required to be integral. To acquire data for this efficient sampling, a dynamic detector shift is required. An implementation of such an acquisition system presents many technical difficulties. See [11] for a detailed treatment of the efficient geometry.

2.2.3. Reflected geometry. Another option for a fan-beam acquisition is to use only the reflections of the sample points. This corresponds to choosing for W,

$$W = \begin{pmatrix} \Delta\beta & 2\Delta\alpha \\ 0 & -\Delta\alpha \end{pmatrix}, \tag{2.17}$$

with $\Delta\beta$ and $\Delta\alpha$ the same as in (2.14). The translates of K by $L_W^{\perp}$ are shown in Figure 2. When the original sample points are given by (2.15) and $\lambda = \Delta\beta/\Delta\alpha$, the reflected samples $(\beta'_{j,\ell}, \alpha'_\ell)$ are

$$\beta'_{j,\ell} = (J/2 + j + 2(\ell + \delta)/\lambda)\Delta\beta, \qquad \alpha'_\ell = -(\ell + \delta)\Delta\alpha, \tag{2.18}$$

where $J = 2\pi/\Delta\beta$. Because L_W must be 2π-periodic in β and π-periodic in α, λ is rational, so the reflected geometry has an interpretation as a dynamic detector shift with the periodicity determined by the denominator after $2/\lambda$ is reduced to lowest terms. A reconstruction algorithm similar to that in section 3.2 can be given for this geometry. However, as this algorithm will be significantly more time consuming than the standard fan beam algorithm, there is no practical incentive to use it.

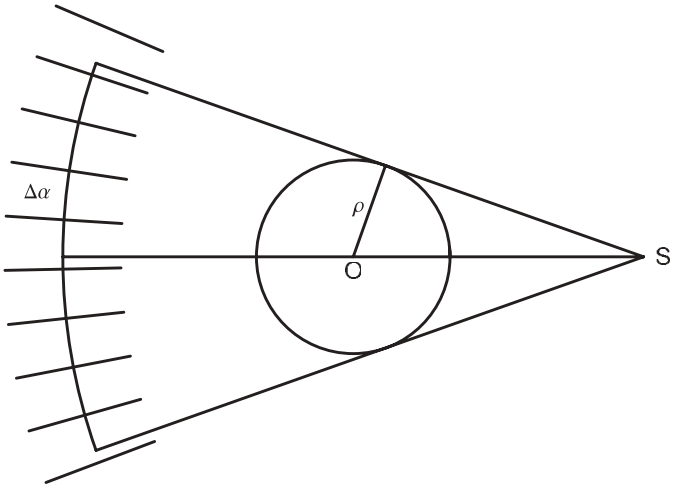


FIG. 3. *A model of the source-detector assembly is shown. All the photons arriving at an interval of width $\Delta\alpha$ between the ticks contribute to the measurement sample located at the center of that interval. The central ray of the fan intersects the central interval of the detector at $\delta\Delta\alpha$ from the center of the interval.*

3. The third generation problem and the quarter detector shift. For a typical third generation CT scanner, the spacing between detector pixels is the limiting factor determining the maximum attainable resolution in reconstructed images. That is, the size of the smallest reconstructible detail, π/Ω, is determined by the detector spacing, $\Delta\alpha$. This in turn determines the source sampling interval, $\Delta\beta$, needed to achieve the π/Ω resolution.

3.1. The detector model. The detector assembly for a third generation CT scanner can be modeled as a contiguous collection of detector elements, each of width $\Delta\alpha$. See Figure 3. For simplicity, it is assumed that each detector element will respond uniformly to all photons arriving within $\Delta\alpha/2$ of the element center. Other models could be considered. Accordingly, the measured data $D'f(\beta,\alpha)$ is a convolution of Df with the detector element response function $(\Delta\alpha)^{-1}\chi_{[-\Delta\alpha/2,\Delta\alpha/2]}$, where χ_I denotes the characteristic function for the interval I:

(3.1)
$$D'f(\beta,\alpha) = (\Delta\alpha)^{-1}\int_{-\Delta\alpha/2}^{\Delta\alpha/2} Df(\beta,\alpha-t)dt = (\Delta\alpha)^{-1} Df * \chi_{[-\Delta\alpha/2,\Delta\alpha/2]}(\beta,\alpha).$$

The Fourier transform of $(\Delta\alpha)^{-1}\chi_{[-\Delta\alpha/2,\Delta\alpha/2]}$ is

(3.2)
$$\hat{\chi}_{[-\Delta\alpha/2,\Delta\alpha/2]}(\xi) = (2\pi)^{-1/2}\mathrm{sinc}(\xi\Delta\alpha/2)/2.$$

Ignoring the contribution from the side lobes of the sinc function allows the finite width detector to be treated as a low-pass filter, effectively making the measured data $2\pi/\Delta\alpha$-bandlimited in the α direction. This hardware-determined bandwidth sets an intrinsic resolution limit for reconstructions from data measured by this acquisition system. On the other hand, a separation of $\Delta\alpha$ between samples implies a maximum sampling-determined bandwidth of $\pi/\Delta\alpha$ in the α direction, which by (2.14), implies a maximum reconstruction resolution of $\pi/\Omega = r\Delta\alpha$, where $\Omega = \pi/(r\Delta\alpha)$. As the

bandwidth in the measured data is twice that which can be exploited by the standard sample scheme, it is natural to ask whether it is possible to reconstruct at the doubled resolution corresponding to the higher data bandwidth. Since this is a two-dimensional reconstruction, doubling the resolution requires a fourfold increase in the number of samples. In the source direction, β, a density doubling can be achieved by finer sampling. An increase in the sample density in the detector direction will be achieved by the judicious use of the reflected sample points.

As previously mentioned, for the parallel beam geometry, the reflected sample points mesh well with the original sample points, resulting in a relatively straight forward algorithm to reconstruct at the hardware-limited resolution. The divergent beam geometry requires a more delicate analysis.

3.2. Symmetry in the divergent beam transform. Following the success in the parallel beam context of using reflected sample points to reconstruct at the hardware-limited resolution, it is natural to ask how to perform the analogous computation for the divergent beam geometry. Unfortunately, with the divergent beam transform and a 1/4 detector shift, the reflected sample points do not align in a rectangular grid, so the straightforward application of the standard divergent beam (Algorithm 5.3 in [13]) can only be viewed, at best, as an approximation. A reconstruction at the hardware-limited resolution can be obtained if the combined original and reflected sets are together viewed as samples on multiple channels for a suitably chosen undersampled lattice. In the language of signal processing, the combined lattices are a periodic, nonuniform sampling set. The undersampled lattice and the number of channels needed depend on the ratio $\theta = \rho/r$.

It is now assumed that $D'f$ is available on the original lattice L_O,

$$L_O = \{(\beta_j, \alpha_\ell) \mid j \in \mathbb{Z}, \ell \in \mathbb{Z}\}, \qquad \beta_j = j\Delta\beta, \quad \alpha_\ell = (\ell + \delta)\Delta\alpha. \tag{3.3}$$

Since L_O must be a sublattice of $2\pi\mathbb{Z} \times \pi\mathbb{Z}$, both of

$$\frac{\pi}{\Delta\alpha} = n \quad \text{and} \quad \frac{2\pi}{\Delta\beta} = m, \tag{3.4}$$

are integral.

Here $\Delta\alpha$ is the detector pixel width and $\Delta\beta$ is the angular sampling interval to be specified shortly. In practice, only $j = 0, \ldots, m-1$ and $\ell = -L, \ldots, L$ for $L > \frac{\sin^{-1}\rho/r}{\Delta\alpha}$ are needed as $D'f$ either is 0 or can be computed by periodicity from the other points.

The resolution implied by the detector sample density is π/Ω, where

$$\Omega \leq \frac{\pi}{r\Delta\alpha}. \tag{3.5}$$

The hardware-limited bandwidth of the data from the third generation detector is 2Ω. A reconstruction at the corresponding resolution will require that $\Delta\beta$ be chosen appropriately. That is,

$$\Delta\beta \leq \frac{r+\rho}{r\rho}\frac{\pi}{2\Omega}. \tag{3.6}$$

The divergent beam reflection property (2.3) is applied to the sample lattice L_O with $\Delta\alpha$ and $\Delta\beta$ given by (3.5) and (3.6), respectively. The reflected sample points are

$$L_R = \left\{(\beta'_{j,\ell}, \alpha'_\ell) \mid j \in \mathbb{Z}, \ell \in \mathbb{Z}\right\}, \tag{3.7}$$

where

$$\beta'_{j,\ell} = \pi + j\Delta\beta + 2(\ell+\delta)\Delta\alpha, \qquad \alpha'_\ell = -(\ell+\delta)\Delta\alpha. \tag{3.8}$$

If $\delta = 0$, then $\beta'_{j,\ell} = \pi + j\Delta\beta + 2\ell\Delta\alpha$ and $\alpha'_\ell = -\ell\Delta\alpha$. Regardless of the possibly increased sample density in the β direction, the α coordinates of the sample points in L_O match those of L_R. As a result, the translates of K in the direction dual to α still overlap, and the data remain insufficient to reconstruct at the higher resolution $\pi/(2\Omega)$. The same occurs for $\delta = \pm 1/2$.

Duplication is avoided and the α coordinates of the reflected points interlace perfectly with the α coordinates of the original samples when $\delta = 1/4$. With $\delta = 1/4$, (3.8) becomes

$$\beta'_{j,\ell} = \pi + j\Delta\beta + (2\ell + 1/2)\Delta\alpha, \qquad \alpha'_\ell = -(\ell + 1/4)\Delta\alpha. \tag{3.9}$$

From (3.4), $\Delta\alpha = \frac{m}{2n}\Delta\beta$ and

$$\beta'_{j,\ell} = \pi + \left(j + \frac{(4\ell+1)m}{4n}\right)\Delta\beta. \tag{3.10}$$

The reflected sample set L_R has a static shift in β of $\pi + \frac{m}{4n}\Delta\beta$ in addition to a dynamic shift of $m\Delta\beta/n$, which is periodic in ℓ. This is illustrated in Figure 4.

To work with this sample set, the union of the original and the reflected sample points is recast as a union of identical rectangular lattices, each shifted by a static distance from the origin. The β separation between lattice points of these rectangular lattices will be $\Delta\beta$, and the α separation will be a multiple of $\Delta\alpha$. Each rectangular lattice provides one channel of a multichannel sampling. Within this framework, Theorem 2.9 can be applied to accurately compute (2.12) as a discrete sum of the sampled divergent beam transform data.

3.2.1. Sample points as a union of rectangular lattices. In this subsection we show that the original sample points together with the reflected points can be reorganized as multiple shifted copies of a rectangular lattice. Thus, sampling on $L_O \cup L_R$ defines a periodic, nonuniform sampling set.

THEOREM 3.1. *Let L_O be the sample lattice arising from a quarter detector shift with a source sampling at twice the standard density, and let L_R be the corresponding reflected lattice. Then, the union, $L_O \cup L_R$, is a disjoint union of $2n'$ translates of the lattice L_P,*

$$L_P = \begin{pmatrix} \Delta\beta & 0 \\ 0 & n'\Delta\alpha \end{pmatrix} \mathbb{Z}^2, \tag{3.11}$$

where n' is the denominator when m/n is reduced to lowest terms.

Proof. The lattices L_O and L_R are given by (3.3), (3.4), (3.7), and (3.10).

Since L_O must be a sublattice of $2\pi\mathbb{Z} \times \pi\mathbb{Z}$, both of

$$n = \frac{\pi}{\Delta\alpha} \geq r\Omega \tag{3.12}$$

and

$$m = \frac{2\pi}{\Delta\beta} \geq \frac{4r\rho\Omega}{r+\rho} \tag{3.13}$$

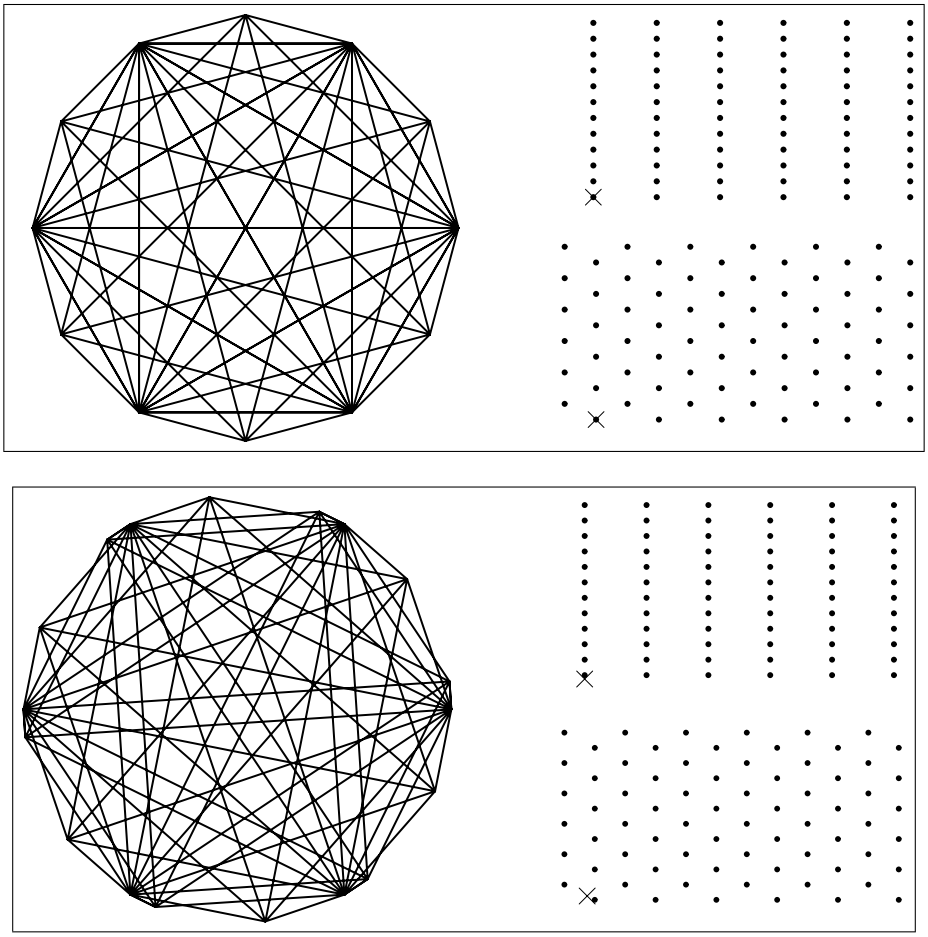


FIG. 4. *For a divergent beam scanner with $\rho/r = 1/3$, the doilygram (the set of lines on which line integrals are available), the standard sample lattice L_O (upper right), and the reflected sample lattice L_R (lower right) are shown for a detector shift of $\delta = 0$ (top) and $\delta = 1/4$ (bottom). On the lattices, the source position (β) is the horizontal coordinate and position along the detector (α) is the vertical coordinate. The origin is marked with $\times$. The doilygram for $\delta = 1/4$ appears denser than $\delta = 0$ because each point in the sample lattice corresponds to a unique line in the doilygram. For $\delta = 0$, some lines are generated by two points in the lattice. Samples of Df on any of the lattices shown can be used to reconstruct an essentially Ω-bandlimited function from its divergent beam transform. The lattices are shown for $\Omega = 8$.*

are integral. Let m/n be expressed in lowest terms as m'/n'. That is,

$$m' = \frac{m}{\gcd(m,n)}, \qquad n' = \frac{n}{\gcd(m,n)}. \tag{3.14}$$

As ℓ varies, a shift $\frac{m'}{n'}\Delta\beta$ is induced in $\beta'_{j,\ell}$. This shift is a rational multiple of the sample interval $\Delta\beta$. From (3.10) and (3.14),

$$\beta'_{j,\ell} = \pi + \beta_{j+\frac{m'}{n'}(\ell+1/4)}, \tag{3.15}$$

and for any j_1, j_2, ℓ_1, ℓ_2,

$$\beta'_{j_1,\ell_1} - \beta'_{j_2,\ell_2} = \left(j_1 - j_2 + \frac{(\ell_1 - \ell_2)m'}{n'} \right) \Delta\beta, \tag{3.16}$$

which is an integral multiple of $\Delta\beta$ if and only if $\ell_1 \equiv \ell_2 \mod n'$. Hence, for each $\ell_0 \in \mathbb{Z}$, the set $L_{\ell_0 R} = \{(b'_{j,\ell}, \alpha'_\ell) \mid \ell \equiv \ell_0 \mod n', j \in \mathbb{Z}\}$ is a shifted copy of the rectangular lattice L_P,

$$L_{\ell_0 R} = \begin{pmatrix} \pi + \left(\frac{(\ell_0 + 1/4)m'}{n'}\right)\Delta\beta \\ (-1/4 - \ell_0)\Delta\alpha \end{pmatrix} + L_P. \tag{3.17}$$

In view of (3.16), a distinct lattice is obtained for each equivalence class in $\mathbb{Z}/(n'\mathbb{Z})$. Therefore, the reflected sample set L_R is the disjoint union of the distinct lattices,

$$L_R = \bigcup_{[\ell_0] \in \mathbb{Z}/(n'\mathbb{Z})} L_{\ell_0 R}. \tag{3.18}$$

The original sample set L_O can easily be represented as a disjoint union of n' shifted copies of the lattice L_P,

$$L_O = \bigcup_{[\ell_0] \in \mathbb{Z}/(n'\mathbb{Z})} \begin{pmatrix} 0 \\ (\ell_0 + 1/4)\Delta\alpha \end{pmatrix} + L_P. \tag{3.19}$$

Combining (3.18) and (3.19) completes the proof. □

Three examples are shown in Figure 5.

3.2.2. The divergent beam multichannel sampling algorithm. Equation (2.12) with Ω replaced by 2Ω is used to reconstruct a bandlimited approximation to f from Divergent beam transform data. To properly compute the continuous integral from discrete data, Theorem 2.9 must be applied.

First, the relevant lattices are identified. $D'f$ is Q-periodic with

$$L_Q = 2\pi\mathbb{Z} \times \pi\mathbb{Z}, \qquad L_Q^\perp = \mathbb{Z} \times 2\mathbb{Z}. \tag{3.20}$$

As $D'f$ is essentially W-bandlimited with K as in (2.11) when Ω is replaced by 2Ω, a reconstruction at the hardware-limited resolution is possible when samples of $D'f$ are available on L_W/L_Q, where, with m and n as in (3.12) and (3.13),

$$L_W = \left(\frac{2\pi}{m}\right)\mathbb{Z} \times \left(\frac{\pi}{2n}\right)\mathbb{Z}, \qquad L_W^\perp = m\mathbb{Z} \times 4n\mathbb{Z}. \tag{3.21}$$

With $\beta_s = s\Delta\beta$ and $\alpha_\ell = \ell\Delta\alpha$, we make the identification

$$L_W/L_Q = \left\{ (\beta_s, \alpha_\ell) \mid s = 0, \ldots, m-1, \ell = \frac{-n}{2}, \frac{-n+1}{2}, \ldots, \frac{n-2}{2}, \frac{n-1}{2} \right\}. \tag{3.22}$$

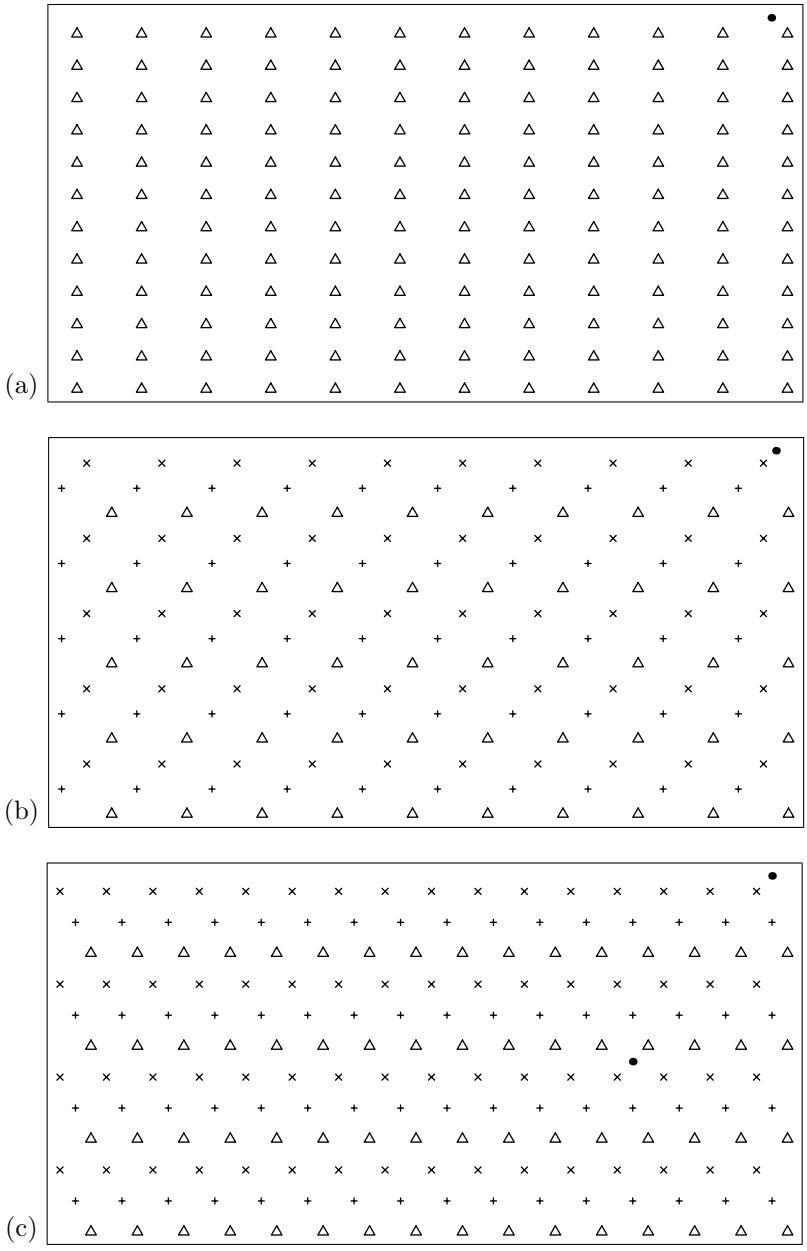


FIG. 5. *Lattices are shown for $\theta = 1/3, 1/5, 1/2$, respectively. For each case, the original lattice L_O is shown by dots. For* (a), *$n' = 1$. The reflected lattice L_R is rectangular and is shown by triangles. For both* (b) *and* (c), *L_R is a union of $n' = 3$ rectangular lattices, shown by the triangles, $+$, and $\times$. For each case, the horizontal spacing of all the rectangular lattices is the same, and the vertical spacing of the rectangular lattices making up L_R is n' times that of L_O.*

By Theorem 3.1, the original and reflected lattices together can be realized as a union of $2n'$ copies of the lattice L_P,

$$L_P = \left(\frac{2\pi}{m}\right)\mathbb{Z} \times \left(\frac{n'\pi}{n}\right)\mathbb{Z}, \qquad L_P^{\perp} = m\mathbb{Z} \times \left(\frac{2n}{n'}\right)\mathbb{Z}, \tag{3.23}$$

the kth copy of L_P being shifted by γ_k. For $k = 0, \ldots, n' - 1$,

$$\gamma_k = \begin{pmatrix} 0 \\ (k+1/4)\frac{\pi}{n} \end{pmatrix}, \qquad \gamma_{k+n'} = \begin{pmatrix} \pi + (k+1/4)\frac{2\pi}{n} \\ -(k+1/4)\frac{\pi}{n} \end{pmatrix}. \tag{3.24}$$

For $\zeta \in L_P^\perp / L_W^\perp$, $\zeta = (u, 2v)^T$, with $u \in m\mathbb{Z}/m\mathbb{Z}$ and $2v \in \frac{2n}{n'}\mathbb{Z}/4n\mathbb{Z}$. To find a representative $\zeta' = (u', v')$ for ζ, observe that if $u' = mt$, with $0 \neq t \in \mathbb{Z}$, then $|u'| \geq |t|\frac{4\pi\rho\Omega}{r+\rho}$ which implies $(mt, v') \notin K$. Hence, $\zeta' = (0, 2v')$. Next, by (2.11), $|2v'| < 2\Omega r$. Any other representative for ζ is of the form $\zeta' = (0, 2v + 4n\tau)$ with $0 \neq \tau \in \mathbb{Z}$. However, $|2v + 4n\tau| \geq |4n\tau| - |2v| > 4n - 2r\Omega \geq 2r\Omega$. Hence, $\zeta' = (0, 2v + 4n\tau) \notin K$. Accordingly, we may take $\zeta' = (0, 2jn/n')$ for $j = -n', \ldots, n' - 1$. Note that the symmetry of K allows an equivalent choice of $j = -n' + 1, \ldots, n'$, and that $\zeta_j \cap K = \emptyset$ for either indexing scheme.

The samples of $D'f$ on each $\gamma_k + L_P$ are samples of the measurement channel, $g_k = h_k * f$, where $h_k(x) = \delta(x - \gamma_k)$. For $\eta \in L_Q^\perp / L_P^\perp$, $\eta = (u, 2v)^T$, with $u \in \mathbb{Z}/m\mathbb{Z}$ and $2v \in (2\mathbb{Z})/(\frac{2n}{n'}\mathbb{Z})$. The k, j element of $\mathbf{H}$ in Definition 2.8 is given by

$$\mathbf{H}_{k,j} = e^{-i\zeta_j' \gamma_k} = e^{-i(0, 2jn/n')\gamma_k}. \tag{3.25}$$

So, for $k = 0, \ldots, n' - 1$,

$$\mathbf{H}_{k,j} = e^{-i(2jn/n')(k+1/4)\pi/n} = e^{-i(k+1/4)2\pi j/n'}, \tag{3.26}$$

$$\mathbf{H}_{k+n',j} = e^{i(2jn/n')(k+1/4)\pi/n} = e^{i(k+1/4)2\pi j/n'}. \tag{3.27}$$

Since

$$\sum_{j=-n'}^{n'-1} (\mathbf{H})_{k_1,j} (\overline{\mathbf{H}})_{k_2,j} = 2n' \delta_{k_1,k_2}, \tag{3.28}$$

$\mathbf{H}^{-T}$ is explicitly computed as $\mathbf{H}^{-T} = \frac{1}{2n'}\overline{\mathbf{H}}$, and therefore, $\lambda_{k,0} = \frac{1}{2n'}$, which is real.

To apply Theorem 2.9, the essential bandregion K, given by (2.11) with Ω replaced by 2Ω, must be shift-convex with respect to $W^\perp$ and $P^\perp$. The discussion of the circumstances under which K is shift-convex will be postponed to section 3.2.3. For now, we proceed under the assumption that K is indeed shift-convex.

In the present setting, Theorem 2.9 reads

$$\begin{aligned} &\int_0^{2\pi} \int_{-\pi/2}^{\pi/2} f_1(\beta, \alpha) f_2(\beta, \alpha) d\alpha d\beta \\ &\quad = \frac{2\pi^2 n'}{mn} \sum_{k=0}^{2n'-1} \frac{1}{2n'} \sum_{(\beta,\alpha) \in \frac{(2\pi/m)\mathbb{Z}}{2\pi\mathbb{Z}} \times \frac{(n'/n)\mathbb{Z}}{\pi\mathbb{Z}}} f_1((\beta, \alpha) - \gamma_k) f_2((\beta, \alpha) - \gamma_k) \\ &\quad = \frac{\pi^2}{mn} \sum_{(\beta,\alpha) \in \frac{L_Q \cup L_R}{2\pi\mathbb{Z} \times \pi\mathbb{Z}}} f_1(\beta, \alpha) f_2(\beta, \alpha). \end{aligned} \tag{3.29}$$

Applying (3.29) to (2.12) with $f_1(\beta, \alpha) = v_{2\Omega}(x\Phi - r \sin\alpha)\cos\alpha$ and $f_2(\beta, \alpha) = Df(\beta, \alpha)$ allows the recovery of the 2Ω bandlimited version of f by a discrete sum of

samples of the hardware-limited $D'f$,

$$
\begin{aligned}
V_{2\Omega} * f(x) &= \frac{r\pi^2}{mn} \sum_{(\beta,\alpha)\in \frac{(2\pi/m)\mathbb{Z}}{2\pi\mathbb{Z}} \times \frac{(n'/n)\mathbb{Z}}{\pi\mathbb{Z}}} v_{2\Omega}(x\Phi - r\sin\alpha) D'f(\beta,\alpha)\cos\alpha \\
&= \frac{r\pi^2}{mn} \sum_{s=0}^{m-1} \sum_{\ell=-n}^{n-1} \Big(v_{2\Omega}(x\Phi_{s,\ell} - r\sin\alpha_\ell) D'f(\beta_s,\alpha_\ell)\cos\alpha_\ell \\
&\qquad + \; v_{2\Omega}(x\Phi'_{s,\ell} - r\sin\alpha'_\ell) D'f(\beta'_{s,\ell},\alpha'_\ell)\cos\alpha'_\ell \Big) \qquad (3.30) \\
&= \frac{r\pi^2}{mn} \sum_{s=0}^{m-1} \sum_{\ell=-n}^{n-1} \Big(v_{2\Omega}(x\Phi_{s,\ell} - r\sin\alpha_\ell)\cos\alpha_\ell \\
&\qquad + \; v_{2\Omega}(x\Phi'_{s,\ell} - r\sin\alpha'_\ell)\cos\alpha'_\ell \Big) D'f(\beta_s,\alpha_\ell),
\end{aligned}
$$

where $\Phi_{s,\ell}$ and $\Phi'_{s,\ell}$ are the unit vectors associated with $\phi_{s,\ell} = \beta_s + \alpha_\ell - \pi/2$ and $\phi'_{s,\ell} = \beta'_{s,\ell} + \alpha'_\ell - \pi/2$, respectively.

The reconstruction exhibited in (3.30) has a computational complexity of $O(\Omega^4)$. Following [13], to develop an algorithm with a more tractable complexity of $O(\Omega^3)$, it is necessary to replace (2.12) by the approximation, valid when both $|x| \ll r$ and $V_{2\Omega}(k)$ is close to zero for $|k|$ near 2Ω [14],

$$
V_{2\Omega} * f(x) = r \int_0^{2\pi} |b-x|^{-2} \int_{-\pi/2}^{\pi/2} v_{2\Omega r}(\sin(\nu(b,x)-\alpha))\cos\alpha D'f(\beta,\alpha)\, d\alpha d\beta. \qquad (3.31)
$$

Here, $b = (r\cos\beta, r\sin\beta)^T$, $b_\perp = (r\cos(\beta+\pi/2), r\sin(\beta+\pi/2))^T$, and

$$
\cos\nu(b,x) = \frac{(b-x)b}{|b-x||b|}, \quad \nu(b,x) \lesseqgtr 0 \quad \text{when } xb_\perp \gtreqless 0. \qquad (3.32)
$$

By Theorem 2.9 and (3.29), the double integral can be computed as a discrete sum. For $|x| < \rho$,

$$
\begin{aligned}
V_{2\Omega} * f(x) &= \frac{r\pi^2}{mn} \sum_{s=0}^{m-1} |b_s - x|^{-2} \sum_{\ell=-n}^{n-1} v_{2\Omega r}(\sin(\nu(b_s,x)-\alpha_\ell))\cos\alpha_\ell D'f(\beta_s,\alpha_\ell) \\
&\quad + \frac{r\pi^2}{mn} \sum_{s=0}^{m-1} \sum_{\ell=-n}^{n-1} |b'_{s,\ell} - x|^{-2} v_{2\Omega r}(\sin(\nu(b'_{s,\ell},x)-\alpha'_\ell))\cos\alpha'_\ell D'f(\beta_s,\alpha_\ell) \\
&= T_1(x) + T_2(x).
\end{aligned} \qquad (3.33)
$$

The first term on the right-hand side, T_1, in (3.33) can be computed by adapting the method to interpolate convolutions suggested in [5] to the divergent beam geometry. As the inner sum is not sufficiently sampled to be a valid approximation to the corresponding integral, artifacts are introduced in the interpolation step. To avoid the artifacts, instead of interpolation with a step size of $\Delta\alpha$, the interpolation is done with a finer step size of $\Delta\alpha/M$ for a sufficiently large M.

As written, the discrete convolution argument cannot be applied to the second term T_2 in (3.33) because $b'_{s,\ell}$ depends on ℓ. To restore a convolution structure, the

sample points are reorganized as a union of n' lattices as in (3.18). After a reindexing, the discrete convolution argument can be applied to each lattice independently, and T_2 is obtained by summing the result,

(3.34)
$$T_2(x) = \frac{r\pi^2}{mn} \sum_{k=0}^{n'-1} \sum_{s=0}^{m-1} \sum_{\ell=-n/n'}^{n/n'-1} |b'_{s,n'\ell+k} - x|^{-2} v_{2\Omega r}(\sin(\nu(b'_{s,n'\ell+k}, x) - \alpha'_{n'\ell+k})) \times \cos\alpha'_{n'\ell+k} D'f(\beta_s, \alpha_{n'\ell+k}).$$

From (3.15), $b'_{s,n'\ell+k} = -b_{s+m'(\ell+(k+1/4)/n')}$ and

(3.35)
$$\begin{aligned} \cos\nu(b'_{s,n'\ell+k}, x) &= \frac{(b'_{s,n'\ell+k} - x)b'_{s,n'\ell+k}}{|b'_{s,n'\ell+k} - x||b'_{s,n'\ell+k}|} \\ &= \frac{(b_{s+m'(\ell+(k+1/4)/n')} + x)b_{s+m'(\ell+(k+1/4)/n')}}{|b_{s+m'(\ell+(k+1/4)/n')} + x||b_{s+m'(\ell+(k+1/4)/n')}|} \\ &= \cos\nu(b_{s+m'(\ell+(k+1/4)/n')}, -x). \end{aligned}$$

Also, the sign of $\nu(b'_{s,n'\ell+k}, x)$ is the same as that of $\nu(b_{s+m'(\ell+(k+1/4)/n')}, -x)$. Hence,

(3.36)
$$\nu(b'_{s,n'\ell+k}, x) = \nu(b_{s+m'(\ell+(k+1/4)/n')}, -x).$$

Of course, it is understood that the arithmetic in the subscripts of β_s and b_s is in $\mathbb{R}/m\mathbb{Z}$,

(3.37)
$$T_2(x) = \frac{r\pi^2}{mn} \sum_{k=0}^{n'-1} \sum_{s=0}^{m-1} \sum_{\ell=-n/n'}^{n/n'-1} |b_{s+m'(\ell+(k+1/4)/n')} + x|^{-2} \times v_{2\Omega r}(\sin(\nu(b_{s+m'(\ell+(k+1/4)/n')}, -x) - \alpha'_{n'\ell+k})) \cos\alpha'_{n'\ell+k} D'f(\beta_s, \alpha_{n'\ell+k}).$$

Reindex the sums over s and ℓ as sums over $t = s + m'\ell$ and ℓ,

(3.38)
$$\begin{aligned} T_2(x) = \frac{r\pi^2}{mn} \sum_{k=0}^{n'-1} \Bigg(& \sum_{t=0}^{m-1} |b_{t+(k+1/4)m'/n'} + x|^{-2} \\ & \times \sum_{\ell=-n/n'}^{n/n'-1} v_{2\Omega r}(\sin(\nu(b_{t+(k+1/4)m'/n'}, -x) - \alpha'_{n'\ell+k})) \\ & \times \cos\alpha'_{n'\ell+k} D'f(\beta_{t-m'\ell}, \alpha_{n'\ell+k}) \Bigg). \end{aligned}$$

For each k, the innermost sum of (3.38) is a discrete convolution with separation $n'\Delta\alpha$. To achieve $O(\Omega^3)$ computation, the convolution at $\nu(b_{t+(k+1/4)m'/n'}, -x)$ is obtained by interpolation from the convolution evaluated at a discrete set of values. As this discrete convolution is $2n'$ times undersampled with respect to the corresponding continuous integral in α, Theorem 2.6 cannot be applied. To avoid the introduction of artifacts, it is again necessary to compute the convolution on an extra fine grid. Introduce a multiplier M^R, and compute the convolution on a grid with a separation of $n'\Delta\alpha/M^R$ between grid points. The algorithm for reconstruction can now be presented.

ALGORITHM 3.2 (divergent beam reconstruction at hardware resolution).

1. *Choose an integer multiplier* M^O. *Typically,* $M^O = 8$ *is large enough. For* $s = 0, \ldots, m-1$ *and* $\ell = -nM^O, \ldots, (n-1)M^O$, *compute the discrete convolutions*

$$h^O_{s,\ell} = \sum_{\mu=-n}^{n-1} v_{2\Omega r}(\sin(\alpha_{\ell/M^O} - \alpha_\mu)) \cos\alpha_\mu D'f(\beta_s, \alpha_\mu). \tag{3.39}$$

2. *Choose an integer multiplier* M^R. *Typically,* $M^R = 8n'$ *should be large enough. For* $k = 0, \ldots, n'-1$ *and* $t = 0, \ldots, m-1$, *compute the discrete convolutions at* $\ell = -nM^R, \ldots, (n-1)M^R$

$$h^R_{k,t,\ell} = \sum_{\mu=-n/n'}^{n/n'-1} \Big(v_{2\Omega r}(\sin(\alpha_{\ell/M^R} - \alpha'_{n'\mu+k})) \cos\alpha_\mu D'f(\beta_{t+m'\mu}, \alpha_{n'\mu+k}) \Big). \tag{3.40}$$

3. *For each* x *at which the reconstruction is needed, evaluate the interpolated discrete backprojection*

$$f_O(x) = \frac{r\pi^2}{mn} \sum_{s=0}^{m-1} |b_s - x|^{-2} \Big((1-\vartheta_s) h^O_{s,\ell_s} + \vartheta_s h^O_{s,\ell_s+1} \Big), \tag{3.41}$$

where

$$\tau_s = \frac{M^O \nu_s(x)}{\Delta\alpha}, \quad \ell_s = \lfloor \tau_s - 1/4 \rfloor, \quad \vartheta_s = \tau_s - (\ell_s + 1/4). \tag{3.42}$$

4. *For each* x *at which the reconstruction is needed, and for each* $k = 0, \ldots, n'-1$, *evaluate the interpolated discrete backprojection*

$$f_{R,k}(x) = \frac{r\pi^2}{mn} \sum_{t=0}^{m-1} |b_{t+(k+1/4)m'/n'} + x|^{-2} \Big((1-\vartheta_t) h^R_{k,t,\ell_t} + \vartheta_t h^R_{k,t,\ell_t+1} \Big), \tag{3.43}$$

where

$$\tau_t = \frac{M^R \nu(b_{t+(k+1/4)m'/n'}, -x)}{\Delta\alpha}, \quad \ell_t = \lfloor \tau_t - 1/4 \rfloor, \quad \vartheta_t = \tau_t - (\ell_t + 1/4). \tag{3.44}$$

5. *The approximation* $f_{2\Omega}(x)$ *to* $V_{2\Omega} * f(x)$ *is given by*

$$f_{2\Omega}(x) = f_O(x) + \sum_{k=0}^{n'-1} f_{R,k}(x). \tag{3.45}$$

The errors in this approximation can arise from a number of sources.

1. The aliasing from the side lobes in the measurement system response function was ignored. One way to reduce this effect is to custom design a detector to minimize the size of the side lobes.

2. The divergent beam transform of a function is not bandlimited, but essentially bandlimited. Reconstructing at a slightly reduced bandwidth from the optimal bandwidth will reduce the aliasing due to the Fourier transform not truly vanishing outside the nominal support region K. Estimates for this type of error are available. See [6, 10], for example.

3. In the derivation of (3.31), the correct kernel $V_{2\Omega|b-x|}$ was replaced by the kernel $V_{2\Omega r}$. This approximation will break down when x gets close to r. Thus, it is necessary to keep $\rho \ll r$.

4. While $Df(\beta_s, \alpha_\ell) = Df(\beta'_{s,\ell}, \alpha'_\ell)$, it is not strictly true that $D'f(\beta_s, \alpha_\ell) = D'f(\beta'_{s,\ell}, \alpha'_\ell)$ as the two are averaged over slightly different sets of lines. Incorporating a source width into the model mitigates the effect of this approximation.

5. If either M^O or M^R is not large enough, the linear interpolation in step 3 or step 4 will effectively further bandlimit the reconstruction, causing artifacts in the reconstructed image.

3.2.3. Shift-convexity of K. In order to apply Algorithm 3.2, the bandregion K must be shift-convex with respect to $W^\perp$ and $P^\perp$. The shape of K is determined by $\theta = \rho/r$, and the scaling of K is determined by the product $r\Omega$. From these parameters, it is not immediately obvious whether K is shift-convex with respect to the sampling lattices. In this section, it will be shown how to choose the parameters m and n to guarantee shift-convexity of K, so Algorithm 3.2 can be applied to recover f at the maximum resolution consistent with the detector hardware.

First, to illustrate the problem which can arise, consider equality in (3.12) and (3.13) with the two examples $\theta = 1/2$ and $\theta = 1/3$. Shift-convexity with respect to $W^\perp$ and $P^\perp$ requires that once K has been shifted by an element of $L_W^\perp$, further shifting away from the origin by elements of $L_P^\perp$ does not reintroduce overlap. The smallest generators for $L_W^\perp$ consistent with (3.12) and (3.13) are $w_1 = (0, 4r\Omega)^T$ and $w_2 = (\frac{\rho}{r+\rho} 4r\Omega, 0)^T$. w_2 has been chosen to be the minimal horizontal shift necessary to ensure $K \cap (K + w_2) = \emptyset$. With equality in (3.12) and (3.13), $m/n = \frac{4\rho}{r+\rho} = \frac{4\theta}{1+\theta}$. For $\theta = 1/2$, $m/n = 8/3$, so $n' = 3$. Hence, $p_1 = w_1/6 = (0, 2r\Omega/3)^T$. As illustrated in Figure 6, further shifting of $K + w_2$ by several multiples of p_1 reintroduces overlap, so K is not shift-convex. On the other hand, for $\theta = 1/3$, $m/n = 1$, so $n' = 1$. Here, $p_1 = w_1/2 = (0, 2r\Omega)$. In this case, $K + w_2 + P_1$ interlaces with K, so no overlap is introduced and K is shift-convex. As θ controls the shape of K, whether this interlacing occurs depends on θ, and also on the size of the vertical shift p_1, which depends on n'.

From the above examples, it can be seen that absent an interlacing of the type seen for $\theta = 1/3$, vertical shifts of $K + w_2$ by multiples of p_1 can reintroduce overlap when the left side of a shifted K covers the upper right side of K. To avoid such overlapping, two approaches are suggested. Each requires an oversampling in β, the amount depending on θ. Oversampling is measured with respect to the optimal sampling density, given by equality in (3.13).

Method 1. Extend w_2 to be the smallest length such that all vertical translations of $K + w_2$ do not overlap K. That is, the inequality of (3.13) is replaced by

$$m = \frac{2\pi}{\Delta\beta} \geq 4\rho\Omega. \tag{3.46}$$

By construction, K is automatically shift-convex, independently of any choice of m and n consistent with (3.12) and (3.46). The shift-convexity comes at the expense of an oversampling in β by a factor of $4\rho\Omega / \frac{4r\rho\Omega}{r+\rho} = 1 + \theta$ with respect to the optimal sampling density.

Since the possibility that K may interlace with its shifts has not been utilized, for some parameter values (such as $\theta = 1/3$), shift-convexity can be achieved with a sparser sampling than this method constructs.

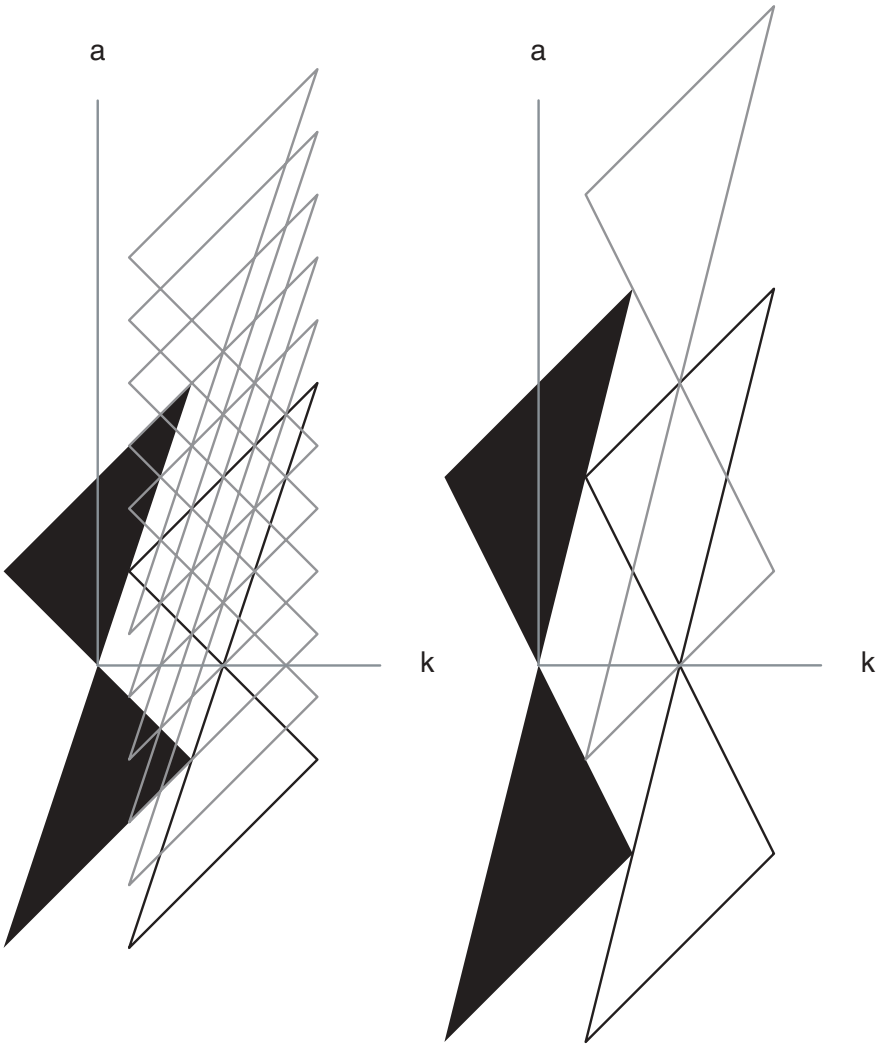


FIG. 6. *The essential bandregion K for the divergent beam transform is shown along with some of its translates. K is solid, $K + w_2$ is outlined in black, and $K + w_2 + jp_1$ with $j \neq 0$ is shown in gray. The left figure was drawn with equality in* (3.12) *and* (3.13) *and with $\theta = 1/2$. Here $n' = 3$. Although $K \cap (K + w_2) = \emptyset$, there is overlap between K and $K + w_2 + jp_1$ for $j = 1, 2, 3$. Hence K is not shift-convex with respect to the corresponding $W^{\perp}$ and $P^{\perp}$. The right figure was drawn with equality in* (3.12) *and* (3.13) *and with $\theta = 1/3$. In this case $n' = 1$, and both $K \cap (K + w_2) = \emptyset$ and $K \cap (K + w_2 + p_1) = \emptyset$. K is shift-convex with respect to the corresponding $W^{\perp}$ and $P^{\perp}$.*

Method 2. The physical detector geometry determines r and $\Delta\alpha$. Accordingly, n is fixed. Since maximizing reconstruction resolution is the objective, Ω is determined by equality in (3.12). $m \in \mathbb{Z}$ can be chosen freely as long as the inequality of (3.13) is satisfied and K is shift-convex.

Algorithm 3.2 is simplest, easiest to implement, and quickest when $n' = 1$. In what follows, it is shown how to choose m so that $n' = 1$ simultaneously with K being shift-convex.

When $n' = 1$, $L_P^{\perp}$ undersamples by a factor of 2, so $p_1 = (0, 2r\Omega)^T$. It is necessary to select $v > 0$ to locate the minimal shift $w_2 = (2vr\Omega, 0)^T$ such that both $K \cap (K +$

$w_2) = \emptyset$ and $K \cap (K + w_2 + p_1) = \emptyset$. The latter holds when either the upper left corner of $K + w_2 + p_1$ is to the right of the line $a = (1 + 1/\theta)k$, or the upper right corner of K is to the left of the line connecting the center of $K + w_2 + p_1$ to the upper left corner of $K + w_2 + p_1$. The coordinates of the upper left and right corners of K are, respectively, $(-2\rho\Omega, 2(r-\rho)\Omega)^T$ and $(2\rho\Omega, 2(r+\rho)\Omega)^T$, and the coordinates of the center and upper left corners of $K + w_2 + p_1$ are, respectively, $(2\upsilon r\Omega, 2r\Omega)^T$ and $(2(\upsilon r - \rho)\Omega, 2(2r-\rho)\Omega)^T$. The first condition becomes (assuming $\upsilon > 0$), $\frac{2(2r-\rho)\Omega}{2(\upsilon r-\rho)\Omega} \leq (1 + 1/\theta)$, or equivalently, since $K \cap (K + w_2) = \emptyset$ implies $\upsilon r > \rho$,

$$\upsilon \geq \frac{3\theta}{1+\theta}. \tag{3.47}$$

The second condition is satisfied when $(2\rho\Omega, 2(r+\rho)\Omega)^T$ is to the left of the line of slope $1 - 1/\theta$ through $(2\upsilon r\Omega, 2r\Omega)^T$. That is, $\frac{2\rho\Omega}{2(\rho-\upsilon r)\Omega} \geq \frac{\theta-1}{\theta}$, or equivalently,

$$\upsilon \geq \frac{\theta}{1-\theta}. \tag{3.48}$$

Thus, for K to be shift-convex with $n' = 1$, (3.13), (3.47), and (3.48) together require m/n to be integral and

$$m \geq 2r\Omega \max\left(\frac{2\theta}{1+\theta}, \min\left(\frac{3\theta}{1+\theta}, \frac{\theta}{1-\theta}\right)\right). \tag{3.49}$$

Inequality (3.49) can be expressed as

$$m \geq n \begin{cases} \frac{4\theta}{1+\theta} & \theta \in (0, 1/3], \\ \frac{2\theta}{1-\theta} & \theta \in (1/3, 1/2], \\ \frac{6\theta}{1+\theta} & \theta \in (1/2, 1). \end{cases} \tag{3.50}$$

For $\theta \leq 1/3$, $\frac{4\theta}{1+\theta} \leq 1$, so $m = n$ satisfies both requirements, for an effective oversampling in β of $\frac{1+\theta}{4\theta}$. For $\theta \in (1/3, 1/2]$, $1 < \frac{2\theta}{1-\theta} \leq 2$, so $m = 2n = 2r\Omega$. This gives an effective oversampling of $\frac{1+\theta}{2\theta}$. For $\theta \in (1/2, 1)$, $2 < \frac{6\theta}{1+\theta} < 3$, so $m = 3n = 3r\Omega$. This gives an effective oversampling of $\frac{3(1+\theta)}{4\theta}$.

Inequality (3.50) governs the additional sampling density beyond that of (3.13) which is needed to avoid overlap. A component of the inefficiency (which, for small θ, is unacceptably large) arises from the requirement that m/n be integral. Which of the two methods is preferable can be ascertained by examining Figure 7 in which the oversampling required for each method is plotted as a function of θ. The maximum oversampling of 1.75 occurs for $\theta = 3/4$. We note that for modern medical CT scanners, typically $\theta \leq 1/3$, and $\theta \leq 1/2$ almost always.

No oversampling occurs when $\theta = 1/3$. As θ increases beyond $1/3$, the oversampling jumps abruptly to $4/3$ because (3.46) requires $m = 2n$ when the right-hand side is slightly larger than 1. In a configuration with θ slightly larger than $1/3$, practically speaking, the relatively large ρ precludes visualization of the full 2Ω bandwidth. Therefore, it is common practice to filter the acquired data to perform a reduced resolution reconstruction. With Ω slightly reduced from the optimal value given by (3.12), $m = n$ can be used. This gain in reconstruction efficiency comes at the cost of a slight loss in reconstruction resolution.

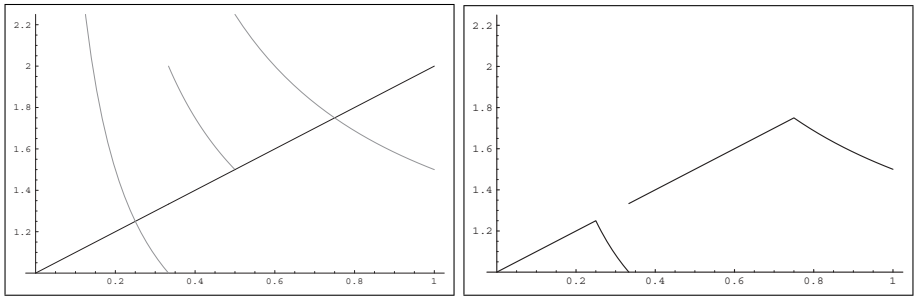


FIG. 7. *The left figure shows the oversampling required in the β direction to ensure the shift-convexity of K is plotted against θ. The oversampling required for Methods* 1 *and* 2 *are shown in black and gray, respectively. The right figure shows the oversampling required when the most efficient method is selected.*

Reduction of the reconstruction bandwidth Ω to reduce oversampling can be interpreted as yet a third method available to guarantee shift-convexity. The specific application requirements will determine what levels of oversampling versus resolution reduction are acceptable.

We close this section by remarking that the increased source density from shift-convexity induced oversampling impacts reconstruction time. The data from each source point must be backprojected. Since the bulk of the processing time is in the backprojection phase, computation time is increased by a factor matching that of the oversampling. Future studies will assess the practical significance of this increased execution time.

4. Numerical experiments. To validate the performance of the divergent beam multichannel sampling algorithm, Algorithm 3.2, third generation projection data from a simulated phantom was reconstructed both with Algorithm 3.2 and with the standard algorithm (Algorithm 5.3 in [13]).

The phantom used was an off-center disk modeling water with two pins of bone. This phantom, shown in Figure 8, was selected to test the recovery of details near the edge of the scan circle, a region in which any lack of conformance to (3.6) will produce artifacts. Also, in this region the approximation of (3.31) incurs the largest error. The scan radius for the phantom is $\rho = 20$. The circle which modeled water had a radius of 6.27208, was centered at $(-9, 9)$, and had attenuation coefficient in CT numbers of 1000. The pins modeling bone had a radius of 0.2, attenuation coefficient 2000, and were centered at $(14.3313, 9)$ and $(-9, 14.3313)$.

The simulated data from a third generation detector were generated with a source radius of $r = 60$. Thus, $\rho/r = 1/3$, which, with equality in (3.12) and (3.13), implies $n' = 1$. Figure 5(a) shows the interlacing of the original and reflected lattices for this geometry. The reconstruction bandwidth for the standard algorithm is $\Omega = 31.59$ corresponding to a resolution of 0.1 in the reconstructed image. The simulated hardware-limited bandwidth was 2Ω for an effective reconstruction resolution of 0.05.

Four reconstructions were performed, all using the cosine filter with a minor correction to reduce ringing. That is, ψ in (2.13) is taken to be $\hat{\psi}(\sigma) = \chi_{[-1,1]}(\sigma) \cos \sigma\pi/2$.

In Figure 9, the reconstruction at bandwidth 2Ω using the divergent beam multichannel sampling algorithm with $M_O = M_R = 8$ is shown. The data set consisted of 410 samples in α and 1884 samples in β. The top two images show the intermediate reconstructions f_O and f_R obtained by using only the original or reflected sample

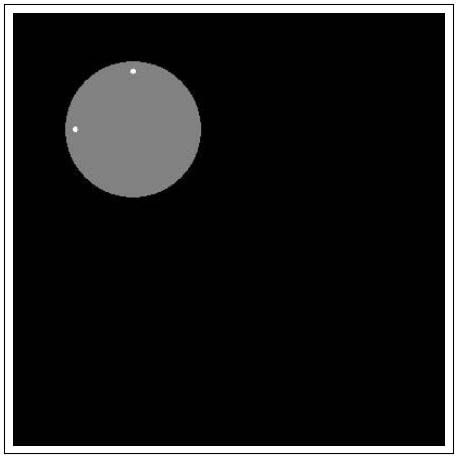


FIG. 8. *The phantom used for numerical experiments is shown. The scan radius for the phantom is* $\rho = 20$. *Data were collected using a source radius of* $r = 60$. *The larger circle modeling water has radius* 6.27208, *is centered at* $(-9, 9)$ *and has attenuation coefficient* 1000. *The pins modeling bone have radius* 0.2, *attenuation* 2000, *and are centered at* $(14.3313, 9)$ *and* $(-9, 14.3313)$. *This phantom was chosen to test the reconstruction algorithms' ability to recover details near the edge of the scan circle, a region in which any lack of conformance to* (3.6) *will produce artifacts. Also, in this region the approximation of* (3.31) *incurs the largest error.*

points. Observe that each of the intermediate images exhibits undersampling artifacts which disappear in the final reconstructed image. The reconstructions in this figure and the following one are all displayed with a window of 5 units centered on 1000.

A comparison between the divergent beam multichannel reconstruction and the standard algorithm appears in Figure 10. In Figure 10(a), the 2Ω reconstruction from the divergent beam multichannel sampling algorithm with 410 α samples and 1884 β samples is copied from Figure 9.

Figure 10(b) shows the standard divergent beam reconstruction at bandwidth Ω from 942 β samples and 410 α samples. It was sampled with the minimum density sufficient for applying the standard divergent beam reconstruction at a bandwidth of Ω. It shows a swirl artifact which is probably from a slight undersampling in β. The reconstructed diameter of the two pins is significantly larger than that of the phantom or that of the multichannel sampling reconstruction. This is due to the lower bandwidth, and hence, inferior resolution.

Figure 10(c) shows the reconstruction obtained when the standard divergent beam algorithm at bandwidth Ω is applied to the same data set used in generating Figure 10(a). This reconstruction is two-times oversampled in β and at the minimum sample density in α. The oversampling in β clears the swirl artifact, but as this reconstruction is still at bandwidth Ω, the smearing artifact does not improve.

Figure 10(d) shows the standard divergent beam algorithm applied at bandwidth 2Ω to data with 818 α and 1884 β samples. In this simulation the 818 α samples were "acquired" with detector elements half the size of those in Figures 10(a), 10(c), and 10(d). This data set is at the minimal density for a 2Ω application of the standard algorithm. As the bandwidth is 2Ω, the pins do not show the smearing artifact.

Comparing Figures 10(a) and 10(d), it can be seen that the multichannel algorithm does effectively reconstruct at a 2Ω bandwidth.

Finally, it has been observed that the residual artifact between the pins is not

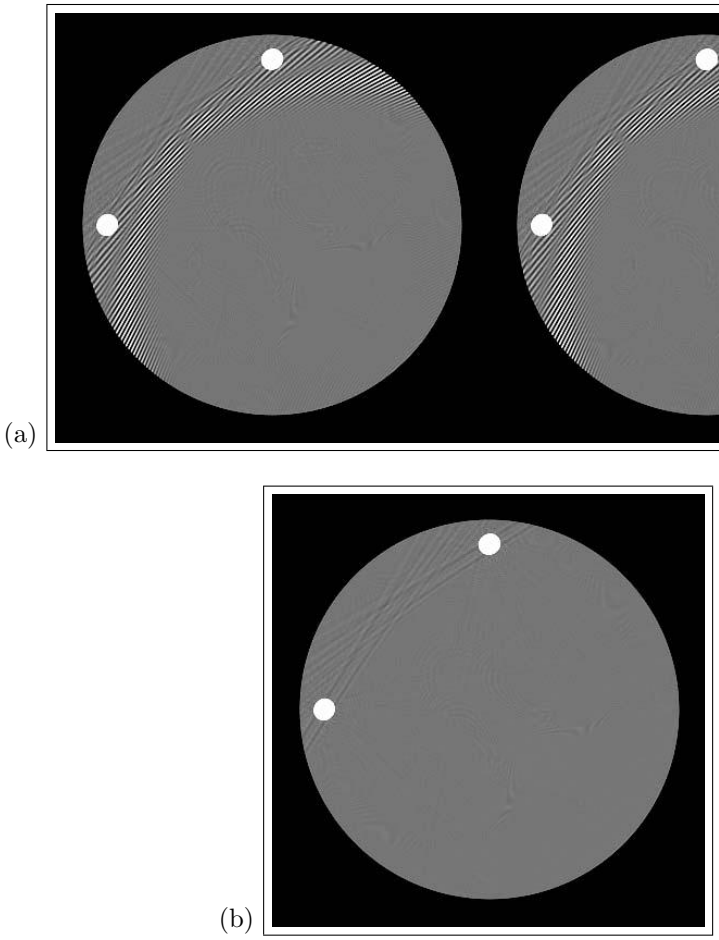


FIG. 9. *The reconstruction of the phantom from simulated third generation divergent beam projection data is shown with a window of* 5 *CT units centered about* 1000 *(water). The divergent beam multichannel sampling algorithm was applied with* $M_O = M_R = 8$ *to reconstruct at bandwidth* 2Ω *from* 1884 β *samples and* 410 α *samples. The upper left image shows the reconstruction* f_O *using data from the original lattice* L_O*. The upper right image shows the reconstruction* f_R *using data from only the reflected lattice. Both* f_O *and* f_R *exhibit artifacts from the undersampling in* α*. The bottom image is the final reconstruction obtained from the sum of* f_O *and* f_R*. The undersampling artifacts disappear as this reconstruction is appropriately sampled.*

a function of the reconstruction algorithm. It is an artifact of the fine sampling incorporated into the third generation projection simulator. Increasing the fine sampling reduces this artifact. Due to the smaller detector element, the data set for Figure 10(d) has an effective fine sampling double that of the other data sets. As a result, the amplitude of the artifact is reduced.

5. Conclusions. A multichannel sampling reconstruction algorithm has been presented to reconstruct data from a third generation CT scanner at the resolution limit of hardware detector. Initial reconstructions have validated the algorithm. Further testing is underway to evaluate the speed and image quality of this algorithm. It is expected that similar techniques can be applied to handle the flying focal spot capability of some modern scanners.

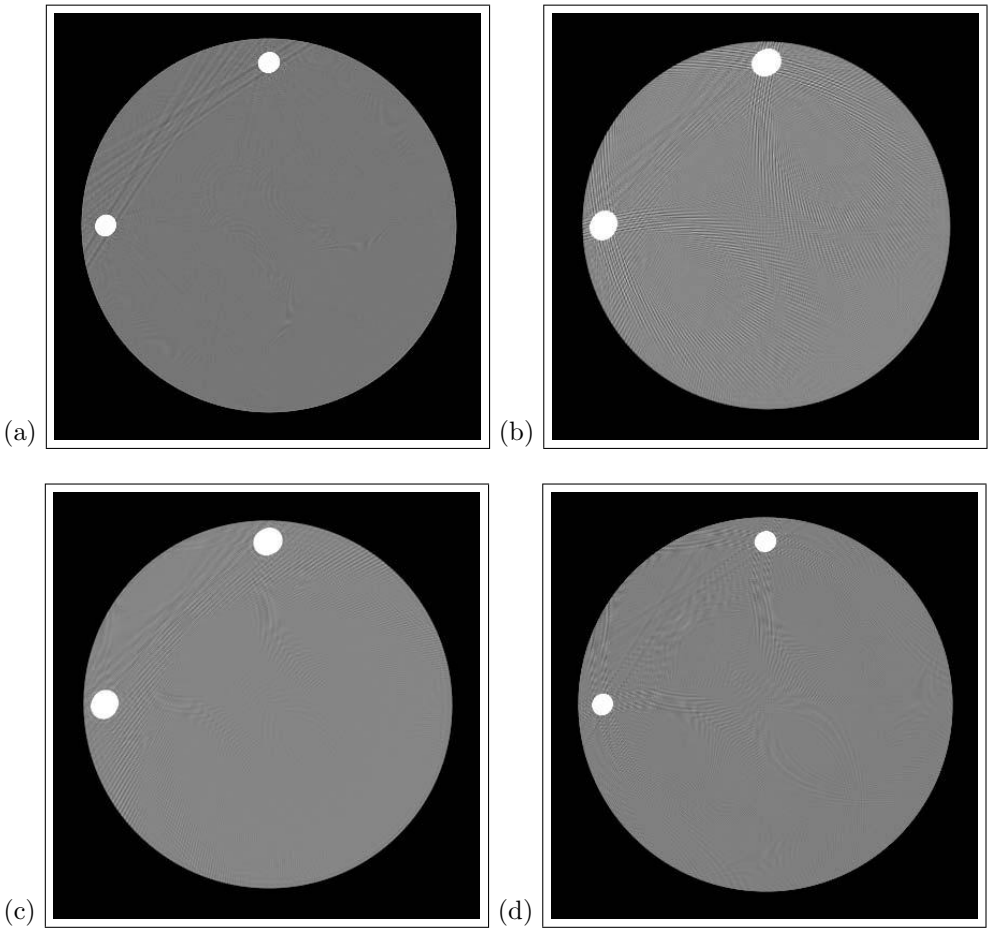


FIG. 10. *Reconstructions of the phantom from simulated third generation divergent beam projection data are shown with a window of* 5 *CT units centered about* 1000 *(water). Each of the reconstructions is centered on the region of interest in the phantom. In* (a), *the divergent beam multichannel sampling algorithm was applied to reconstruct at bandwidth* 2Ω *from* 1884 β *samples and* 410 α *samples. In* (b), *the standard divergent beam reconstruction was applied at bandwidth* Ω *from* 942 β *samples and* 410 α *samples. It is sampled with the minimum density sufficient for applying the standard divergent beam reconstruction at a bandwidth of* Ω. *In* (c), *the standard divergent beam algorithm was applied to reconstruct at a bandwidth* Ω *from the same data set used for* (a). *This reconstruction is two-times oversampled in* β, *but at the minimum sample density in* α. *In* (d) *the standard divergent beam algorithm was applied at bandwidth* 2Ω *to* 1884 β *samples and* 818 α *samples. The detector element was half the size of the detector element for the other three data sets.*

REFERENCES

[1] J. L. BROWN, JR. AND K. SA'NGSARI, *Sampling reconstruction of N-dimensional images after multi-linear filtering*, IEEE Trans. Circuits Syst., 36 (1989), pp. 1035–1038.

[2] A. M. CORMACK, *Sampling the radon transform with beams of finite width*, Phys. Med. Biol., 23 (1978), pp. 1141–1148.

[3] K. F. CHEUNG, *A multidimensional extension of Papoulis' generalized sampling expansion with the application in minimum density sampling*, in Advanced Topics in Shannon Sampling and Interpolation Theory, Springer-Verlag, New York, NY, 1993, pp. 85–119.

[4] L. DESBAT, *Efficient sampling on coarse grids in tomography*, Inverse Problems, 9 (1993), pp. 251–269.

[5] A. Faridani, *Reconstructing from efficiently sampled data in parallel-beam computed tomography*, in Inverse Problems and Imaging, G. F. Roach, ed., Pitman Res. Notes Math. Ser. 245, Longman Press, London, 1991, pp. 68–102.
[6] A. Faridani, *An application of a multidimensional sampling theorem to computed tomography*, Contemp. Math. 113, AMS, Providence, RI, 1990, pp. 65–80.
[7] A. Faridani, *A generalized sampling theorem for locally compact abelian groups*, Math. Comp., 63 (1994), pp. 307–327.
[8] A. Faridani, *Sampling in parallel-beam tomography*, in Inverse Problems, Tomography, and Image Processing, A. G. Ramm, ed., Plenum, New York, NY, 1998, pp. 33–53.
[9] S. H. Izen, *Generalized sampling expansion on lattices*, IEEE Trans. Signal Process., to appear.
[10] F. Natterer, *The Mathematics of Computerized Tomography*, Wiley, New York, 1986.
[11] F. Natterer, *Sampling in fan beam tomography*, SIAM J. Appl. Math, 53 (1993), pp. 358–380.
[12] F. Natterer, *Sampling Functions with Symmetries*, preprint, 1998, http://wwwmath.uni-muenster.de/math/inst/num/Preprints/1999/natterer_1/paper.pdf.
[13] F. Natterer and F. Wübbeling, *Mathematical Methods in Image Reconstruction*, SIAM Monogr. Math. Model. Comput. 5, SIAM, Philadelphia, 2001.
[14] F. Noo, C. Bernard, F. X. Litt, and P.Marchot, *A comparison between filtered backprojection algorithm and direct algebraic method in fan beam CT*, Signal Proc., 51 (1996), pp. 191–199.
[15] V. Palamodov, *Localization of harmonic decomposition of the radon transform*, Inverse Problems, 11 (1995), pp. 1025–1030.
[16] A. Papoulis, *Generalized sampling expansion*, IEEE Trans. Circuits Syst., 24 (1979), pp. 652–654.
[17] D. P. Petersen and D. Middleton, *Sampling and reconstruction of wave-number-limited functions in N-dimensional Euclidean spaces*, Inform. Contr., 5 (1962), pp. 279–323.

SIAM J. APPL. MATH.
Vol. 65, No. 3, pp. 1053–1079

PERIODICALLY GENERATED PROPAGATING PULSES*

L. L. BONILLA†, M. KINDELAN†, AND J. B. KELLER‡

Abstract. Certain equations with integral constraints have as solutions time-periodic pulses of a fieldlike unknown while a currentlike unknown oscillates periodically with time. A general asymptotic theory of this phenomenon, the generalized Gunn effect, has been found recently. Here we extend this theory to the case of nonlinearities having only one stable zero, which is the case for the usual Gunn effect in n-GaAs. Our ideas are presented in the context of a simple scalar model where the waves can be constructed analytically and explicit expressions for asymptotic approximations can be found.

Key words. reaction-diffusion-convection equations, propagation of pulses and wavefronts, piecewise linear model, Gunn effect

AMS subject classifications. 34E15, 92C30

DOI. 10.1137/S0036139903434948

1. Introduction.

The Gunn effect is the periodic oscillation of the current in a passive external circuit attached to a dc-voltage biased semiconductor whose electron drift velocity has a single maximum as a function of the electric field (and therefore the curve of electron velocity versus electric field has negative slope for field values on a certain interval, a fact called *negative differential mobility*) [22, 25]. During each period of the oscillation, a pulse of the electric field is created at the injecting contact, moves through the semiconductor, and is annihilated at the receiving contact. While originally observed in bulk n-GaAs samples, similar current oscillations, mediated by pulse dynamics in dc-voltage biased semiconductors, have been found in many materials, several of which lack negative differential mobility [1]. Instead, other processes (impact ionization at impurities [24], nonlinear capture coefficients [21], nonlinear recombination processes, etc.) may be responsible for a current vs. local electric field characteristic curve displaying a local maximum followed by a region of negative slope (negative differential conductivity).

Propagation of pulses occurs in many systems of interest in biology, physics, and so on: morphogen pulses or spikes in activator-inhibitor reaction-diffusion systems modeling cell development or chemical reactors [16, 12, 13, 23], propagation of nerve impulses along myelinated or unmyelinated fibers [18, 20, 17], pulse propagation through cardiac cells [18], calcium release waves in living cells [9], semiconductor superlattices [8, 26], and oscillatory instabilities of the current in bulk semiconductors with an N-shaped current-field characteristic [1, 22, 25]. These distributed systems can be spatially discrete or continuous and can be described by a variety of model equations. Sometimes a pulse is created from an appropriate initial condition and reaches a stable shape, moving uniformly until it arrives at a boundary. Sometimes understanding pulse dynamics is the key to describing a more complicated evolution of the system. A good example is the Gunn effect in bulk semiconductors.

*Received by the editors September 24, 2003; accepted for publication (in revised form) April 19, 2004; published electronically March 31, 2005. This research was supported by Spanish MCyT grant BFM2002-04127-C02-01 and by European Union grant HPRN-CT-2002-00282.
http://www.siam.org/journals/siap/65-3/43494.html
†Escuela Politécnica Superior, Universidad Carlos III de Madrid, Avenida de la Universidad 30, 28911 Leganés, Spain (bonilla@ing.uc3m.es, kinde@ing.uc3m.es).
‡Departments of Mathematics and Mechanical Engineering, Stanford University, Stanford, CA 94305 (keller@math.stanford.edu).

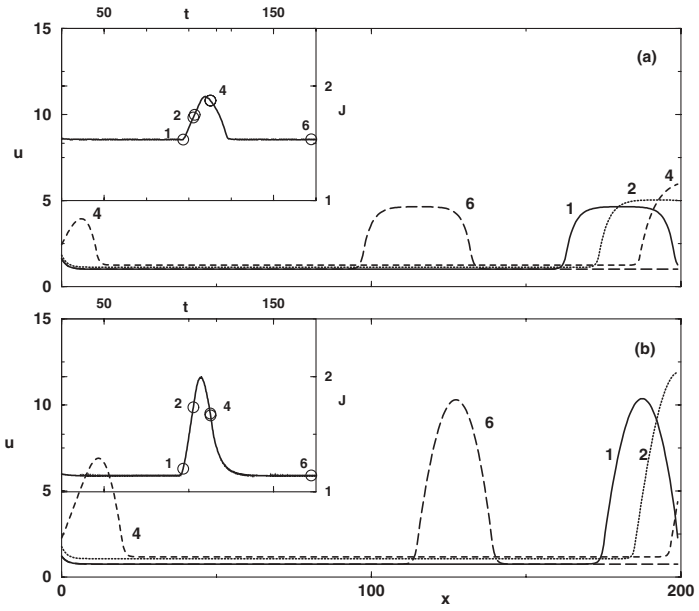


FIG. 1. *Self-sustained oscillations of the "current" $J(t)$ (see inset, which shows one period of $J(t)$) mediated by pulses of the "field" $u(x,t)$, which are solutions of the model described in section* 2. *The pulse profiles correspond to the times marked in the inset. Case* (a): *bistable source. Case* (b): *source with a single stable zero.*

While Gunn-like instabilities have been known for a long time, only recently have pulse annihilation and creation at boundaries been studied by asymptotic methods [14, 5, 6]. These theories treated the case in which the relevant nonlinear source term has two stable zeros. Then there are stable wavefronts joining these two zeros, and a moving pulse is a flat region of high field bounded by two wavefronts. The pulse changes its size if its leading and trailing wavefronts move at different speeds. Figure 1 shows the time-periodic oscillation of a "current" $J(t)$ accompanied by the repeated generation and motion of flat-top pulses of an "electric field" $u(x,t)$, which are solutions of model equations described in section 2. Figure 1(a) corresponds to the case of a bistable nonlinear source. The asymptotic analysis of the Gunn effect is based on the dynamics of such pulses [5]. However, the source term (velocity-field characteristic curve) in very relevant materials, such as bulk n-GaAs [22], wide-miniband GaAs/AlAs superlattices [11, 15], and semi-insulating GaAs [21], does not have two stable zeros. Instead, the nonlinearity has a single stable zero, so the previous theories, based on two moving wavefronts, are not valid. Figure 1(b) shows the self-sustained oscillations corresponding to this case. Can we find an asymptotic theory of pulse mediated oscillations in this case? The answer is yes, as we show in this paper.

To emphasize that our analysis applies to a class of models, we shall analyze a simpler problem than that of the Gunn effect. We will study a nonlinear reaction-diffusion-convection equation with an integral constraint [6] and a piecewise linear source term. Then the pulses can be constructed analytically, their size can change as they move, and explicit expressions for the asymptotic approximations can be found. Such a construction was used by Rinzel and Keller for the FitzHugh–Nagumo equation [20]. For Kroemer's model, a piecewise linear electron velocity was used to calculate the exact form and speed of a steadily propagating pulse [10].

The outline of the paper is as follows. Section 2 presents our simplified model. It also reviews the key ideas of previous asymptotic theories, valid for an N-shaped nonlinearity with two stable zeros. Section 3 discusses the construction of stationary solutions, and the kinematics of wave fronts, in the limit of long samples. When the nonlinearity is piecewise linear, the pulses can be found analytically. Section 4 discusses the dynamics of a single pulse moving from the injecting to the receiving boundary. We show that the pulse changes form and speed adiabatically, following the instantaneous value of the current. In section 5, we complete our description of the oscillations by explaining what happens when the pulse reaches the receiving boundary and how a new pulse is created at the injecting boundary. Section 6 contains our conclusions. The appendices are devoted to technical matters.

2. Simple scalar model. The model consists of a one-dimensional nonlinear parabolic equation for $u(x,t)$ (the "electric field") with an unknown forcing term $J(t)$ (the "current"). There is also an integral constraint (the "voltage bias condition"), as well as boundary and initial conditions,

$$\frac{\partial u}{\partial t} + K\,\frac{\partial u}{\partial x} = \frac{\partial^2 u}{\partial x^2} + J - g(u), \quad 0 < x < L, \tag{2.1}$$

$$\frac{1}{L}\int_0^L u(x,t)\,dx = \phi, \tag{2.2}$$

$$u(0,t) = \rho\,J(t), \qquad \rho > 0, \qquad \frac{\partial u}{\partial x}(L,t) = 0, \tag{2.3}$$

$$u(x,0) = f(x) \geq 0, \qquad 0 < x < L. \tag{2.4}$$

Here K, ρ, and ϕ are positive constants.

This model becomes the Kroemer model of the Gunn effect if the advection constant K in (2.1) is replaced by $g(u)$ [2, 3, 4, 14, 22, 25]. For the Kroemer model, existence and uniqueness of solutions have been studied by Liang [19]. Furthermore, linear stability of the stationary and moving pulse solutions has been analyzed in [3]; see also [22, 25]. A bifurcation analysis of the self-sustained oscillations due to pulse dynamics near critical values of ϕ can be found in [4]. Asymptotic analyses of the Gunn effect for the Kroemer model with a bistable $g(u)$ can be found in [14, 5].

2.1. Bistable source $g(u)$. The simplest case to consider is that of an N-shaped nonlinearity $g(u)$ (the "velocity-field characteristic") for $u \geq 0$, with a local maximum $g_M = g(u_M)$, $u_M > 0$, followed by a local minimum $g_m = g(u_m) > 0$, $u_m > u_M$. Then $J - g(u)$ may have up to three positive zeros for $J > 0$, namely, $u_1(J) < u_2(J) < u_3(J)$. For large enough L and K, $g_M/u_M < \rho < g_m/u_m$, and for ϕ in a certain subinterval of $(u_M, u_3(g_M))$, there are stable time periodic solutions of (2.1)–(2.3) of Gunn type. As shown in Figure 1(a), while $J(t)$ oscillates periodically, pulses of $u(x,t)$ are created at $x = 0$, move towards $x = L$, and disappear there.

The analysis of the model is simple in the asymptotic limit

$$0 < \epsilon \equiv \frac{1}{L} \ll 1. \tag{2.5}$$

In this limit (2.1)–(2.2) may be written as

$$\frac{\partial u}{\partial s} + K\,\frac{\partial u}{\partial y} = \epsilon \frac{\partial^2 u}{\partial y^2} + \frac{J - g(u)}{\epsilon}, \tag{2.6}$$

$$\int_0^1 u(y,s)\,dy = \phi, \tag{2.7}$$

where

$$y = \epsilon x, \quad s = \epsilon t. \tag{2.8}$$

Equation (2.6) is a parabolic equation with fast reaction and slow diffusion terms. As $\epsilon \to 0+$, $u(y,s)$ is typically a piecewise constant function taking on the *order* 1 *values* $u_1(J)$ or $u_3(J)$ in intervals of length $y = O(1)$. The extrema of these intervals are typically moving internal layers. These layers are important because they bound pulses, and the self-sustained oscillation we want to describe is due to recycling and motion of pulses at the boundaries. A pulse is a region where u is $u_3(J)$, separated by moving wavefronts from two other regions where u is $u_1(J)$. There are two wavefronts bounding the pulse. In the backfront, u increases from $u_1(J)$ to $u_3(J)$; this front moves with a speed $c_+(J)$. The forefront moves at speed $c_-(J)$, and in it u decreases from $u_3(J)$ to $u_1(J)$. Forefront and backfront are heteroclinic trajectories connecting the two saddles $(u_1(J), 0)$ and $(u_3(J), 0)$ in an appropriate phase plane $(u, du/d\chi)$, where $\chi = [y - Y(s)]/\epsilon$ is a moving coordinate ($\chi = 0$ at the wavefront and $dY/ds = c_\pm(J)$). The instantaneous value of $J(s)$ is determined by using the integral condition (2.7). Typically J obeys the simple equation

$$\frac{dJ}{ds} = A(J)\,[n_+ c_+(J) - n_- c_-(J)], \tag{2.9}$$

where $A(J) > 0$ is a known function of J, and n_+ and n_- are the numbers of wavefronts with increasing and decreasing u profiles, respectively. For high-field domains, $n_+ - n_- = 0, 1$ [6, 5]. The fixed points of (2.9) correspond to the equal area rule

$$\int_{u_1}^{u_3} [g(u) - c]\,du = 0 \tag{2.10}$$

if $n_+ = n_-$, or to possible plateaus in the shape of $J(s)$ otherwise. Many questions on the stability of the pulses and their evolution can be simply answered by analyzing (2.9) and using the asymptotic procedure explained in [6, 5]. The description of pulse creation and annihilation at the boundaries requires a finer analysis, as explained in [6, 5].

2.2. Saturating source. If $g(u)$ saturates, i.e., $g(u) \to$ constant as $u \to \infty$, no $u_3(J)$ exists and the previous construction is no longer possible. What is the correct asymptotic description of a pulse in this case? Let us anticipate the answer here. As Figure 1(b) shows, a pulse is a traveling wave whose profile has a single maximum, and tends to $u = u_1(J)$ as $x \to \pm\infty$. The pulse is bounded by a leading front (forefront, moving at speed c_-) and a trailing front (backfront, moving at speed c_+). Two parameters uniquely determine the pulse: its maximum height, u_m, and the instantaneous value of J. Given u_m and J, we can find the wavefronts enclosing the pulse by simple phase plane constructions. Consider the phase plane $(u, du/d\chi)$,

$\chi = x - X(t)$, where $X(t)$ is the instantaneous position of a wavefront and dX/dt its speed. There is a unique speed $c_+ = c_+(J, u_m)$ for which a separatrix issuing from the saddle $(u_1(J), 0)$ on the upper half plane reaches the u axis at $(u_m, 0)$. This separatrix constitutes the backfront of the pulse, and a similar construction supplies its forefront moving at speed $c_-(J, u_m)$. In general, $c_+ \neq c_-$, which implies that our pulse changes its size as it moves. How do we characterize the dynamics of pulses?

Suppose that there is a single pulse moving in the sample. We need to determine the instantaneous values of J and u_m, for they characterize the pulse completely . The pulse width, l, changes as $dl/dt = c_- - c_+$. On the other hand, l may be determined by a line integral on the corresponding phase plane trajectories which form the pulse. Then $l = l(J, u_m)$. The dc bias condition yields a connection between u_m and J, $u_m = U(J)$. Then the pulse width is a function of J only, $\varphi(J) = l(J, U(J))$. Therefore, since $dl/ds = \varphi'(J)\, dJ/ds = c_- - c_+$, we get

$$\frac{dJ}{ds} = \frac{c_+(J, U(J)) - c_-(J, U(J))}{-\varphi'(J)}. \tag{2.11}$$

Typically the fixed point of this equation, $J = J^*$, is such that $c_+ = c_-$ is a globally stable solution, so that J tends exponentially fast to J^*. The corresponding pulse moves steadily without changing its size. This pulse is the homoclinic orbit in the phase plane, usually given by an equal-area rule, and has been exhaustively studied by previous authors. Notice that the present construction explains why this steadily moving pulse is stable, thereby clarifying an old issue at the heart of the Gunn effect [25, 2]. When the pulse reaches the receiving contact, a different stage of the Gunn oscillation begins. This stage and others needed to fully describe the Gunn oscillation will be explained later.

A subtle point follows. Due to the integral condition (2.2), the pulse height and width (in the (y, s) scales) are $O(\epsilon^{-\frac{1}{2}}) \gg 1$ and $O(\epsilon^{\frac{1}{2}})$, respectively, while outside the pulse, $u = u_1 = O(1)$ and $J = O(1)$. Thus our leading order approximation for $u(y, s)$ is not uniformly of the same order in space. Successive approximations of a single pulse lead to the following ansatz for u:

$$u(y, s; \epsilon) \sim u^{(0)}(y, s; \epsilon) + \epsilon\, u^{(1)}(y, s; \epsilon).$$

Here each $u^{(j)}(y, s; \epsilon)$ may be of different order in ϵ for different values of the space and time variables. However, we shall impose that

$$\frac{u^{(1)}}{u^{(0)}} = O(1)$$

uniformly in y and s as $\epsilon \to 0+$. This situation results in a changing (self-adjusting) time scale for the evolution of J described by (2.11). See [7] for the description of a similar situation in combustion theory.

3. Boundary layers and wavefronts. The model (2.1)–(2.4) was introduced in order to argue that the asymptotics of the Gunn effect is universal within a class of models [6]. The nonlinearity $g(u)$ was originally N-shaped: it had three branches for $u > 0$ (with a maximum at $u_M > 0$ and a minimum at $u_{\min} > u_M$ such that $g(u_M) > g(u_{\min}) > 0$, with $g(u) \to \infty$ as $u \to \infty$). In the present paper, we shall assume that $g(u)$ is a smooth function having a single maximum at $u_M > 0$, $g(u_M) = g_M > 0$, such that $g'(0) = \beta > 0$, and $\lim_{u\to\infty} g(u) = \alpha \in (0, g_M)$. To

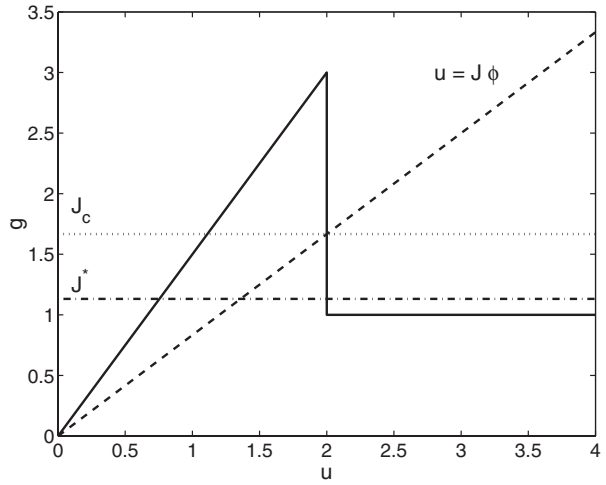


FIG. 2. *The nonlinearity function $g(u)$ (solid) and the curve corresponding to the boundary conditions $J = u/\rho$ (dashed). The parameter values are $\phi = 1.6$, $L = 200$, $K = 2.0$, $u_M = 2.0$, $\beta = 1.5$, $\alpha = 1.0$, $\rho = 1.2$.*

obtain explicit analytic expressions, we shall use a piecewise linear version of $g(u)$,

$$g(u) = \beta u\,\theta(u_M - u) + \alpha\,\theta(u - u_M), \tag{3.1}$$

where $\theta(x)$ is the Heaviside unit step function. See Figure 2, where we have also shown a typical straight line $J = u/\rho$ which gives the value of u at the left boundary. Notice that the straight line intersects $g(u)$ at $(u_M, \rho u_M)$, with $\alpha < \rho u_M < g_M$. We want to find stable solutions $(u(x,t), J(t))$ of (2.1) and (2.2) under the boundary conditions (2.3) and the initial condition (2.4) for appropriate positive values of the bias ϕ in the asymptotic limit $L \to \infty$.

3.1. Outer limit and boundary layers. If $0 < \phi < u_M$, there is a single stationary solution of (2.1)–(2.3) which is easily constructed. Let $y = x/L \equiv \epsilon x$, $0 < \epsilon \ll 1$. Introducing this scaling in (2.1) yields the leading order equation

$$J - g(u) = 0 \tag{3.2}$$

outside the boundary layers at $y = 0, 1$. For $0 < J < g_M$ and $0 < y < 1$, (3.2) has the solutions $u_1(J) < u_2(J)$. Of these we may choose $u = u_1(J)$, $0 < y < 1$, as the outer limit. Inserting it into (2.2), we find $u_1(J) + O(\epsilon) = \phi$ and then (3.2) yields

$$J = g(\phi). \tag{3.3}$$

For the piecewise linear $g(u)$, $J = \beta\,\phi$, and $u_1(J) = J/\beta$.

The boundary layers at $y = 0, 1$ are solutions of the problems

$$\frac{\partial^2 U}{\partial \xi^2} \mp K\frac{\partial U}{\partial \xi} + J - g(U) = 0, \quad 0 < \xi < \infty, \tag{3.4}$$

$$U(0) = \rho J, \quad U(\infty) = u_1(J). \tag{3.5}$$

Here $\xi = x = y/\epsilon$ for the injecting boundary layer at $y = 0$, and $\xi = L - x = (1-y)/\epsilon$ for the receiving boundary layer at $y = 1$. The minus (resp., plus) sign in (3.4)

corresponds to the injecting (resp., receiving) boundary. Clearly, the shape of the unique solution of (3.4) and (3.5) (for $\xi = x$) depends on whether J is smaller or larger than $J = J_c = u_M/\rho$. When $0 < J < J_c$, the boundary layer profile monotonically decreases from $u = \rho J$ to $u_1 = J/\beta$. For the piecewise linear $g(u)$, we have

$$U(\xi) = \frac{J}{\beta} + \left(\rho J - \frac{J}{\beta}\right) \exp\left(-\frac{\sqrt{K^2 + 4\beta} \mp K}{2}\,\xi\right), \tag{3.6}$$

$$\int_0^\infty (U - u_1)\, d\xi = \frac{2\left(\rho J - \frac{J}{\beta}\right)}{\sqrt{K^2 + 4\beta} \mp K}. \tag{3.7}$$

However, if $J > J_c$, the boundary layer profile reaches a maximum before decreasing to $u_1 = J/\beta$. Numerical simulations show that in this case, the stationary solution of the model becomes unstable to Gunn-type oscillations. Given (3.3), this occurs for $\phi > u_1(J_c)$.

3.2. Wavefronts.

3.2.1. Bistable source. Let us first review how to compute the wavefronts when $g(u)$ is N-shaped. Then (3.2) has three solutions $u_1(J) < u_2(J) < u_3(J)$ for $g(u_m) < J < g(u_M)$. As explained in [6], the building blocks of the Gunn-oscillation asymptotics are wavefronts connecting $u_1(J)$ and $u_3(J)$. These wavefronts adjust themselves instantaneously to the value of J, as this unknown evolves on a slower time scale (see below). A wavefront centered at $x = X_\pm(t)$ is a monotone function of $\chi = x - X_\pm(t)$ such that

$$u(-\infty; c_+) = u_1(J), \quad u(+\infty; c_+) = u_3(J) \quad \text{and}$$
$$u(-\infty; c_-) = u_3(J), \quad u(+\infty; c_-) = u_1(J).$$

For the simple model used here, there is a relation between the wavefronts $u(\chi; c_+)$ and $u(\chi; c_-)$.

THEOREM 1. *Let $u(\chi; c_\pm)$ be the wavefront satisfying*

$$\frac{d^2u}{d\chi^2} - (K - c_\pm)\frac{du}{d\chi} + J - g(u) = 0, \tag{3.8}$$

$$u(-\infty; c_+) = u_1(J), \qquad u(+\infty; c_+) = u_3(J), \tag{3.9}$$

$$u(-\infty; c_-) = u_3(J), \qquad u(+\infty; c_-) = u_1(J), \tag{3.10}$$

where $\chi = x - X_\pm(t)$ ($X_\pm$ is the position of the front at time t, determined by imposing $u(0) = u^0$. $dX_\pm/dt = c_\pm$). Then we have

$$u(\chi; c_-) = u(-\chi; c_+), \qquad c_+ + c_- = 2K. \tag{3.11}$$

The proof is evident. This theorem shows that we only need to construct $u(\chi; c_+)$ and find c_+ in order to have $u(\chi; c_-)$ and c_-.

3.2.2. Saturating source. Now let $g(u)$ be a function with only two branches such as (3.1). A wavefront is the only monotone trajectory connecting $(u_1(J), 0)$ and a given point on the u axis $(u_m, 0)$. There is again a symmetry result for these wavefronts.

THEOREM 2. *Let $u(\chi;c_+)$ be the wavefront satisfying* (3.8), *$\partial u/\partial\chi > 0$ for $-\infty < \chi \equiv x - X_+(t) < \chi_m$, and*

$$\begin{aligned} u(-\infty;c_+) &= u_1(J), \qquad u(0;c_+) = u^0, \\ u(\chi_m;c_+) &= u_m, \qquad \frac{\partial u}{\partial\chi}(\chi_m;c_+) = 0. \end{aligned} \tag{3.12}$$

Here $dX_+/dt = c_+$, and $2\chi_m = l(J,u_m) > 0$ is a function of J and u_m. Then the wavefront satisfying (3.8), *$\partial u/\partial\chi < 0$ for $-\chi_m < \chi \equiv x - X_-(t) < \infty$, $c_- = dX_-/dt$, and*

$$\begin{aligned} u(-\chi_m;c_-) &= u_m, \qquad \frac{\partial u}{\partial\chi}(-\chi_m;c_-) = 0, \\ u(0;c_-) &= u^0, \qquad u(+\infty;c_-) = u_1(J), \end{aligned} \tag{3.13}$$

is such that (3.11) *holds.*

Again the proof is immediate.

Let us now choose a certain $u_m = U(J)$ for each $\phi > u_M$ so that the bias condition (2.2) holds for a pulse made out of the following:

- a backfront $u(\chi;c_+)$ at $x = X_+(t)$, $\chi = x - X_+(t)$; and
- a forefront $u(-\chi;c_+)$ at $X_- = X_+ + 2\chi_m$.

Now $\chi = x - X_-(t)$. The pulse $u(\chi;c_+)$ can be constructed explicitly for the piecewise linear $g(u)$ of (3.1). If we choose $u^0 = u_M$, the solution of (3.8) and (3.9) which is continuous and has a continuous first derivative at $\chi = 0$ is

$$u(\chi;c_+) = u_1(J) + [u_M - u_1(J)]\, e^{\lambda_+\chi}, \quad \chi < 0, \tag{3.14}$$

$$u(\chi;c_+) = u_M + \frac{J-\alpha}{K-c_+}\,\chi + B_+\left[e^{(K-c_+)\chi} - 1\right], \quad \chi > 0. \tag{3.15}$$

Here $u_1(J) = J/\beta$, and

$$\lambda_+ = \frac{K-c_+}{2} + \sqrt{\left(\frac{K-c_+}{2}\right)^2 + \beta}, \tag{3.16}$$

$$B_+ = \frac{1}{K-c_+}\left[\lambda_+[u_M - u_1(J)] - \frac{J-\alpha}{K-c_+}\right], \tag{3.17}$$

$$\begin{aligned} \chi_m &= \frac{1}{K-c_+}\,\ln\left[-\frac{J-\alpha}{(K-c_+)^2\,B_+}\right] \\ &= -\frac{1}{K-c_+}\,\ln\left[1 - \frac{\lambda_+(u_M-u_1)(K-c_+)}{J-\alpha}\right], \end{aligned} \tag{3.18}$$

$$u_m = u_M + \frac{J-\alpha}{K-c_+}\,\chi_m + B_+\left[e^{(K-c_+)\chi_m} - 1\right]. \tag{3.19}$$

If these expressions are inserted in the bias condition (2.2), c_+, χ_m, and u_m may be determined as functions of J for a fixed ϕ. Figures 3 and 4 show the phase planes and the trajectories corresponding to $u(\chi;c_+)$ and $u(\chi;c_-)$, respectively, for a given value of $J = 1.3$. Figure 5 shows χ_m and u_m as functions of J.

There are two important approximations to the wavefronts of Theorem 2, which yield either approximately triangular pulses ($c_+ \neq c_-$, $u_m \gg 1$) or the homoclinic pulse ($c_+ = c_- = K$). These two limiting cases are described in detail in Appendix A.

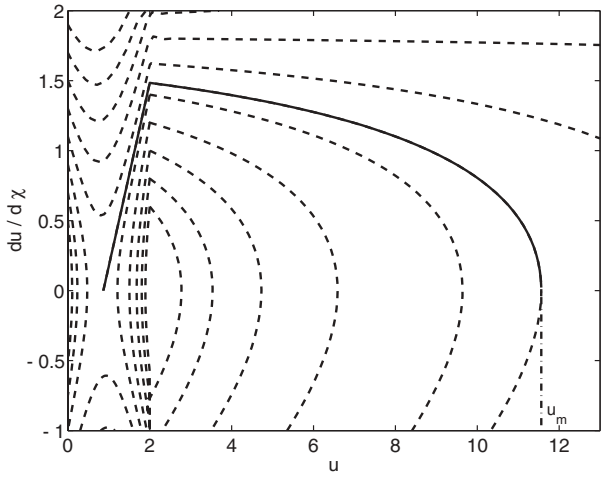


FIG. 3. *Phase plane* $(u, du/d\chi)$ *and trajectory corresponding to* $u(\chi; c_+)$ *for* $J = 1.3$, $c_+ = 1.8359$. *Other parameter values are as in Figure* 2.

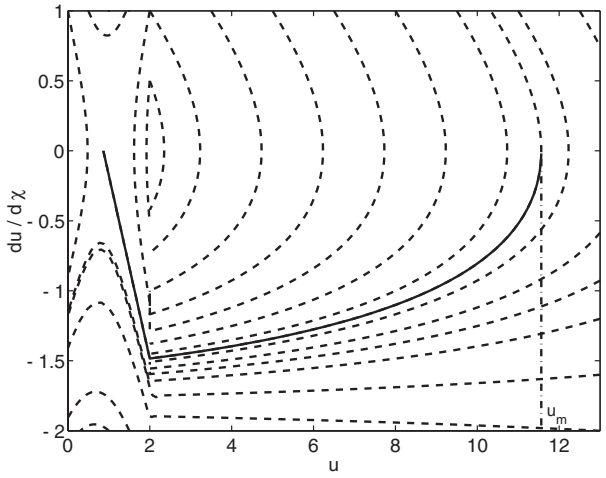


FIG. 4. *Phase plane* $(u, du/d\chi)$ *and trajectory corresponding to* $u(\chi; c_-)$ *for* $J = 1.3$, $c_- = 2.1641$. *Other parameter values are as in Figure* 2.

4. Pulse dynamics.

4.1. One pulse far from the boundaries. Let us consider a single pulse moving far from the boundaries, as described in the previous section. Its height and width are established by imposing the integral condition (2.2). The result will show that the pulse is a tall and narrow moving layer which changes size as it moves. At certain stages of the periodic Gunn oscillation (see below), the pulse height is $u = O(\epsilon^{-\frac{1}{2}})$ and its width is $\Delta y = O(\epsilon^{\frac{1}{2}})$, so that the pulse excess area (in y space units) is $O(1)$. The inner core of the pulse contributes an order 1 amount to the bias whereas its exponential tails approaching $u_1(J)$ yield an $O(\epsilon)$ correction to its excess area. Thus we suggest a decomposition of the solution u into an outer solution valid outside the pulse and boundary layers, and inner solutions comprising the front(s)

and boundary layers. In this section, we shall calculate the leading order term in each asymptotic expansion explicitly. The first correction to these results can be found in Appendix B.

4.1.1. Outer solution. The outer solution is

$$u^{outer} = u^{(0)}(y,s) + \epsilon\, u^{(1)}(y,s) + O(\epsilon^2), \tag{4.1}$$

$$J = J^{(0)}(s) + \epsilon\, J^{(1)}(s) + O(\epsilon^2). \tag{4.2}$$

Here $u^{(0)} = O(1)$ yields an order 1 contribution to the integral condition (2.2), of the same order as that provided by integration of the excess area of the pulse inner core. $\epsilon\, u^{(1)}$ yields an order ϵ contribution to (2.2), of the same order as that provided by integration of the excess area of the pulse tails. Inserting this ansatz into (2.6), we obtain

$$J^{(0)} - g(u^{(0)}) = 0, \tag{4.3}$$

$$J^{(1)} - g'(u^{(0)})\, u^{(1)} = \frac{\partial u^{(0)}}{\partial s} + K\, \frac{\partial u^{(0)}}{\partial y}. \tag{4.4}$$

Solving (4.3) and (4.4) yields

$$u^{(0)} = u_1(J^{(0)}) = \frac{J^{(0)}}{\beta}, \tag{4.5}$$

$$u^{(1)} = \frac{J^{(1)} - \frac{1}{g_1'}\frac{dJ^{(0)}}{ds}}{g_1'} = \frac{1}{\beta}\left(J^{(1)} - \frac{1}{\beta}\,\frac{dJ^{(0)}}{ds}\right). \tag{4.6}$$

$J^{(0)}$ and $J^{(1)}$ will be found later from the integral constraint.

4.1.2. Two-term description of the pulse. The pulse described in the previous section may be considered a moving inner layer solution. Its height is much larger than 1 and its width much smaller than 1 (the precise orders will be determined later). We shall assume

$$u^{inner} \sim P^{(0)}(y,s;\epsilon) + \epsilon\, P^{(1)}(y,s;\epsilon), \tag{4.7}$$

where $P^{(0)}$ is the pulse solution of (3.8) described in Theorem 2:

$$\begin{aligned} P^{(0)}(y,s;\epsilon) &= u(x - X_+; c_+)\,\theta[\chi_m - (x - X_+)] \\ &\quad + u(X_+ + 2\chi_m - x; c_+)\,\theta[x - X_+ - \chi_m]. \end{aligned} \tag{4.8}$$

$P^{(0)}(y,s;\epsilon)$ and $P^{(1)}(y,s;\epsilon)$ depend on ϵ in such a way that

$$\frac{P^{(1)}(y,s;\epsilon)}{P^{(0)}(y,s;\epsilon)} = O(1) \quad \text{as} \qquad \epsilon \to 0+$$

uniformly in y, s. The inner expansion (4.7) is chosen so that the pulse inner core in $P^{(0)}$ yields an order 1 contribution to the bias condition, whereas its tails together with the inner core of $\epsilon P^{(1)}$ yield an $O(\epsilon)$ contribution. The latter is of the same order as the contribution of the outer solution, $\epsilon u^{(1)}$, to the bias.

In (4.8), the location of the fronts and their velocities are

$$X_+ \sim X_+^{(0)}(t) + \epsilon X_+^{(1)}(t),$$
$$c_+ \sim c_+^{(0)} + \epsilon c_+^{(1)}, \tag{4.9}$$

and similarly for X_- and c_-. Inserting (4.2) and (4.7)–(4.9) in (2.1) and (3.9), we obtain for the trailing front

$$\frac{\partial^2 P^{(0)}}{\partial \chi^2} - (K - c_\pm^{(0)}) \frac{\partial P^{(0)}}{\partial \chi} + J^{(0)} - g(P^{(0)}) = 0, \tag{4.10}$$
$$P^{(0)}(-\infty; c_+^{(0)}) = u^{(0)} = u_1, \quad P^{(0)}(0; c_+^{(0)}) = u_M,$$

$$P^{(0)}(\chi_m; c_+^{(0)}) = u_m, \qquad \frac{\partial P^{(0)}}{\partial \chi}(\chi_m; c_+^{(0)}) = 0, \tag{4.11}$$

whose solution is (4.8), and

$$\frac{\partial^2 P^{(1)}}{\partial \chi^2} - (K - c_\pm^{(0)}) \frac{\partial P^{(1)}}{\partial \chi} - g'(P^{(0)}) P^{(1)} = \frac{\partial P^{(0)}}{\partial s} - J^{(1)} - c_\pm^{(1)} \frac{\partial P^{(0)}}{\partial \chi}, \tag{4.12}$$

$$P^{(1)}(-\infty; c_+^{(0)}) = u^{(1)}, \qquad \frac{\partial P^{(1)}}{\partial \chi}(\chi_m; c_+^{(0)}) = 0. \tag{4.13}$$

The leading front of the pulse obeys similar expressions. The correction $P^{(1)}$ is calculated in Appendix B, in which explicit formulas for piecewise linear $g(u)$ are given.

4.2. General equation for $J^{(0)}$. Using Theorem 2, we can calculate the bias condition for a single pulse moving far from the boundaries as

$$\begin{aligned}\phi = {} & u_1(J^{(0)}) + 2\epsilon \int_0^{\chi_m} (P^{(0)} - u_1)\, d\chi \\ & + 2\epsilon \int_{-\infty}^{0} (P^{(0)} - u_1)\, d\chi + \epsilon \frac{J^{(1)} - g_1' J_s^{(0)}}{{g_1'}^2} \\ & + 2\epsilon^2 \int_0^{\chi_m} \left(P^{(1)} - \frac{J^{(1)} - g_1' J_s^{(0)}}{{g_1'}^2} \right) d\chi \\ & + \epsilon \int_0^{\infty} [U_L(\xi) - u_1]\, d\xi \\ & + \epsilon \int_0^{\infty} [U_R(\xi) - u_1]\, d\xi + O(\epsilon^2).\end{aligned} \tag{4.14}$$

Here the bias is the sum of the areas of the regions inside and outside the moving pulse. The leading order contributions to these areas are the first two terms on the right-hand side, which are of order 1. They are (i) the leading order contribution of the outer solution and (ii) the leading order contribution of the inner core of the pulse. The other terms are $O(\epsilon)$ and correspond to the following:

(i) the tails of the pulse to leading order,

$$2\epsilon \int_{-\infty}^{0} (P^{(0)} - u_1)\, d\chi = \frac{2\epsilon(u_M - u_1)}{\lambda_+},$$

(ii) the second order contribution to the outer solution,

(iii) the second order contribution to the area of the inner core of the pulse,

$2\epsilon^2 \int_0^{\chi_m} P^{(1)}\, d\chi = O(\epsilon)$ (we have ignored a much smaller term of order $\epsilon^2\chi_m$), (iv) the layer at the left boundary, (3.6) and (3.7):

$$\epsilon \int_0^{\infty} (U_L - u_1)\, d\xi = \frac{2\epsilon\,(\rho - \beta^{-1})\, J^{(0)}}{\sqrt{K^2 + 4\beta} - K}.$$

The area of the injecting (left) boundary layer becomes of order 1 when a new pulse is being shed; otherwise it is of order ϵ, as indicated above.

If the pulse is far from the boundaries, only the first two terms are of order 1, and we have

$$\begin{aligned}\phi &= u_1(J^{(0)}) + 2\epsilon \left[\chi_m(u_M - u_1 - B_+)\right. \\ &\qquad \left. + \frac{(J^{(0)} - \alpha)\chi_m^2}{2(K - c_+^{(0)})} + \frac{B_+ \left(e^{(K - c_+^{(0)})\chi_m} - 1\right)}{K - c_+^{(0)}}\right] + O(\epsilon) \\ &= u_1 + 2\epsilon \left[\chi_m(u_M - u_1 - B_+) + \frac{(J^{(0)} - \alpha)\chi_m^2}{2(K - c_+^{(0)})}\right. \\ &\qquad \left. + \frac{B_+\lambda_+(u_M - u_1)}{J^{(0)} - \alpha}\right] + O(\epsilon). \qquad (4.15)\end{aligned}$$

Now we can proceed as sketched in section 2. Equation (3.18) allows us to obtain $c_+^{(0)}$ as a function of $J^{(0)}$ and χ_m,

$$c_+^{(0)} = \Xi(J^{(0)}, \chi_m), \qquad (4.16)$$

for a fixed value of ϕ. Inserting this function in (4.15), we can determine χ_m as a function of $J^{(0)}$. Then (3.19) yields u_m as a function of $J^{(0)}$. The results are certain functions

$$\chi_m = \frac{\varphi(J^{(0)})}{2}, \quad u_m = U(J^{(0)}) \qquad (4.17)$$

that have been depicted in Figure 5. Time differentiation of $2\chi_m = \varphi(J^{(0)})$ yields (2.11)

$$\frac{dJ^{(0)}}{ds} = 2\frac{c_+^{(0)}(J^{(0)}, U(J^{(0)})) - K}{-\varphi'(J^{(0)})}, \qquad (4.18)$$

which describes the time evolution of $J^{(0)}$. Provided that $J^{(0)}(0) > J^*$, $J^{(0)}$ decreases exponentially fast to J^* such that $c_+^{(0)} = c_-^{(0)} = K$. The pulse then moves at constant $J^{(0)} = J^*$ and speed K, and it is a homoclinic orbit of the phase plane (3.8) with $c = K$, as shown in Figure 6.

Figure 7 compares the leading order asymptotic solution with the numerical solution. The upper part shows the time evolution of $J(t)$ and the lower part $u(x,t)$ at the times marked in the upper figure. Notice that initially $J(t)$ decreases exponentially fast to J^*. Consider an instant (point 6) at which the wave is fully developed and far from the boundaries. We observe that the profile of the asymptotic solution

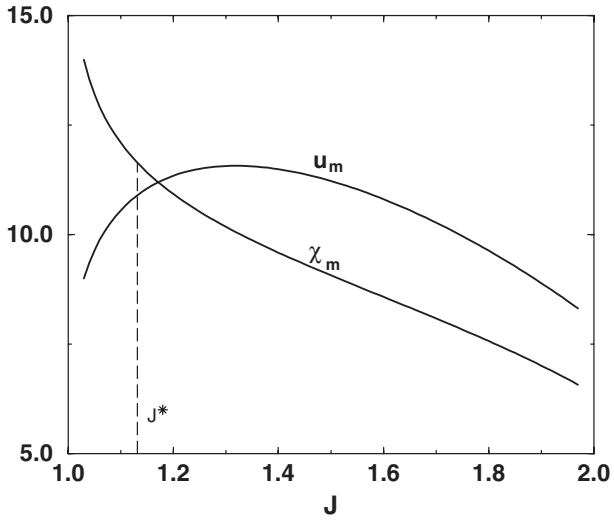


FIG. 5. *Half-width χ_m and maximum height u_m of the single pulse as a function of J for the same parameter values as in Figures 2–4.*

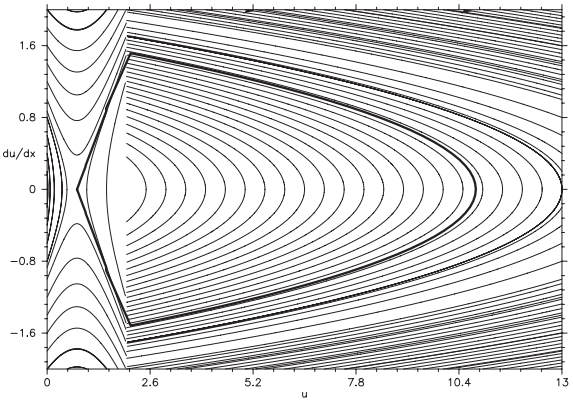


FIG. 6. *Phase plane $(u, du/d\chi)$ and homoclinic orbit for $J^* = 1.13182$, $c = 2$. Other parameter values are as in Figure 2.*

has a larger height and is slightly thinner than the numerical solution. Why? We have neglected $O(\epsilon)$ terms (boundary layers, wave tails, etc.) when calculating the integral condition with the leading order asymptotic solution. Then $J^{(0)}(t)$ is slightly larger than the numerically calculated $J(t)$. The asymptotic profile fully agrees with the homoclinic solution described in Appendix A by (A.10), (A.11), and (A.12) (also calculated excluding order ϵ effects).

The previous ideas are correct if we can show that $J^{(0)}$ evolves on a slow time scale, say $\sigma = \sqrt{\epsilon}(t - t_0)$ or $\tau = \epsilon(t - t_0)$. To see this we should analyze (4.15) and the previous equations with a little more care. We shall show that there are two limiting cases for which the pulse can be easily calculated and (2.11) explicitly obtained to leading order:

- $J^{(0)} - J^* = O(1)$, $K - c_+^{(0)} = O(1)$, and $(J^{(0)} - \alpha)\chi_m \gg 1$. The limiting pulse is a triangular wave formed by pieces of phase plane trajectories which

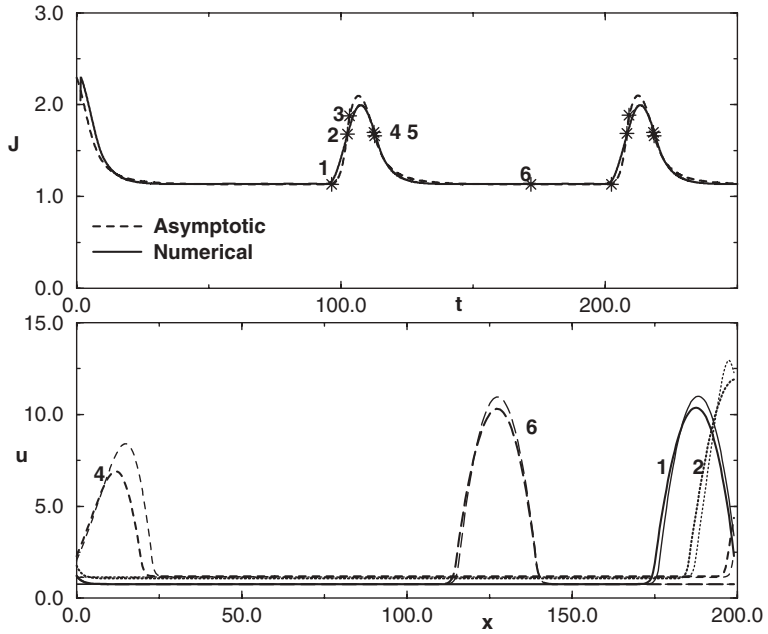


FIG. 7. *Upper: Time evolution of J. Times marked correspond to the following:* (1) *wave reaches receiving contact,* (2) *J reaches J_c,* (3) *$\chi_m < \chi_L$,* (4) *wave completely exits,* (5) *$J < J_c$,* (6) *fully developed wave. Lower: Profiles of $u(x,t)$ at the times marked in the upper figure.*

do not tend exponentially to infinity as $\chi \to \pm\infty$.

- $(K - c_+^{(0)}) \ll (J^{(0)} - \alpha) \ll 1$, $\chi_m \gg 1$. The limiting pulse is the homoclinic trajectory with $c = K$.

In the first case, $J^{(0)}$ evolves on the slow time scale $\sigma = \sqrt{\epsilon}(t - t_0)$. In the second case, the time over which $J^{(0)}$ varies appreciably is $o(\epsilon^{-1})$. One period of the Gunn oscillation contains stages during which these limiting cases are good approximations to (2.11). In fact the Gunn oscillation may be considered as transitions from one limiting case to the other depending on the number of pulses existing at the time. Then $J^{(0)}$ changes slowly with time, which justifies our wavefront constructions.

Appendix C analyzes the dynamics of pulses corresponding to the two limiting cases of triangular and homoclinic pulses. At this point, we call attention to two peculiar features of our results:

1. Our leading order approximation for the solution is of order one outside a pulse, but it is much larger inside it: of order $\epsilon^{-\frac{1}{2}}$.
2. The proper time scale for the variation of J is $\epsilon^{\frac{1}{2}}t$ if the pulse may be approximated by a triangular wave. It is slower as the pulse approaches a homoclinic pulse. Both approximations are limits of the same equation, (2.11). Thus this equation contains more than one asymptotic limit and its time scale is changing as time changes.

5. Pulse dynamics near the boundaries. In the previous section, we described the dynamics of a single pulse far from the boundaries. Basically the pulse approaches the homoclinic pulse on a slow time scale: at each time on the scale t the pulse follows adiabatically the instantaneous value of J. J changes on the scale $\sigma = \sqrt{\epsilon}t$ or even more slowly as described by (2.11), where φ is a function of J and ϵ

($\varphi(J)$ is of order $\epsilon^{-\frac{1}{2}}$ or larger). In this section we shall describe what occurs when the pulse reaches the receiving boundary at $x = L$ and beyond, until a period of the Gunn oscillation is completed.

5.1. Pulse disappearing through the receiving boundary. Let us assume that a single pulse has reached its asymptotic shape ($J^{(0)} = J^*$) and advances with speed K until its forefront reaches $X_- = L$ at time t_1. Afterwards, it begins leaving the sample. As time elapses the wave *exits* through the receiving boundary, and therefore the area under the wave decreases. Since the total area has to satisfy the bias condition, this loss of area has to be compensated by a corresponding increase in u_1 so that (2.2) still holds.

Let us denote by u_L the value of the pulse inner solution $P^{(0)}(y, s)$ at the receiving boundary. u_L is obtained from (3.15) for $u(x - X_-^{(0)}; c_-^{(0)}) = u(X_-^{(0)} - x; c_+^{(0)}) = u(2\chi_m - \chi; c_+^{(0)})$ when $\chi - 2\chi_m = L - X_-^{(0)}$. The corresponding argument of $u(X_-^{(0)} - x; c_+^{(0)})$ is $\chi_L = X_-^{(0)} - L > 0$. When $0 < \chi_L < \chi_m$, the bias condition (2.2) yields

$$\phi = u_1 + \Phi^{(0)}(J, c_+^{(0)}, \chi_m, \chi_L) + \epsilon\, \Phi^{(1)}(J, c_+^{(0)}, \chi_m, \chi_L, J^{(1)}) + O(\epsilon^2), \tag{5.1}$$

$$\begin{aligned}\frac{\Phi^{(0)}(J, c_+^{(0)}, \chi_m, \chi_L)}{\epsilon} &= 2\int_0^{\chi_m} [u(\chi; c_+^{(0)}) - u_1]\, d\chi - \int_0^{\chi_L} [u(\chi; c_+^{(0)}) - u_1]\, d\chi \\ &= (2\chi_m - \chi_L)\,(u_M - u_1 - B_+) + \frac{(J - \alpha)}{(K - c_+^{(0)})}\left(\chi_m^2 - \frac{\chi_L^2}{2}\right) \\ &\quad + \frac{B_+\left(2e^{(K - c_+^{(0)})\chi_m} - e^{(K - c_+^{(0)})\chi_L} - 1\right)}{K - c_+^{(0)}},\end{aligned} \tag{5.2}$$

$$\begin{aligned}\Phi^{(1)}(J, c_+^{(0)}, \chi_m, \chi_L, J^{(1)}) &= \int_{-\infty}^{0} (P^{(0)} - u_1)\, d\chi + \frac{J^{(1)} - g_1' J_s^{(0)}}{{g_1'}^2} \\ &\quad + 2\epsilon\int_0^{\chi_m} P^{(1)}\, d\chi - \epsilon\int_0^{\chi_L} P^{(1)}\, d\chi + \epsilon\int_0^{\infty} [U_L(\xi) - u_1]\, d\xi\end{aligned} \tag{5.3}$$

instead of (4.14). In this equation, $\Phi^{(0)}$ and $\Phi^{(1)}$ are of order 1 because the integrations of $P^{(0)}$ and $P^{(1)}$ over the inner core of the pulse are of order ϵ^{-1}. Equation (5.1) yields

$$\phi = u_1 + \Phi^{(0)}(J^{(0)}, c_+^{(0)}, \chi_m, \chi_L), \tag{5.4}$$

$$\Phi^{(1)}(J^{(0)}, c_+^{(0)}, \chi_L, \chi_m, J^{(1)}) = 0. \tag{5.5}$$

We shall now find the evolution equation for $J^{(0)}$ by a procedure similar to that used to find (4.18). The right-hand side of (5.4) depends on $J^{(0)}$, $c_+^{(0)}$, χ_m, and χ_L. χ_m is a function of $J^{(0)}$ and $c_+^{(0)}$ given by (3.18). As wavefront velocity $c_+^{(0)}$, we shall use the function of $J^{(0)}$ (for a fixed value of ϕ) that was determined at the end of section 3. Then the right-hand side of (5.4) is a function of $J^{(0)}$ and χ_L only (for fixed ϕ):

$$\phi = \mathcal{B}(J^{(0)}, \chi_L) \equiv u_1(J^{(0)}) \quad + \Phi^{(0)}(J^{(0)}, c_+^{(0)}(J^{(0)}), \chi_m(J^{(0)}, c_+^{(0)}(J^{(0)})), \chi_L). \tag{5.6}$$

χ_L can be explicitly calculated from

$$\frac{d\chi_L}{dt} = c_-^{(0)} = 2K - c_+^{(0)}, \quad \chi_L(t_1) = 0, \tag{5.7}$$

where $c_+^{(0)}$ is our known function of $J^{(0)}$. We can obtain a closed system of equations for χ_L and $J^{(0)}$ by differentiating (5.6) with respect to time and then using (5.7). The result is

$$\frac{\partial \mathcal{B}}{\partial J^{(0)}}\,\frac{dJ^{(0)}}{dt} \sim \epsilon\,(u_L - u_1)\,(2K - c_+^{(0)}). \tag{5.8}$$

Here we have used that (5.2) and (5.6) imply $\partial\mathcal{B}/\partial\chi_L = -(u_L - u_1)$.

The time evolution of $J^{(0)}$ is found by solving this equation while the wave disappears through the receiving contact. Having found the solution to leading order, (5.5) yields the correction $J^{(1)}$. Numerical solution of (5.1)–(5.8) shows that $J^{(0)}$ increases with time, as shown in the region between times 1 and 2 in Figure 7. Notice that this increase agrees with the numerical solution of (2.1)–(2.4).

Depending on the bias ϕ, one of the following two events may happen first:

(i)$J^{(0)}$ reaches J_c, or

(ii)$\chi_L = \chi_m$.

In both cases the stage described by the previous equations ends. In case (i), a new wave is created at $x = 0$, whereas in case (ii) (5.1) should be changed to

$$\phi = u_1 + \Psi^{(0)}(J^{(0)}, c_+^{(0)}, \chi_L) + \epsilon\,\Psi^{(1)}(J^{(0)}, c_+^{(0)}, \chi_L, J^{(1)}) + O(\epsilon^2), \tag{5.9}$$

$$\Psi^{(0)}(J^{(0)}, c_+^{(0)}, \chi_L) = \epsilon \int_0^{\chi_L} [u(\chi; c_+^{(0)}) - u_1]\,d\chi$$

$$= \epsilon \left[\frac{(J^{(0)} - \alpha)\chi_L^2}{2(K - c_+^{(0)})} + \chi_L(u_M - u_1 - B_+) + \frac{B_+ \left(e^{(K - c_+^{(0)})\chi_L} - 1\right)}{K - c_+^{(0)}} \right], \tag{5.10}$$

$$\Psi^{(1)}(J, c_+^{(0)}, \chi_L, J^{(1)}) = \int_{-\infty}^{0} [u(\chi; c_+^{(0)}) - u_1]\,d\chi + \frac{J^{(1)} - g_1' J_s^{(0)}}{{g_1'}^2}$$

$$+ \epsilon \int_0^{\chi_L} P^{(1)}\,d\chi + \epsilon \int_0^{\infty} [U_L(\xi) - u_1]\,d\xi, \tag{5.11}$$

where now $\chi_L = L - X_+^{(0)} > 0$ and

$$\frac{d\chi_L}{dt} = -c_+^{(0)}, \quad \chi_L(t_2) = \chi_m(t_2). \tag{5.12}$$

Here t_2 is the time at which the maximum of the pulse reaches $x = L$ (equivalently $\chi_L = \chi_m$ in the previous stage). Arguments similar to those used in the previous stage lead to

$$\left(\frac{1}{\beta} + \frac{\partial \tilde{\Psi}^{(0)}}{\partial J^{(0)}}\right) \frac{dJ^{(0)}}{dt} \sim \epsilon\,(u_L - u_1)\,c_+^{(0)}, \tag{5.13}$$

where $\tilde{\Psi}^{(0)}(J^{(0)}, \chi_L) = \Psi^{(0)}(J^{(0)}, c_+^{(0)}(J^{(0)}), \chi_L)$.

This stage lasts until $J^{(0)}$ reaches J_c and a new pulse is shed. If the bias is small enough, the solution of these equations may be such that $J^{(0)}$ never reaches J_c and it eventually decreases with time. In such case, the pulse exits and leaves a stable stationary state in its wake after $\chi_L = 0$. Notice that setting $\chi_L = \chi_m = 0$ in (3.18) implies a front velocity (A.7). Thus the velocity of the disappearing wavefront approaches that of the triangular wave, although the shapes of the respective wavefronts may differ greatly. As in the previous stage, we find $J^{(1)}$ by solving the equation $\Psi^{(1)}(J^{(0)}, c_+^{(0)}, \chi_L, J^{(1)}) = 0$.

5.2. Pulse shedding at the injecting boundary. If $J(t)$ reaches J_c at $t = t_2$, the boundary layer becomes unstable and a new pulse starts being shed at the injecting boundary. The boundary layer profile, $U(x,t)$, solves the following semi-infinite integrodifferential problem:

$$\frac{\partial U}{\partial \sigma} + K\frac{\partial U}{\partial x} - \frac{\partial^2 U}{\partial x^2} + g(U) = J(\sigma), \tag{5.14}$$

$$U(0,\sigma) = \rho\, J, \tag{5.15}$$

whose solution exhibits an explosion-type behavior and a rapid growth of the area enclosed by the boundary layer which can no longer be neglected. This increase in area is to be compared with the area released by the disappearing pulse at $x = L$. Initially, the area released is larger, and this net area loss has to be compensated by an equal increase in the area of the outer solution, so that J continues to increase, although at a slower rate. After a short time, the growth of the boundary layer is larger than the area released, so that J reaches a maximum and starts decreasing.

The structure of the injecting boundary layer when $J > J_c$ is as follows:

1. $u = u(x; J, U_m)$ is quasi-stationary for $0 < x < X_m$ such that (3.4) holds with $u(0; J, U_m) = \rho J$, $u(X_m; J, U_m) = U_m$, with $\partial u(X_m; J, U_m)/\partial x = 0$. The boundary layer solution is the trajectory of the phase plane corresponding to (3.4) which leaves the vertical line $u = \rho J$ at $x = 0$ and intersects the u axis at $u = U_m$. Notice that this trajectory is uniquely determined by giving J and U_m.
2. For $x > X_m$, the boundary layer is a wavefront of type $u(\chi; c_-)$ moving at speed C_-. This speed and the forefront are uniquely determined by J and U_m: $u = u(\xi; C_+)$, $\xi = Xn_- - x$, with $C_+ = 2K - C_-$, $u(-\infty; C_+) = u_1(J)$, $u(0; C_+) = u_M$, $u(\xi_m; C_+) = U_m$, $\partial u(\xi_m; C_+)/\partial \xi = 0$.

The bias condition (including the injecting boundary layer) is

$$\phi = u_1 + \tilde{\Psi}^{(0)}(J, \chi_m, \chi_L) + \epsilon\, A^{(0)}(t) + O(\epsilon), \tag{5.16}$$

$$A^{(0)} = \int_0^\infty [U(\xi) - U_\infty]\, d\xi, \tag{5.17}$$

provided that the maximum of the old pulse has left the sample. Time differentiation of (5.16) yields

$$\left(\frac{1}{\beta} + \frac{\partial \tilde{\Psi}^{(0)}}{\partial J^{(0)}}\right) \frac{dJ^{(0)}}{dt} \sim -\epsilon \frac{dA^{(0)}}{dt} + \epsilon\,(u_L - u_1)\, c_+^{(0)}. \tag{5.18}$$

Equations (5.12), (5.16), and (5.18), together with (5.2), (5.10), and (4.17), constitute a closed system of equations for the unknowns J, χ_L, and χ_m. During this stage, J initially increases and then decreases until either it again reaches J_c or the old wave completely disappears. In the latter case, the evolution of $J(t)$ is still given by (5.18) without the last term which represents the area lost by the disappearing wave. The evolution during this stage can be observed between points 3 and 4 in Figure 7.

Also, Figure 8 shows in detail the evolution of the boundary layer profile from the time that $J > J_c$. Initially, u grows at the injecting boundary because J increases. Furthermore, the slope $\partial u(0,t)/\partial x$ increases and becomes positive at a certain time. Then a wavelike structure is created. The leading front of this wave moves away from the boundary while the backfront is attached to it. The wave continues its growth

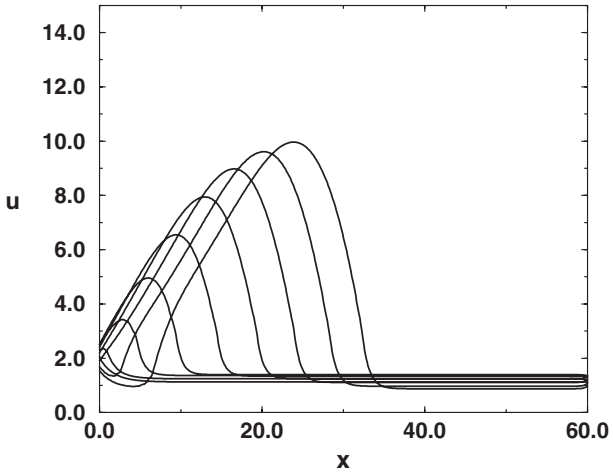


FIG. 8. *Boundary layer profile evolution during the shedding stage.*

until the time when J becomes again smaller than J_c. Then, the slope at the boundary becomes negative, and the wave detaches and moves away from the boundary as a solitary wave. At that time, the quasi-stationary part of the boundary layer which joins $x = 0$ to the maximum at $x = X_m$ becomes the backfront of a detached pulse. Then, we again have the same equation, (2.11), describing the first stage, and a cycle of the Gunn oscillation has been completed. The time evolution during this stage can be observed in Figure 7 for times larger than t_5.

6. Conclusions. In this paper we have asymptotically described one period of a Gunn-type oscillation in a simple model. The nonlinearity of the model is such that at most two constant solutions are possible for each value of J (the currentlike unknown). The model consists of a parabolic equation for the fieldlike unknown, $u(x,t)$, and an integral constraint (bias condition) which determines $J(t)$. Appropriate boundary and initial conditions are imposed. The key new idea of our analysis is that a pulse that changes shape as it advances may be constructed by fixing only two parameters: J and the pulse maximum, u_m. If the pulse width is small compared to the sample length, then L, J, and u_m change on a slow time scale. The trailing front of the pulse is part of a separatrix joining a saddle point to $(u_m, 0)$ with $du/d\chi > 0$ on the $(u, du/d\chi)$ phase plane. The initial and final points determine the backfront speed c_+ as a function of J and u_m. Similarly the forefront of a pulse is constructed and its speed c_- determined. Then equations for J and u_m are obtained by time-differentiation of the bias condition and of the pulse width. The time derivative of the latter is $c_- - c_+$ and J tends toward a fixed value corresponding to a rigidly moving pulse with $c_- = c_+$. Other stages of a Gunn oscillation, including wave creation and annihilation at the boundaries, are analyzed by similar methods. Our theory compares well with direct numerical simulations.

We have found an asymptotic theory of the "Gunn effect" in a simple piecewise linear model, whose main step is a construction of pulses and a derivation of an equation for the "current." An analysis of the stability of these solutions is an open problem, although there is some work on this problem in the related Kroemer model [3, 25]. That the profile of pulses can become oscillatory for appropriate parameter values has been shown by Sun et al. for some reaction-diffusion models [23]. We expect

that the present method yield results independent of the model equations within a class thereof displaying the Gunn instability [6, 21, 22, 25, 26]. Studies of other systems that can be understood by the dynamics of pulses are in progress.

Appendix A. Limiting cases.

A.1. Triangular pulses. The bias condition (2.2) determines the orders of magnitude of χ_m and u_m in terms of the small parameter ϵ. Let us assume that $(K - c_+)$ and $(J - \alpha)$ are $O(1)$, whereas $\chi_m \gg 1$. Then (3.17), (3.18), and (3.19) imply that

$$\begin{aligned} u_m &= u_M + \frac{J-\alpha}{K-c_+}\,\chi_m - \frac{\lambda_+(u_M-u_1)}{K-c_+} \\ &\sim \frac{J-\alpha}{K-c_+}\,\chi_m. \end{aligned} \tag{A.1}$$

The wavefront $u(\chi; c_+)$ is given by (3.14) for $\chi < 0$ and

$$u = u_m + \frac{J-\alpha}{K-c_+}\,(\chi - \chi_m) + B_+\left[e^{(K-c_+)\chi} - e^{(K-c_+)\chi_m}\right] \tag{A.2}$$

$$\sim \frac{J-\alpha}{K-c_+}\,\chi \tag{A.3}$$

for $\chi > 0$, where (A.1) has been used. The bias condition (2.2) then yields

$$\frac{\phi - u_1}{\epsilon} \sim \frac{J-\alpha}{K-c_+}\,\chi_m^2, \tag{A.4}$$

$$\chi_m \sim \sqrt{\frac{(\phi-u_1)\,(K-c_+)}{\epsilon\,(J-\alpha)}}. \tag{A.5}$$

These equations imply that χ_m and u_m are $O(\epsilon^{-\frac{1}{2}})$, while the proper time scale over which J varies is $t = O(\epsilon^{-\frac{1}{2}})$.

We can obtain (A.1)–(A.5) directly from (3.8) and the bias condition. If $\chi_m \gg 1$, so is u_m. Then we shall select uniquely the wavefront solution of (3.8), $u(\chi; c_+)$, so that it satisfies $u(-\infty; c_+) = u_1(J)$ and does not grow exponentially as $\chi \to +\infty$. Similarly, $u(\chi; c_-) = u(-\chi; c_+)$ does not grow exponentially as $\chi \to -\infty$ and satisfies $u(+\infty; c_-) = u_1(J)$. As in Theorem 1, we still have $c_+ + c_- = 2K$.

$u(\chi; c_+)$ satisfies (3.14) and

$$u(\chi; c_+) = u_M + \frac{J-\alpha}{K-c_+}\,\chi \quad \text{if} \quad \chi > 0. \tag{A.6}$$

Continuity of $du/d\chi$ at $\chi = 0$ directly yields

$$\left(u_M - \frac{J}{\beta}\right)\lambda_+ = \frac{J-\alpha}{K-c_+},$$

which in turn implies

$$c_+ = K - \frac{J-\alpha}{\sqrt{\beta\,(u_M-u_1)\,\left(u_M - \frac{\alpha}{\beta}\right)}} > 0. \tag{A.7}$$

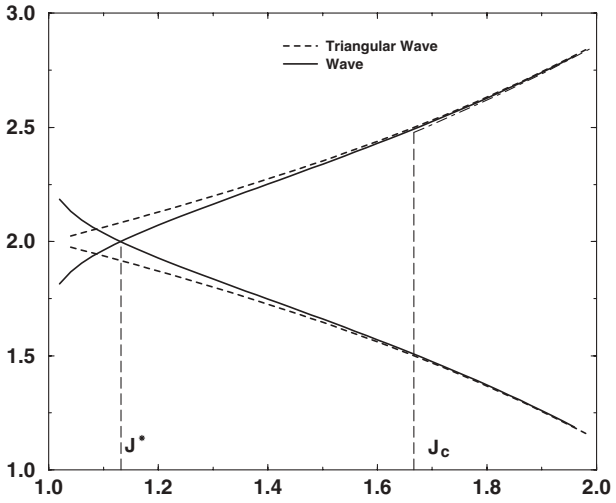


FIG. 9. *Velocities of the backfront, c_+, and of the forefront, c_-, as functions of J for $\phi = 1.6$ and the same parameter values as in Figure 2. We have also shown the corresponding approximate values $c_\pm$ obtained for the triangular pulse described in the text.*

Figure 9 compares the actual values of $c_\pm$ with the approximation (A.7). Notice that both lines are reasonably close for J sufficiently higher than J^*.

We can now form a pulse by joining the backfront $u(\chi; c_+)$, $\chi = x - X_+ < \chi_m$ to the forefront $u(2\chi_m - \chi; c_+)$. This pulse asymptotically approaches an isosceles triangle of basis $(X_- - X_+) = 2\chi_m$ and height approximately given by

$$u_m = u(\chi_m; c_+) = u\left(\frac{X_- - X_+}{2}; c_+\right) \sim \frac{(J-\alpha)\,(X_- - X_+)}{2(K - c_+)}. \tag{A.8}$$

See Figure 10, which compares the triangular pulse to the real pulse and the homoclinic pulse for the same values of J and ϕ. Here we have used (A.6) and assumed that χ_m (the location of the maximum, equal to the pulse half-width) is very large. To be precise, we assume that $(K - c_+) = O(1)$ and that $(J - \alpha)\,\chi_m \gg 1$.

Notice that the way we have constructed $u(\chi; c_+)$ is immediately applicable to a general smooth nonlinearity $g(u)$ of the same type. We have thus the following result.

RESULT. *Let the characteristic curve $g(u)$ be an odd function of u with a positive local maximum after which it monotonically decays to a positive constant as $u \to +\infty$. The approximate backfront $u(\chi; c_+)$ is the unique solution of (3.8) which, for an appropriate choice of the velocity c_+, satisfies $u(-\infty; c_+) = u_1(J)$ and does not grow exponentially as $\chi \to +\infty$.*

A.2. The homoclinic pulse. If $c_+ = c_- = K$, the pulse is a homoclinic orbit of the phase plane (3.8),

$$\frac{\partial^2 u}{\partial \zeta^2} + J - g(u) = 0, \tag{A.9}$$

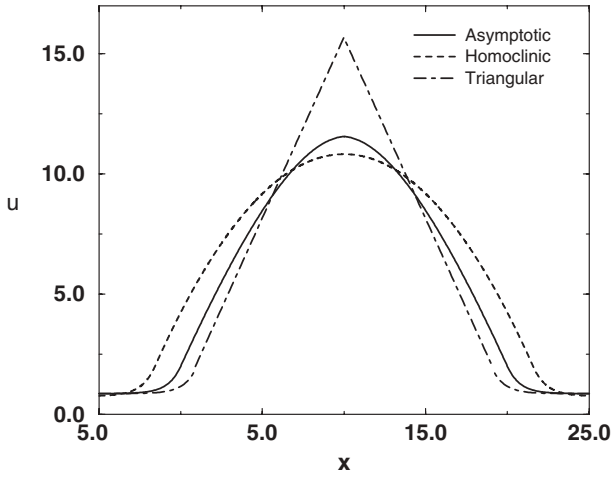


FIG. 10. *Shape of the pulse for $J = 1.3$ and the parameter values of Figure 2. Also shown are the corresponding triangular and homoclinic pulses.*

with $u(\pm\infty) = u_1(J) = J/\beta$ (see Figure 6). Here $\zeta \equiv x - X_0$, and $X_0 = (X_+ + X_-)/2$ is the location of the maximum of the pulse, u_m. The solution is

$$u(\zeta) = \frac{J}{\beta} + \left(u_M - \frac{J}{\beta}\right) e^{-\sqrt{\beta}(|\zeta| - \zeta_0)} \quad \text{for} \quad |\zeta| > \zeta_0, \tag{A.10}$$

$$u(\zeta) = u_m - \frac{J-\alpha}{2}\,\zeta^2 \quad \text{for} \quad |\zeta| < \zeta_0, \tag{A.11}$$

$$u_m = u_M + \frac{J-\alpha}{2}\,\zeta_0^2. \tag{A.12}$$

ζ_0 is determined by imposing continuity of $du/d\zeta$ at $\zeta = \pm\zeta_0$:

$$\zeta_0 = \frac{\sqrt{\beta}\left(u_M - \frac{J}{\beta}\right)}{J-\alpha}. \tag{A.13}$$

Now we may find $J = J^*$ from the bias condition. The result is

$$J^* = \alpha + \beta^{\frac{3}{4}} \sqrt{\frac{2\epsilon\left(u_M - \frac{\alpha}{\beta}\right)^3}{3\left(\phi - \frac{\alpha}{\beta}\right)}} + O(\epsilon), \tag{A.14}$$

$$\zeta_0 = \beta^{-\frac{1}{4}} \sqrt{\frac{3\left(\phi - \frac{\alpha}{\beta}\right)}{2\epsilon\left(u_M - \frac{\alpha}{\beta}\right)}} + O(1), \tag{A.15}$$

$$u_m = \frac{\beta^{\frac{1}{4}}}{2} \sqrt{\frac{3}{2\epsilon}\left(\phi - \frac{\alpha}{\beta}\right)\left(u_M - \frac{\alpha}{\beta}\right)} + O(1). \tag{A.16}$$

Figure 11 compares J^* to the approximation (A.14). Notice that a simple phase plane

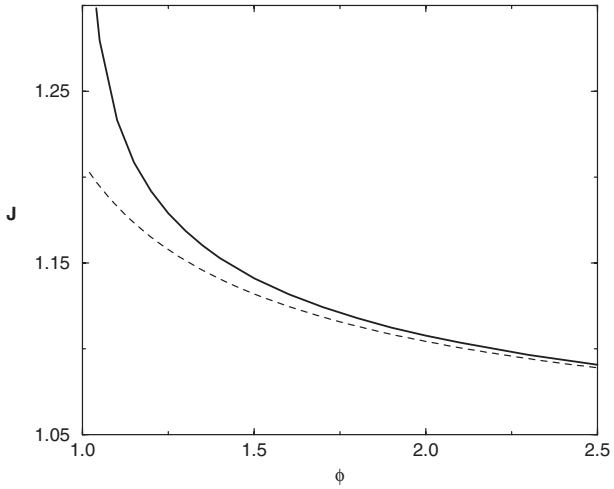


FIG. 11. *Value $J = J^*$ for the homoclinic pulse as a function of the bias ϕ. The dashed curve corresponds to the approximation* (A.14).

argument indicates that J^* obeys the following equal-area rule:

$$J^* = \frac{1}{u_m - u_1} \int_{u_1}^{u_m} g(u)\, du, \tag{A.17}$$

where u_m is given by (A.16). Figure 10 compares the actual pulse, the homoclinic pulse with $c = K$ and the triangular wave for the same values of J and ϕ.

Appendix B. Explicit calculation of $P^{(1)}$ and $c_+^{(1)}$. First of all, $P^{(1)} = P_\chi^{(0)} \equiv \partial P^{(0)}/\partial\chi$ is a solution of the homogeneous problem (4.12) (with zero right-hand side) and (4.13). The solvability conditions for the nonhomogeneous problem yield

$$\left[P_\chi^{(0)}\, P_\chi^{(1)} - P_{\chi\chi}^{(0)}\, P^{(1)}\right]_{\chi=\chi_m} e^{-(K-c_+^{(0)})\chi_m}$$

$$= \int_{-\infty}^{\chi_m} P_\chi^{(0)} \left[P_s^{(0)} - J^{(1)} - c_+^{(1)} P_\chi^{(0)}\right] e^{-(K-c_+^{(0)})\chi} d\chi, \tag{B.1}$$

$$c_+^{(1)} + c_-^{(1)} = 0. \tag{B.2}$$

Let us define

$$P^{(1)} \equiv p(\chi; c_+^{(0)})\, \frac{\partial P^{(0)}}{\partial \chi}. \tag{B.3}$$

Inserting (B.3) in (4.12), we obtain an equation which can be solved by two quadratures. We find

$$\frac{\partial p}{\partial \chi} = \frac{e^{(K-c_+^{(0)})\,\chi}}{P_\chi^{(0)\,2}} \int_{-\infty}^{\chi} e^{-(K-c_+^{(0)})\,\chi} P_\chi^{(0)} \left[P_s^{(0)} - J^{(1)} - c_+^{(1)} P_\chi^{(0)}\right] d\chi. \tag{B.4}$$

Here subscripts indicate partial derivatives with respect to the corresponding variable. It is easy to check that this expression satisfies (B.1). Further integration yields

(for $\chi < 0$)

$$p = \frac{(\beta J^{(1)} - J_s^{(0)})\, e^{-\lambda_+ \chi}}{\beta^2 \lambda_+ (u_M - u_1)} + \frac{\lambda_{+s} \chi^2}{2\lambda_+ \sqrt{(K - c_+^{(0)})^2 + 4\beta}}$$

$$- \frac{\chi}{\sqrt{(K - c_+^{(0)})^2 + 4\beta}} \left[\frac{J_s^{(0)}}{\beta \lambda_+ (u_M - u_1)} + \frac{\lambda_{+s}}{\lambda_+ \sqrt{(K - c_+^{(0)})^2 + 4\beta}} + c_+^{(1)} \right] + q, \tag{B.5}$$

where q is a constant. For $\chi > 0$, $P_{\chi\chi}^{(0)} = B_+ (K - c_+^{(0)})^2 e^{(K - c_+^{(0)})\chi}$, and we can simplify the expression for p by integrating by parts. The result is

$$p = \frac{1}{B_+ (K - c_+^{(0)})^2 P_\chi^{(0)}} \left\{ P_\chi^{(0)} \int_0^\chi e^{-(K - c_+^{(0)})\chi} \left[P_s^{(0)} - J^{(1)} - c_+^{(1)} P_\chi^{(0)} \right] d\chi - \int_{-\infty}^\chi e^{-(K - c_+^{(0)})\chi} P_\chi^{(0)} \left[P_s^{(0)} - J^{(1)} - c_+^{(1)} P_\chi^{(0)} \right] d\chi \right\} + Q. \tag{B.6}$$

Continuity of p at $\chi = 0$ yields

$$q = Q - \frac{J^{(1)} - \frac{J_s^{(0)}}{\beta} + \frac{\beta I_0}{B_+ (K - c_+^{(0)})^2}}{\lambda_+ \beta (u_M - u_1)}, \tag{B.7}$$

where I_0 is the following integral:

$$I_0 \equiv \int_{-\infty}^0 e^{-(K - c_+^{(0)})\chi} P_\chi^{(0)} \left[P_s^{(0)} - J^{(1)} - c_+^{(1)} P_\chi^{(0)} \right] d\chi$$

whose value can be computed as

$$I_0 \;=\; I_{00} \;+\; I_{0J}\, J^{(1)} \;+\; I_{0c}\, c_+^{(1)}, \tag{B.8}$$

where

$$I_{00} = \frac{\lambda_+^2 (u_M - u_1)}{\sqrt{(K - c_+^{(0)})^2 + 4\beta}} \left[\frac{\lambda_+ J_s^{(0)}}{\beta^2} + \frac{(u_M - u_1)\, c_{+s}}{(K - c_+^{(0)})^2 + 4\beta} \right],$$

$$I_{0J} = -\frac{\lambda_+^2 (u_M - u_1)}{\beta},$$

$$I_{0c} = -\frac{\lambda_+^2 (u_M - u_1)^2}{\sqrt{(K - c_+^{(0)})^2 + 4\beta}}.$$

Continuity of $P_\chi^{(1)}$ at $\chi = 0$ implies

$$P_\chi^{(0)}(0)\, [p_\chi(0+) - p_\chi(0-)] = \qquad -p(0) \left[P_{\chi\chi}^{(0)}(0+) - P_{\chi\chi}^{(0)}(0-) \right]. \tag{B.9}$$

The second argument $c_+^{(0)}$ has been omitted in all the functions in this formula. The jump discontinuity of the second derivative $P_{\chi\chi}^{(0)}$ at $\chi = 0$ implies that p_χ also has a jump discontinuity at $\chi = 0$. Substituting (B.5) and (B.6) in (B.9), we obtain

$$QB_+(K - c_+^{(0)})^2 = q\lambda_+^2(u_M - u_1) - \frac{1}{\sqrt{(K - c_+^{(0)})^2 + 4\beta}} \times \left[\frac{J_s^{(0)}}{\beta} + (u_M - u_1)\lambda_+ c_+^{(1)} + \frac{\lambda_{+\,s}(u_M - u_1)}{\sqrt{(K - c_+^{(0)})^2 + 4\beta}}\right]. \tag{B.10}$$

We can obtain Q from (B.7), (B.8), and (B.10). After some algebra, the result is

$$Q \,=\, Q_0 \,+\, Q_J\, J^{(1)} \,+\, Q_c\, c_+^{(1)} \tag{B.11}$$

with

$$\upsilon\, Q_0 \,=\, \frac{J_s^{(0)}}{\beta}\left[\frac{1}{\sqrt{(K - c_+^{(0)})^2 + 4\beta}} \,-\, \frac{\lambda_+}{\beta}\right] + \frac{\lambda_+\, I_{00}}{B_+(K - c_+^{(0)})^2} \,-\, \frac{\lambda_+\,(u_M - u_1)\, c_{+\,s}^{(0)}}{[(K - c_+^{(0)})^2 + 4\beta]^{\frac{3}{2}}},$$

$$\upsilon\, Q_J \,=\, \frac{\lambda_+\, I_{0J}}{B_+(K - c_+^{(0)})^2} \,+\, \frac{\lambda_+}{\beta},$$

$$\upsilon\, Q_c \,=\, \frac{\lambda_+\, I_{0c}}{B_+(K - c_+^{(0)})^2} \,+\, \frac{\lambda_+\,(u_M - u_1)}{\sqrt{(K - c_+^{(0)})^2 + 4\beta}},$$

$$\upsilon \,=\, \lambda_+^2\,(u_M - u_1) \,-\, B_+(K - c_+^{(0)})^2.$$

The velocity correction $c_+^{(1)}$ can be obtained from the condition $\partial P^{(1)}/\partial\chi = 0$ at $\chi = \chi_m$. Thus

$$Q + \frac{1}{B_+(K - c_+^{(0)})^2}\int_0^{\chi_m} e^{-(K - c_+^{(0)})\,\chi} P_\chi^{(0)} \times \left[P_s^{(0)} - J^{(1)} - c_+^{(1)} P_\chi^{(0)}\right] d\chi = 0. \tag{B.12}$$

Inserting (B.11) in this equation, we get $c_+^{(1)}$. After simplification, the result is

$$\gamma\, c_+^{(1)} \,=\, c_{+0} \,+\, c_{+J}\, J_c^{(1)} \tag{B.13}$$

with

$$c_{+0} = \frac{J_s^{(0)}}{K - c_+^{(0)}} \left[z_1 + \frac{z_2}{K - c_+^{(0)}} + (\chi_m + z_2) \left(\frac{\lambda_+}{\beta} + \frac{1}{K - c_+^{(0)}} \right) \right]$$
$$+ \frac{c_{+\,s}^{(0)}}{(K - c_+^{(0)})^2} \left[2(\chi_m + z_2) \left(\frac{J^{(0)} - \alpha}{K - c_+^{(0)}} - \frac{\beta(u_M - u_1)}{\sqrt{(K - c_+^{(0)})^2 + 4\beta}} \right) \right.$$
$$\left. + (J^{(0)} - \alpha) \left(z_1 + \frac{z_2}{K - c_+^{(0)}} \right) - \frac{B_+}{2} (K - c_+^{(0)})^2 \chi_m^2 \right]$$
$$- B_+ Q_0 (K - c_+^{(0)})^2,$$
$$c_{+J} = -z_2 - B_+ Q_J (K - c_+^{(0)})^2,$$
$$\gamma = \frac{J^{(0)} - \alpha}{K - c_+^{(0)}} z_2 - B_+ (K - c_+^{(0)}) \chi_m + B_+ Q_c (K - c_+^{(0)})^2,$$
$$z_1 \equiv \frac{\chi_m \exp[(K - c_+^{(0)})\chi_m]}{(K - c_+^{(0)})} = -\frac{\chi_m B_+ (K - c_+^{(0)})}{J^{(0)} - \alpha},$$
$$z_2 \equiv \frac{\exp[(K - c_+^{(0)})\chi_m] - 1}{(K - c_+^{(0)})} = -\frac{\lambda_+ (u_M - u_1)}{J^{(0)} - \alpha}.$$

Here we have used (3.16)–(3.18) to simplify the result. $J^{(1)}$ is found from the equation

$$0 = 2 \int_{-\infty}^{0} (P^{(0)} - u_1)\, d\chi + \frac{J^{(1)} - g_1' J_s^{(0)}}{{g_1'}^2}$$
$$+ 2\epsilon \int_0^{\chi_m} P^{(1)}\, d\chi + \int_0^{\infty} (U_L(\xi) - u_1)\, d\xi + \int_0^{\infty} (U_R(\xi) - u_1)\, d\xi \tag{B.14}$$

after substitution of (4.14), (4.15), and (B.13).

Appendix C. Pulse dynamics: Limiting cases.

C.1. Triangular pulse. If the limiting pulse is triangular, the approximate evolution equation for J may be obtained by time-differentiating (A.5) and using (A.7):

$$\frac{dJ}{dt} = -\sqrt{\epsilon}\, B(J)\, [K - c_+(J)], \tag{C.1}$$

$$B(J) = \frac{4\beta(\beta u_M - \alpha)^{\frac{1}{4}} (\phi - u_1)^{\frac{1}{2}} (u_M - u_1)^{\frac{5}{4}}}{2u_M - \phi - u_1}. \tag{C.2}$$

For typical values of the parameters such as those in Figure 2, $B(J) > 0$, so that J tends to the solution of $c_+(J) = K$. Triangular pulses are good approximations for J sufficiently large, which means that the solution of (C.1) decreases towards $J = \alpha$ according to (A.7). Of course, before this value can be reached, (C.1) ceases to be valid and we revert to the general equation for J, (2.11), whose fixed point is J^*.

C.2. Homoclinic pulse. Let us now assume that $(K - c_+) \ll (J - \alpha) \ll 1$, whereas $\chi_m \gg 1$. Then (3.17) and (3.18) imply that

$$\chi_m \sim \frac{\lambda_+ (u_M - u_1)}{J - \alpha} + \frac{\lambda_+^2 (u_M - u_1)^2 (K - c_+)}{2(J - \alpha)^2}. \tag{C.3}$$

Here $\lambda_+ \sim \sqrt{\beta}$. Notice that (C.3) becomes (A.13) as $(K - c_+) \to 0$. We now insert this approximation in (3.19) after (3.17) has been substituted. The result is

$$u_m \sim u_M + \frac{\lambda_+^2 (u_M - u_1)^2}{2(J - \alpha)} \sim \frac{1}{2}(J - \alpha)\chi_m^2. \tag{C.4}$$

The bias condition (4.15) may now be approximated by using (C.3) and (C.4) to obtain

$$\frac{\phi - u_1}{\epsilon} \sim \frac{2\beta^{\frac{3}{2}} (u_M - u_1)^3}{3(J - \alpha)^2}.$$

Then

$$\begin{aligned} (J - \alpha) &\sim \beta^{\frac{3}{4}} \sqrt{\frac{2\epsilon (u_M - u_1)^3}{3(\phi - u_1)}} \\ &\sim \beta^{\frac{3}{4}} \sqrt{\frac{2\epsilon \left(u_M - \frac{\alpha}{\beta}\right)^3}{3\left(\phi - \frac{\alpha}{\beta}\right)}}. \end{aligned} \tag{C.5}$$

Inserting (C.5) in (C.3) and (C.4), we obtain

$$\chi_m \sim \beta^{-\frac{1}{4}} \sqrt{\frac{3(\phi - u_1)}{2\epsilon\,(u_M - u_1)}}, \tag{C.6}$$

$$u_m \sim \frac{\beta^{\frac{1}{4}}}{2} \sqrt{\frac{3}{2\epsilon}\,(\phi - u_1)\,(u_M - u_1)}. \tag{C.7}$$

Equations (C.5)–(C.7) are the same as (A.14)–(A.16) for the homoclinic pulse.

As explained before, $d\chi_m/dt = (c_- - c_+)/2 = K - c_+$. Therefore, the derivative of (C.6) with respect to time yields

$$\frac{dJ}{dt} \sim -\sqrt{\frac{8\epsilon}{3}}\, \frac{\beta^{\frac{5}{4}}\, (\phi - u_1)^{\frac{1}{2}} (u_M - u_1)^{\frac{3}{2}}}{u_M - \phi}\, (K - c_+). \tag{C.8}$$

This equation has the same form as (2.11), and shows that the unknown $J(t)$ varies on a slow time scale $t = O(1/\sqrt{\epsilon(K - c_+)})$. In the present limit, $(K - c_+) \ll (J - \alpha) = O(\sqrt{\epsilon})$, so that the corresponding time scale is slower than $\tau = \epsilon t$.

A glance to (C.8) shows that J decreases exponentially fast to J^* such that $c_+ = c_- = K$. The resulting pulse is the homoclinic orbit of the phase plane (3.8) with $c = K$ described in the previous section.

REFERENCES

[1] V. L. Bonch-Bruevich, I. P. Zvyagin, and A. G. Mironov, *Domain Electrical Instabilities in Semiconductors*, Consultants Bureau, New York, 1975.

[2] L. L. Bonilla, *Solitary waves in semiconductors with finite geometry and the Gunn effect*, SIAM J. Appl. Math., 51 (1991), pp. 727–747.

[3] L. L. Bonilla, F. Higuera, and S. Venakides, *The Gunn effect: Instability of the steady state and stability of the solitary wave in long extrinsic semiconductors*, SIAM J. Appl. Math., 54 (1994), pp. 1521–1541.

[4] L. L. Bonilla and F. J. Higuera, *The onset and end of the Gunn effect in extrinsic semiconductors*, SIAM J. Appl. Math., 55 (1995), pp. 1625–1649.

[5] L. L. Bonilla, I. R. Cantalapiedra, G. Gomila, and J. M. Rubí, *Asymptotic analysis of the Gunn effect with realistic boundary conditions*, Phys. Rev. E, 56 (1997), pp. 1500–1510.

[6] L. L. Bonilla and I. R. Cantalapiedra, *Universality of the Gunn effect: Self-sustained oscillations mediated by solitary waves*, Phys. Rev. E, 56 (1997), pp. 3628–3632.

[7] L. L. Bonilla, A. L. Sánchez, and M. Carretero, *The description of homogeneous branched-chain explosions with slow radical recombination by self-adjusting time scales*, SIAM J. Appl. Math., 61 (2000), pp. 528–550.

[8] L. L. Bonilla, *Theory of nonlinear charge transport, wave propagation and self-oscillations in semiconductor superlattices*, J. Phys. Cond. Matter, 14 (2002), pp. R341–R381.

[9] A. E. Bugrim, A. M. Zhabotinsky, and I. R. Epstein, *Calcium waves in a model with a random spatially discrete distribution of Ca2+ release sites*, Biophys. J., 73 (1997), pp. 2897–2906.

[10] P. N. Butcher, W. Fawcett, and C. Hilsum, *A simple analysis of stable domain propagation in the Gunn effect*, Brit. J. Appl. Phys., 17 (1966), pp. 841–850.

[11] M. Büttiker and H. Thomas, *Current instability and domain propagation due to Bragg scattering*, Phys. Rev. Lett., 38 (1977), pp. 78–80.

[12] A. Doelman, W. Eckhaus, and T. J. Kaper, *Slowly modulated two-pulse solutions in the Gray–Scott model* I: *Asymptotic construction and stability*, SIAM J. Appl. Math., 61 (2000), pp. 1080–1102.

[13] H. Hempel, I. Schebesch, and L. Schimansky-Geier, *Traveling pulses in reaction-diffusion systems under global constraints*, Eur. Phys. J. B, 2 (1998), pp. 399–407.

[14] F. J. Higuera and L. L. Bonilla, *Gunn instability in finite samples of GaAs.* II: *Oscillatory states in long samples*, Physica D, 57 (1992), pp. 161–184.

[15] K. Hofbeck, J. Grenzer, E. Schomburg, A. A. Ignatov, K. F. Renk, D. G. Pavel'ev, Yu. Koschurinov, B. Melzer, S. Ivanov, S. Schaposchnikov, and P. S. Kop'ev, *High-frequency self-sustained current oscillation in an Esaki-Tsu superlattice monitored via microwave emission*, Phys. Lett. A, 218 (1996), pp. 349–353.

[16] D. Iron and M. Ward, *A metastable spike solution for a nonlocal reaction-diffusion model*, SIAM J. Appl. Math., 60 (2000), pp. 778–802.

[17] J. P. Keener, *Propagation and its failure in coupled systems of discrete excitable cells*, SIAM J. Appl. Math., 47 (1987), pp. 556–572.

[18] J. P. Keener and J. Sneyd, *Mathematical Physiology*, Springer-Verlag, New York, 1998.

[19] J. Liang, *On a nonlinear integrodifferential semiconductor model*, SIAM J. Math. Anal., 25 (1994), pp. 1375–1392.

[20] J. Rinzel and J. B. Keller, *Traveling waves solutions of a nerve conduction equation*, Biophys. J., 13 (1973), pp. 1313–1337.

[21] V. A. Samuilov, *Nonlinear and chaotic charge transport in semi-insulating semiconductors*, in Nonlinear Dynamics and Pattern Formation in Semiconductors and Devices, F.-J. Niedernostheide, ed., Springer-Verlag, Berlin, 1995, pp. 220–249.

[22] M. P. Shaw, H. L. Grubin, and P. R. Solomon, *The Gunn-Hilsum Effect*, Academic Press, New York, 1979.

[23] W. Sun, T. Tang, M. J. Ward, and J. Wei, *Numerical challenges for resolving spike dynamics for two one-dimensional reaction-diffusion systems*, Stud. Appl. Math., 111 (2003), pp. 41–84.

[24] S. W. Teitsworth, *The physics of space charge instabilities and temporal chaos in extrinsic semiconductors*, Appl. Phys. A, 48 (1989), pp. 127–136.

[25] A. F. Volkov and Sh. M. Kogan, *Physical phenomena in semiconductors with negative differential conductivity*, Sov. Phys. Usp., 11 (1969), pp. 881–903. (Usp. Fiz. Nauk., 96 (1968), pp. 633–672.)

[26] A. Wacker, *Semiconductor superlattices: A model system for nonlinear transport*, Phys. Rep., 357 (2002), pp. 1–111.

SIAM J. APPL. MATH.
Vol. 65, No. 3, pp. 1080–1099

GENERALIZED AZIMUTHAL SHEAR DEFORMATIONS IN COMPRESSIBLE ISOTROPIC ELASTIC MATERIALS*

ELEFTHERIOS KIRKINIS† AND HUNGYU TSAI‡

Abstract. In this article we study the azimuthal shear deformations in a compressible isotropic elastic material. This class of deformations involves an azimuthal displacement as a function of the radial and axial coordinates. The equilibrium equations are formulated in terms of the Cauchy–Green strain tensors, which form an overdetermined system of partial differential equations for which solutions do not exist in general. By means of a Legendre transformation, necessary and sufficient conditions for the material to support this deformation are obtained explicitly, in the sense that every solution to the azimuthal equilibrium equation will satisfy the remaining two equations. Additionally, we show how these conditions are sufficient to support *all* currently known deformations that locally reduce to simple shear. These conditions are then expressed both in terms of the invariants of the Cauchy–Green strain and stretch tensors. Several classes of strain energy functions for which this deformation can be supported are studied. For certain boundary conditions, exact solutions to the equilibrium equations are obtained.

Key words. nonlinear elasticity, Legendre transforms, constitutive laws, azimuthal shear, quasi-linear partial differential equations

AMS subject classifications. 44A15, 274B20, 274D10, 35J25

DOI. 10.1137/S0036139903438077

1. Introduction.

It is well known [8] that the only deformations possible in all isotropic *compressible* elastic materials are homogeneous. Therefore, the analysis of nonhomogeneous deformations can only be accomplished if one concentrates on specific classes of strain energy functions. A growing body of literature that addresses these issues has been developed in recent years. The reader is referred to [12] for a review and relevant references.

One important nonhomogeneous deformation that attracts our attention is the *generalized azimuthal shear* deformation. Its perceived importance stems from the fact that it presents an analogous (but more complicated) kinematic structure to that of the *antiplane shear* deformation; hence, it may emerge as the impetus of developments and analysis on issues such as loss of ellipticity, crack problems cavitation, and phase transitions. The generalized azimuthal shear is an *isochoric* deformation of the form

$$r = R, \quad \theta = \Theta + g(R, Z), \quad z = Z, \tag{1.1}$$

with (R, Θ, Z) and (r, θ, z) being the cylindrical polar coordinates in the unstressed natural configuration and the deformed configuration, respectively. This deformation (or its Z-independent specialization) may also appear under the names of *circular* or *rotational* shear. The function $g(R, Z)$ has to be determined by the equilibrium equations and it depends on the form of strain energy function employed. This deformation belongs to a class of isochoric deformations that locally reduce to simple shear;

*Received by the editors December 1, 2003; accepted for publication (in revised form) March 15, 2004; published electronically March 31, 2005.
http://www.siam.org/journals/siap/65-3/43807.html
†Department of Applied Mathematics, University of Washington, Seattle, WA 98195 (kirkinis@amath.washington.edu).
‡Department of Mechanical Engineering, Michigan State University, East Lansing, MI 48824 (hytsai@egr.msu.edu).

the other well-known deformations in this class are the *helical shear* deformation [3] and the *antiplane shear* deformation [10, 17].

Research in this area was initiated by Knowles [17], in his study of antiplane shear, who recognized that compressible materials can undergo isochoric deformations (which locally reduce to simple shear) only if the strain energy function is consistent with certain restrictive conditions.

A special case of the deformation under study is the *pure azimuthal* shear deformation ($g = g(R)$), for which conditions on the strain energy function were derived by Polignone and Horgan [20] and explicit necessary and sufficient conditions were found by Beatty and Jiang [2]. More recently, Horgan and Saccomandi [13], gave a detailed theoretical analysis and computed all mechanical quantities of interest for the pure azimuthal shear deformation in the case of incompressible materials exhibiting limiting chain extensibility.

A second special case of the deformation under consideration is the *pure torsion* deformation ($g = \tau Z$), investigated by Beatty and Jiang [2], who also derived several constitutive assumptions to support this special deformation. For the same problem, explicit necessary and sufficient conditions on the strain energy function were derived by Polignone and Horgan [18]. Throughout the present study appropriate comparisons with the deformations of *helical shear* and *antiplane shear* will also be made.

Closed form and approximate solutions for special forms of the azimuthal displacement $g = g(R, Z)$, namely pure azimuthal shear and combined pure azimuthal shear/pure torsion, were considered by Tao, Rajagopal, and Wineman [22]. They considered a generalized power law neo-Hookean material in the framework of the incompressible theory, without reference, however, to the locally simple shear character of the deformation as well as the need for discussion of necessary and sufficient conditions on the form of the strain energy function.

In this paper we present the first result available in the literature for the deformation (1.1), in the context of the finite theory of elasticity. Section 2 describes the kinematics, stress, and equilibrium associated with the deformation (1.1). The equilibrium equations are expressed in terms of the Cauchy stress tensor $\boldsymbol{\sigma}$. New universal relations are derived, and a discussion of uniqueness of solution and ellipticity of the governing displacement equations is included. Section 3 describes the transformation method of the equilibrium equations from the reference configuration space to the reference strain space. The equilibrium equations then are obtained in strain space coordinates and their form is used in section 4 to derive necessary and sufficient conditions for the strain energy function to admit the generalized azimuthal shear deformation. In the same section, a straightforward comparison with results from the current literature shows that the above conditions are clearly sufficient to support all currently known isochoric deformations that locally reduce to simple shear, though there might be others whose structure has not been examined yet. These results are then utilized to obtain conditions in terms of the principal invariants of the Cauchy–Green and stretch tensors, in a fashion similar to the discussion of helical shear in [15]. These simple restrictions are then combined, first to determine classes of strain energy functions for which generalized azimuthal shear deformations are possible and second, in section 5, to obtain closed form solutions of $g(R, Z)$ for particular members of these classes. The solutions are subjected to some physically realistic boundary conditions which give further insight into this newly studied deformation. Further, the solutions are compared with the special forms of pure azimuthal shear and pure torsion, analyzed in the literature recently, and some Riemann type similarity solu-

tions are obtained that serve as a test for the validity of the necessary and sufficient conditions we derived earlier. Finally, in this section the effect of torsion on the form of the pure azimuthal displacement is examined for two members of the previously determined energy classes and comparison with previous results of the pure azimuthal shear deformation is made, in the spirit of the work by Tao, Rajagopal, and Wineman [22]. Finally, in section 6 we close with some concluding remarks and compare further the conditions on the strain energy for the generalized azimuthal shear problem with its counterparts, e.g., the antiplane and helical shear deformations. Throughout this article, the notation used closely follows the one adopted by Tsai and Fan [23].

2. Kinematics, strain energy, stress, and equilibrium. We consider a body composed of a compressible, nonlinearly elastic material, occupying the following cylindrical region in its natural (unstressed) configuration:

$$A \le R \le B, \quad 0 \le \Theta \le 2\pi, \quad 0 \le Z \le L, \tag{2.1}$$

where R, Θ, Z are the cylindrical coordinates associated with the reference configuration denoted by $\mathcal{B}_o$. The general azimuthal shear deformation is defined by

$$r = R, \quad \theta = \Theta + g(R, Z), \quad z = Z, \tag{2.2}$$

where r, θ, z are the cylindrical coordinates of a material point in the current configuration denoted by $\mathcal{B}$. In vector notation, the reference and current configurations of the body are related through a mapping $\boldsymbol{\chi}$,

$$\boldsymbol{\chi} : \mathcal{B}_o \to \mathcal{B}, \quad \text{such that} \quad \boldsymbol{x} = \boldsymbol{\chi}(\boldsymbol{X}), \tag{2.3}$$

where $\boldsymbol{x} = r\boldsymbol{e}_r + z\boldsymbol{e}_z$ and $\boldsymbol{X} = R\boldsymbol{E}_R + Z\boldsymbol{E}_Z$ are the position vectors of a particle in the current and reference configurations, respectively, while $\boldsymbol{e}_r, \boldsymbol{e}_z, \boldsymbol{E}_R, \boldsymbol{E}_Z$ are the corresponding radial and axial unit cylindrical polar vectors. The mapping $\boldsymbol{\chi}$ is assumed to be at least twice continuously differentiable. The deformation gradient associated with the generalized azimuthal shear deformation (1.1) obtains the form

$$\boldsymbol{F} = \boldsymbol{Q} + \mathbf{e}_\theta \otimes (R\, g_{,R}\mathbf{E}_R + R\, g_{,Z}\mathbf{E}_Z), \tag{2.4}$$

where

$$\boldsymbol{Q} = \mathbf{e}_r \otimes \mathbf{E}_R + \mathbf{e}_\theta \otimes \mathbf{E}_\Theta + \mathbf{e}_z \otimes \mathbf{E}_Z \tag{2.5}$$

is a local rotation of angle $g(R, Z)$ about Z-axis, mapping the cylindrical axes from the referential basis $\{\mathbf{E}_R, \mathbf{E}_\Theta, \mathbf{E}_Z\}$ to the current basis $\{\mathbf{e}_r, \mathbf{e}_\theta, \mathbf{e}_z\}$. With the notation

$$\omega_r = R\, g_{,R}, \quad \omega_z = R\, g_{,Z}, \tag{2.6}$$

we introduce the *shear strain vectors*, $\boldsymbol{\Omega}$ and $\boldsymbol{\omega}$, associated with the reference and current configurations, respectively, given by

$$\boldsymbol{\Omega} = \omega_r\mathbf{E}_R + \omega_z\mathbf{E}_Z, \quad \boldsymbol{\omega} = \omega_r\mathbf{e}_r + \omega_z\mathbf{e}_z = \boldsymbol{Q}\boldsymbol{\Omega}. \tag{2.7}$$

It then follows that the deformation gradient can be written as

$$\boldsymbol{F} = (\boldsymbol{I} + \mathbf{e}_\theta \otimes \boldsymbol{\omega})\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{I} + \mathbf{E}_\Theta \otimes \boldsymbol{\Omega}). \tag{2.8}$$

Clearly, the deformation is isochoric. It consists of the rotation $\boldsymbol{Q}$ followed by a simple shear, with the amount of shear $\omega \equiv |\boldsymbol{\omega}|$,

$$\omega = (\omega_r^2 + \omega_z^2)^{1/2} = R(g_{,R}^2 + g_{,Z}^2)^{1/2}, \tag{2.9}$$

along the $\mathbf{e}_\theta$ direction. Or equivalently, the deformation consists of a simple shear of amount $\Omega \equiv |\boldsymbol{\Omega}|$, along the direction of $\mathbf{E}_\theta$, followed by the rotation $\boldsymbol{Q}$. Notice that both representations correspond to the same amount of shear $\omega = \Omega$, with glide planes normal to $\boldsymbol{\omega}$ and $\boldsymbol{\Omega}$, respectively. The notation ω and Ω for the amount of shear will be employed interchangeably when need arises.

For the generalized azimuthal deformation (2.2), the left Cauchy–Green strain tensor $\boldsymbol{B} = \boldsymbol{F}\boldsymbol{F}^T$ can be written in terms of the current basis $\{\mathbf{e}_r, \mathbf{e}_\theta, \mathbf{e}_z\}$ as

$$\boldsymbol{B} = \boldsymbol{I} + \boldsymbol{\omega} \otimes \mathbf{e}_\theta + \mathbf{e}_\theta \otimes \boldsymbol{\omega} + \omega^2 \mathbf{e}_\theta \otimes \mathbf{e}_\theta. \tag{2.10}$$

Its invariants are simply

$$I_1 = I_2 = I = \omega_r^2 + \omega_z^2 + 3 = \omega^2 + 3, \quad I_3 = 1. \tag{2.11}$$

Assume the hyperelastic material is homogeneous and isotropic, so its strain energy function Wcan be expressed in terms of the invariants as

$$W = \bar{W}(I_1, I_2, I_3). \tag{2.12}$$

For compatibility with the infinitesimal theory, the strain energy function must satisfy the following restrictions at $I_1 = I_2 = 3$ and $I_3 = 1$,

$$\bar{W}(3,3,1) = 0, \quad \bar{W}_1 + \bar{W}_2 = -(\bar{W}_2 + \bar{W}_3) = \frac{\mu}{2}, \tag{2.13}$$

$$\bar{W}_{11} + 4\bar{W}_{12} + 4\bar{W}_{22} + 2\bar{W}_{13} + 4\bar{W}_{23} + \bar{W}_{33} = \frac{\kappa}{4} + \frac{\mu}{3}, \tag{2.14}$$

where $\bar{W}_i = \partial \bar{W}/\partial I_i$ and $\bar{W}_{ij} = \partial^2 \bar{W}/\partial I_i \partial I_j$; κ and μ are the bulk and shear moduli, respectively.

For the generalized azimuthal deformation (2.2), the Cauchy stress tensor

$$\boldsymbol{\sigma} = 2\bar{W}_1 \boldsymbol{B} + 2\bar{W}_2(I_1 \boldsymbol{B} - \boldsymbol{B}^2) + 2\bar{W}_3 \boldsymbol{I} \tag{2.15}$$

can obtain the alternative representation

$$\boldsymbol{\sigma} = \hat{\sigma}\boldsymbol{I} + \hat{\mu}(\boldsymbol{\omega} \otimes \mathbf{e}_\theta + \mathbf{e}_\theta \otimes \boldsymbol{\omega}) + (\hat{\mu} - \hat{\beta})\omega^2 \mathbf{e}_\theta \otimes \mathbf{e}_\theta - \hat{\beta}\boldsymbol{\omega} \otimes \boldsymbol{\omega}, \tag{2.16}$$

in terms of the material response functions

$$\hat{\sigma}(\omega) = 2[\bar{W}_1 + (I-1)\bar{W}_2 + W_3]\Big|_{I_1 = I_2 = I = 3+\omega^2, I_3 = 1}, \tag{2.17}$$

$$\hat{\mu}(\omega) = 2(\bar{W}_1 + \bar{W}_2)\Big|_{I_1 = I_2 = I = 3+\omega^2, I_3 = 1}, \tag{2.18}$$

$$\hat{\beta}(\omega) = 2\bar{W}_2\Big|_{I_1 = I_2 = I = 3+\omega^2, I_3 = 1} \tag{2.19}$$

by use of relation (2.10). The components of the Cauchy stress in cylindrical coordinates now take the form

$$\sigma_{rr} = \hat{\sigma} - \hat{\beta}\omega_r^2, \quad \sigma_{\theta\theta} = \hat{\sigma} + (\hat{\mu} - \hat{\beta})\omega^2, \quad \sigma_{zz} = \hat{\sigma} - \hat{\beta}\omega_z^2 \tag{2.20}$$

and

$$\sigma_{r\theta} = \hat{\mu}\omega_r, \quad \sigma_{z\theta} = \hat{\mu}\omega_z, \quad \sigma_{rz} = -\hat{\beta}\omega_r\omega_z. \tag{2.21}$$

It will be convenient in what follows to introduce the notation

$$\hat{W}(\omega) = \bar{W}(3 + \omega^2, 3 + \omega^2, 1). \tag{2.22}$$

The material's shear stress response $\hat{\tau}(\omega)$ with respect to simple shear is then given by

$$\hat{\tau}(\omega) = \hat{W}'(\omega) = 2\omega(\bar{W}_1 + \bar{W}_2)\Big|_{I_1=I_2=3+\omega^2, I_3=1}. \tag{2.23}$$

The response function $\hat{\mu}$ defined above is then the secant modulus with respect to simple shear

$$\hat{\mu} = \hat{\tau}(\omega)/\omega = \hat{W}'(\omega)/\omega. \tag{2.24}$$

One can readily show that for the deformation at hand, $\hat{\tau} = \sqrt{\sigma_{r\theta}^2 + \sigma_{z\theta}^2} = \hat{\mu}\omega$ is the resolved shear stress. Note that as $\omega \to 0$, the secant modulus recovers the shear modulus of the infinitesimal theory, $\hat{\mu}(0) = \mu$. Also note that, for consistency with (2.13) and (2.14), $\hat{W}(\omega)$ must satisfy

$$\hat{W}(0) = 0, \quad \hat{W}'(0) = 0, \quad \hat{W}''(0) = \mu > 0. \tag{2.25}$$

Further, we assume $\hat{W}$ satisfies the following condition:

$$\hat{W}'(\omega) > 0 \quad \text{for } \omega > 0. \tag{2.26}$$

This can be viewed as a specialization to simple shear of the Baker–Ericksen inequality. We may also assume that $\hat{W}''(\omega) > 0$ so that $\hat{\tau}$ is a monotonic increasing function of ω; hence increasing shear corresponds to increasing stress, although this restriction can be relaxed if need be.

In view of (2.11) and (2.12), the stress components (2.20)–(2.21) depend on ω and in turn depend on R and Z, and equivalently on r and z by (2.2). The equilibrium equations $\operatorname{div}\boldsymbol{\sigma} = \mathbf{0}$ specialize to

$$\frac{\partial\sigma_{rr}}{\partial r} + \frac{\partial\sigma_{rz}}{\partial z} + \frac{1}{r}(\sigma_{rr} - \sigma_{\theta\theta}) = 0, \tag{2.27}$$

$$\frac{\partial\sigma_{r\theta}}{\partial r} + \frac{\partial\sigma_{\theta z}}{\partial z} + \frac{2}{r}\sigma_{r\theta} = 0, \tag{2.28}$$

$$\frac{\partial\sigma_{rz}}{\partial r} + \frac{\partial\sigma_{zz}}{\partial z} + \frac{1}{r}\sigma_{rz} = 0. \tag{2.29}$$

Using the notation in (2.21), it can be shown that the azimuthal equation (2.28) can be written as

$$\frac{\partial}{\partial R}\Big(R^3\hat{\mu}\Big(R\sqrt{g_{,R}^2 + g_{,Z}^2}\Big)g_{,R}\Big) + \frac{\partial}{\partial Z}\Big(R^3\hat{\mu}\Big(R\sqrt{g_{,R}^2 + g_{,Z}^2}\Big)g_{,Z}\Big) = 0, \tag{2.30}$$

where we have recognized the fact that $r = R$ and $z = Z$ from (2.2). A more compact expression takes the following form:

$$\nabla \cdot (R^3 \hat{\mu}(R|\nabla g|)\nabla g) = 0, \tag{2.31}$$

where ∇ should be interpreted as the gradient operator in the two-dimensional (R, Z)–space.

Throughout the rest of this article we will assume that the equation governing the form of the azimuthal displacement is locally elliptic at a solution g at the reference point $\mathbf{X}$ [7]. It is not difficult to show that this requirement is equivalent to the two inequalities

$$\hat{W}''(\omega) > 0, \quad \hat{W}'(\omega)/\omega > 0 \tag{2.32}$$

for all $\omega > 0$, which we have already adopted. These two inequalities coincide with those imposed by Knowles [17] for the antiplane shear problem to ensure the ellipticity of the governing axial equation and the conditions imposed in [15] for the problem of helical shear, to ensure uniqueness of solution for moderate values of the applied loading. Furthermore, a concise expression adopted in [23], incorporating both relations in (2.32), can be written in the form

$$\hat{\mu}(\hat{\mu}\omega)' > 0, \tag{2.33}$$

where the prime signifies differentiation with respect to the shear ω.

It would be beneficial to consider the equations of equilibrium in an alternative form for reference in the discussion of boundary value problems for the generalized azimuthal shear problem in section 5. In terms of the nominal stress tensor $\boldsymbol{S} = \boldsymbol{F}^{-1}\boldsymbol{\sigma}$, the equilibrium equations $\mathrm{Div}\boldsymbol{S} = \mathbf{0}$ obtain the form

$$\frac{\partial S_{Rr}}{\partial R} + \frac{\partial S_{Zr}}{\partial Z} - g_{,R}S_{R\theta} - g_{,Z}S_{Z\theta} + \frac{1}{R}(S_{Rr} - S_{\Theta\theta}) = 0, \tag{2.34}$$

$$\frac{\partial S_{R\theta}}{\partial R} + \frac{\partial S_{Z\theta}}{\partial Z} + g_{,R}S_{Rr} + g_{,Z}S_{Zr} + \frac{1}{R}(S_{R\theta} + S_{\Theta r}) = 0, \tag{2.35}$$

$$\frac{\partial S_{Rz}}{\partial R} + \frac{\partial S_{Zz}}{\partial Z} + \frac{1}{R}S_{Rz} = 0, \tag{2.36}$$

where the components of the nominal stress tensor $\boldsymbol{S}$ are given by

$$S_{Rr} = 2\bar{W}_1 + 2\bar{W}_2(2 + \omega_z^2) + 2\bar{W}_3, \quad S_{R\theta} = 2\omega_r\left(\bar{W}_1 + \bar{W}_2\right), \tag{2.37}$$

$$S_{Rz} = S_{Zr} = -2\omega_z\omega_r\bar{W}_2, \quad S_{Z\theta} = 2\omega_z\left(\bar{W}_1 + \bar{W}_2\right), \tag{2.38}$$

$$S_{\Theta r} = -2(\bar{W}_2 + \bar{W}_3)\omega_r, \quad S_{\Theta z} = -2(\bar{W}_2 + \bar{W}_3)\omega_z, \tag{2.39}$$

$$S_{\Theta\theta} = 2(I - 2)\bar{W}_1 + 2(I - 1)\bar{W}_2 + 2\bar{W}_3, \tag{2.40}$$

$$S_{Zz} = 2\bar{W}_1 + 2(2 + \omega_r^2)\bar{W}_2 + 2\bar{W}_3, \tag{2.41}$$

evaluated for $I_1 = I_2 = I = 3 + \omega_r^2 + \omega_z^2 = 3 + \omega^2, I_3 = 1$. For nominal stress, we have $S_{R\theta} = \hat{\mu}\omega_r$ and $S_{Z\theta} = \hat{\mu}\omega_r$. From (2.21) and (2.38), it is clear that $\sigma_{r\theta} = S_{r\theta}$ and

$\sigma_{z\theta} = S_{Z\theta}$. It should be noted here that (2.34)–(2.36) can be recovered as a special case of the more general form of the equilibrium equations (2.14a)–(2.14c) of reference [21] associated with the more general form of the deformation field

$$r = r(R, \Theta, Z), \quad \theta = \theta(R, \Theta, Z), \quad z = z(R, \Theta, Z). \tag{2.42}$$

In [21] the authors emphasize the nominal stress (material) formulation of the equilibrium equations in contrast to a Cauchy stress (spatial) formulation and describe why the former representation is more advantageous in specific cases.

3. Transformation of the equations of equilibrium. In general, the equations of equilibrium (2.27)–(2.29) form an overdetermined system for the azimuthal displacement $g(R, Z)$. Specifically, for a suitable secant modulus, the azimuthal equation (2.30) determines a unique g. This solution does not always satisfy the other two equations unless additional restrictions are imposed on the form of the strain energy function W. When these conditions are met, every solution of the azimuthal equation (2.28) will automatically satisfy the other two equations of equilibrium. Then, the system (2.27)–(2.29) reduces to a single quasi-linear partial differential equation for one unknown function g involving only a single constitutive function—the secant shear modulus $\hat{\mu}$. This is exactly the motivation behind the consideration of such specialized deformation classes: to have a single, well-behaved equation for the displacement. The other two equilibrium equations can be ignored, since they are satisfied automatically. Restrictions on the stored energy function always ensure this to be the case.

In what follows we follow the approach by Knowles [16] and Tsai and Fan [23] and implement one-to-one and smooth mappings from the configuration space to a region in the shear strain space. However, in contrast to the problem of antiplane shear, here the basis vectors in the deformed configuration depend nonlinearly on the reference coordinates R and Z, through their dependence on the unknown function $g(R, Z)$, and this nonlinearity carries over to their counterparts in the strain space.

Let $\mathcal{S}_o$ and $\mathcal{S}$ denote the reference and current strain spaces, respectively. We introduce the *reduced shear* vectors, $\boldsymbol{\Gamma} \in \mathcal{S}_o$ and $\boldsymbol{\gamma} \in \mathcal{S}$, associated with the reference and current configurations, respectively, which are defined by

$$\boldsymbol{\Gamma} = \nabla_X g = g_{,R}\mathbf{E}_R + g_{,Z}\mathbf{E}_Z = \boldsymbol{\Omega}/R, \tag{3.1}$$

$$\boldsymbol{\gamma} = \nabla_x g = g_{,R}\mathbf{e}_r + g_{,Z}\mathbf{e}_z = \boldsymbol{\omega}/R = \boldsymbol{Q}\boldsymbol{\Gamma}. \tag{3.2}$$

Note that g will be considered as a scalar function of $\mathbf{X} \in \mathcal{B}_o$ (the reference configuration space) when associated with $\boldsymbol{\Gamma}$, while it will be considered as a scalar function of $\mathbf{x} \in \mathcal{B}$ (the current configuration) when associated with $\boldsymbol{\gamma}$. For simplicity, we use the same notation for various functional representations of the same quantity unless otherwise indicated, so that, for example, $g(\mathbf{X}) = g(\mathbf{x}) = g(R, Z)$. We can consider the two strain vectors as being related by a mapping $\boldsymbol{\psi}$,

$$\boldsymbol{\psi} : \mathcal{S}_o \to \mathcal{S}, \quad \text{such that} \quad \boldsymbol{\gamma} = \boldsymbol{\psi}(\boldsymbol{\Gamma}) \tag{3.3}$$

and $\boldsymbol{\psi}(\boldsymbol{\Gamma}) = \boldsymbol{Q}\boldsymbol{\Gamma}$. Comparing the mappings in (2.3) and (3.3) we notice that the effect of using variables in the strain space has led to a linear relation $\boldsymbol{\gamma} = \boldsymbol{Q}\boldsymbol{\Gamma}$ between the vectors that define deformation in this space. We continue this analysis by constructing a new spherical coordinate system in the shear strain space $\mathcal{S}_o$ with coordinates (Γ, Φ, Θ) in the reference strain space whose radial unit vector points to the direction of the reduced strain vector $\boldsymbol{\Gamma}$. Therefore, we have

$$\Gamma = |\boldsymbol{\Gamma}|, \qquad \Phi = \arctan\frac{\omega_r}{\omega_z} \tag{3.4}$$

and Θ is the same as in $\mathcal{B}_o$. The associated unit vectors in $\mathcal{S}_o$ can be expressed in terms of their cylindrical counterparts in $\mathcal{B}_o$,

$$\mathbf{E}_\Gamma = \sin\Phi\mathbf{E}_R + \cos\Phi\mathbf{E}_Z, \tag{3.5}$$

$$\mathbf{E}_\Phi = \cos\Phi\mathbf{E}_R + \sin\Phi\mathbf{E}_Z, \tag{3.6}$$

with the azimuthal unit vector $\mathbf{E}_\Theta$ the same as the one in the configuration space. Therefore, in the discussion above we introduced a transformation from the reference configuration space to the reference strain space, characterized by the mapping $\hat{\boldsymbol{\Xi}}$,

$$\hat{\boldsymbol{\Xi}} : \mathcal{B}_o \to \mathcal{S}_o \quad \text{such that} \quad \hat{\boldsymbol{\Xi}}(\mathbf{X}) = \boldsymbol{\Gamma} = \Gamma\mathbf{E}_\Gamma. \tag{3.7}$$

The inverse mapping is denoted by $\hat{\mathbf{X}} : \mathcal{S}_o, \to \mathcal{B}_o$ so that

$$\hat{\mathbf{X}}(\hat{\boldsymbol{\Xi}}(\mathbf{X})) = \mathbf{X} \qquad \text{and} \qquad \hat{\boldsymbol{\Xi}}(\hat{\mathbf{X}}(\boldsymbol{\Gamma})) = \boldsymbol{\Gamma}. \tag{3.8}$$

The maps defined above are summarized on the following diagram:

$$\begin{array}{ccc} \mathcal{B}_o & \xrightarrow{\boldsymbol{\chi}} & \mathcal{B} \\ \hat{\boldsymbol{X}} \uparrow \downarrow \hat{\boldsymbol{\Xi}} & & \downarrow \\ \mathcal{S}_o & \xrightarrow[\boldsymbol{\psi}]{} & \mathcal{S} \end{array}.$$

The analysis above has prepared the grounds for the introduction of the Legendre transformation of the function $g : \mathcal{B}_o \to \mathbb{R}$, given by the conjugate function $G : \mathcal{S}_o \to \mathbb{R}$,

$$G(\Gamma, \Phi) = \mathbf{X} \cdot \boldsymbol{\Gamma} - g(R, Z). \tag{3.9}$$

From this expression, the following connections can be derived:

$$R\mathbf{E}_R + Z\mathbf{E}_Z = \mathbf{X} = \nabla_\Gamma G = G_{,\Gamma}\mathbf{E}_\Gamma + \frac{1}{\Gamma}G_{,\Phi}\mathbf{E}_\Phi, \tag{3.10}$$

and hence R and Z can be expressed in terms of Γ and Φ:

$$R = R(\Gamma, \Phi) = \sin\Phi G_{,\Gamma} + \frac{1}{\Gamma}\cos\Phi G_{,\Phi}, \tag{3.11}$$

$$Z = Z(\Gamma, \Phi) = \cos\Phi G_{,\Gamma} - \frac{1}{\Gamma}\sin\Phi G_{,\Phi}. \tag{3.12}$$

We denote the gradient of the mapping $\hat{\boldsymbol{\Xi}}$ in (3.7) by

$$\boldsymbol{H}(\mathbf{X}) = \nabla_X \hat{\boldsymbol{\Xi}}(\mathbf{X}), \quad \forall\, \mathbf{X} \in \mathcal{B}_o. \tag{3.13}$$

The assumption that the mapping be smoothly invertible requires that the Jacobian $\det \boldsymbol{H} \neq 0$; this is exactly the necessary and sufficient condition for the existence of the Legendre transformation (3.9). It follows from (3.8) that the gradient of the inverse mapping is

$$\nabla_\Xi \hat{\mathbf{X}}(\boldsymbol{\Gamma}) = \boldsymbol{H}^{-1}(\mathbf{X}) \text{ evaluated at } \mathbf{X} = \hat{\mathbf{X}}(\boldsymbol{\Gamma}). \tag{3.14}$$

Using this mapping we can now calculate the components of the gradient of the vector $\mathbf{X}$ in the reference strain basis as

$$\boldsymbol{H}^{-1} = \begin{pmatrix} G_{,\Gamma\Gamma} & \left(\frac{1}{\Gamma}G_{,\Phi}\right)_{,\Gamma} & 0 \\ \left(\frac{1}{\Gamma}G_{,\Phi}\right)_{,\Gamma} & \frac{1}{\Gamma}\left(\frac{1}{\Gamma}G_{,\Phi\Phi}+G_{,\Gamma}\right) & 0 \\ 0 & 0 & \frac{R(\Gamma,\Phi)}{\Gamma\sin\Phi} \end{pmatrix}, \tag{3.15}$$

where $R(\Gamma,\Phi)$ is given by (3.11). For the calculation of the equilibrium equations we will also need the inverse of the gradient of $\hat{\mathbf{X}}$ given by

$$\boldsymbol{H} = \frac{1}{\bar{D}}\begin{pmatrix} \frac{1}{\Gamma}\left(\frac{1}{\Gamma}G_{,\Phi\Phi}+G_{,\Gamma}\right) & -\left(\frac{1}{\Gamma}G_{,\Phi}\right)_{,\Gamma} & 0 \\ -\left(\frac{1}{\Gamma}G_{,\Phi}\right)_{,\Gamma} & G_{,\Gamma\Gamma} & 0 \\ 0 & 0 & \frac{\Gamma\sin\Phi}{R(\Gamma,\Phi)}\bar{D} \end{pmatrix}, \tag{3.16}$$

where $\bar{D} = \bar{D}(\Gamma,\Phi)$ is the determinant of the upper-left two-by-two submatrix of $\boldsymbol{H}^{-1}$, i.e.,

$$\bar{D}(\Gamma,\Phi) = \frac{1}{\Gamma}G_{,\Gamma\Gamma}\Big(\frac{1}{\Gamma}G_{,\Phi\Phi}+G_{,\Gamma}\Big) - \left[\Big(\frac{1}{\Gamma}G_{,\Phi}\Big)_{,\Gamma}\right]^2. \tag{3.17}$$

The nominal stress tensor in the strain space coordinates, $\boldsymbol{S}^*(\boldsymbol{\Gamma}) = \boldsymbol{S}(\hat{\mathbf{X}}(\boldsymbol{\Gamma}))$ can be written as

$$\begin{aligned} \boldsymbol{S}^* &= (\hat{\sigma}-\hat{\beta}R^2\Gamma^2)\mathbf{E}_\Gamma\otimes\mathbf{e}_\gamma + \hat{\sigma}\mathbf{E}_\Phi\otimes\mathbf{e}_\phi + (\hat{\sigma}-\hat{\beta}R^2\Gamma^2)\mathbf{E}_\Theta\otimes\mathbf{e}_\theta \\ &+\hat{\mu}R\Gamma\mathbf{E}_\Gamma\otimes\mathbf{e}_\theta + (\hat{\mu}-\hat{\sigma}+\hat{\beta}R^2\Gamma^2)R\Gamma\mathbf{E}_\Theta\otimes\mathbf{e}_\gamma, \end{aligned} \tag{3.18}$$

where the response functions $\hat{\beta}$, $\hat{\sigma}$, and $\hat{\mu}$ are evaluated at $\omega = R\Gamma$. The radial, polar, and azimuthal components of the equilibrium equations, $\nabla_\Xi[\boldsymbol{S}^*]\cdot\boldsymbol{H} = \mathbf{0}$, in the *current* strain space, respectively, obtain the form

$$(\hat{\sigma}-\hat{\beta}R^2\Gamma^2)_{,\Gamma}\bar{E} - (\hat{\sigma}-\hat{\beta}R^2\Gamma^2)_{,\Phi}\bar{B} - \hat{\beta}\Gamma^2R^2\bar{A} - \bar{D}\hat{\mu}\Gamma^2R\sin\Phi = 0, \tag{3.19}$$

$$(-\hat{\sigma}_{,\Gamma}+\hat{\beta}R^2\Gamma)\Gamma\bar{B} + \bar{A}\hat{\sigma}_{,\Phi} + \bar{D}(\hat{\beta}-\hat{\mu})\Gamma^2R\cos\Phi = 0, \tag{3.20}$$

$$(\hat{\mu}R\Gamma)_{,\Gamma}\bar{E} + \hat{\mu}R\Gamma\bar{A} - (\hat{\mu}R\Gamma)_{,\Phi}\bar{B} + 2\bar{D}\hat{\mu}\Gamma\sin\Phi = 0, \tag{3.21}$$

where $\bar{A},\bar{B},\bar{E},\bar{D}$ are functions of Γ,Φ defined as follows: $\bar{A}(\Gamma,\Phi) = \frac{1}{\Gamma}G_{,\Gamma\Gamma}$, $\bar{B}(\Gamma,\Phi) = \frac{1}{\Gamma}\left(\frac{1}{\Gamma}G_{,\Phi}\right)_{,\Gamma}$, $\bar{E}(\Gamma,\Phi) = \frac{1}{\Gamma}\left(\frac{1}{\Gamma}G_{,\Phi\Phi}+G_{,\Gamma}\right)$, and $\bar{D}$ is defined in (3.17). In addition, use has been made of Eulerian strain space quantities and their derivatives with respect to Lagrangian strain space variables, which appear in the appendix. We now consider the real amount of shear $\Omega = R\Gamma$, where the upper case notation is kept throughout this section to emphasize use of quantities in the reference configuration. Using this notation, the equilibrium equations can be transformed into the following equivalent set:

$$\bar{D}\Gamma\sin\Phi\left[(\hat{\sigma}-\hat{\beta}\Omega^2)_{,\Omega}-\hat{\mu}\Omega\right] - \hat{\beta}\Omega^2\bar{A} + \bar{E}R[\hat{\sigma}-\hat{\beta}\Omega^2]_{,\Omega} = 0, \tag{3.22}$$

$$\bar{D}\Gamma\cos\Phi\left[\hat{\sigma}_{,\Omega}+\Omega(\hat{\beta}-\hat{\mu})\right] + R\Gamma\bar{B}[-\hat{\sigma}_{,\Omega}+\hat{\beta}\Omega] = 0, \tag{3.23}$$

$$\bar{D}\Gamma\sin\Phi\left[(\hat{\mu}\Omega)_{,\Omega}+2\hat{\mu}\right] + \hat{\mu}\Omega\bar{A} + \bar{E}R[\hat{\mu}\Omega]_{,\Omega} = 0. \tag{3.24}$$

In the next section, we derive the necessary and sufficient conditions satisfied by the constitutive functions $\hat{\mu}, \hat{\beta}, \hat{\sigma}$ such that the above system collapses into a single equation. Materials that satisfy these conditions are referred to as being able to support a state of generalized azimuthal shear.

4. Necessary and sufficient conditions for materials to support generalized azimuthal shear. In what follows we establish necessary and sufficient conditions on the strain energy function of an isotropic material to support a state of generalized azimuthal shear (1.1) in the sense that every solution to the azimuthal equation of equilibrium will satisfy the remaining two equations. The derivation here differs from those by Knowles [16] and Tsai and Fan [23] in that we do not seek to satisfy a solution of a special form for the azimuthal equation, but we essentially look at a superset of conditions whose combination will provide the required results. Our main result in this section is as follows.

THEOREM 4.1. *A homogeneous, compressible isotropic hyperelastic material with stored-energy function characterized by* (2.12)*, can support a state of generalized azimuthal shear* (1.1) *if and only if the constitutive functions* $\hat{\sigma}$*,* $\hat{\beta}$*, and* $\hat{\mu}$ *characterized by* (2.17), (2.18)*, and* (2.19) *satisfy*

$$\hat{\sigma}_{,\Omega} = \hat{\beta}\Omega, \quad \hat{\beta} = \frac{\hat{\mu}}{2}. \tag{4.1}$$

Proof. To prove sufficiency, we first consider a material that satisfies conditions (4.1). Then the polar component of the equilibrium equations (3.23) is automatically satisfied. Also from the relations (4.1), it follows that

$$\frac{1}{\Omega}(\hat{\sigma} - \hat{\beta}\Omega^2)_{,\Omega} = -(\hat{\beta}\Omega)_{,\Omega} = -\frac{1}{2}(\hat{\mu}\Omega)_{,\Omega} \quad \text{and} \quad \hat{\beta}\Omega = \frac{1}{2}(\hat{\mu}\Omega). \tag{4.2}$$

It then follows that the radial equation (3.22) is equivalent to the azimuthal equation (3.24). Thus, any solution to the azimuthal equation (3.24) satisfies (3.22) and (3.23).

To prove necessity, we consider the necessary and sufficient conditions derived from the literature for pure torsion [18] and pure azimuthal shear [2], which, in our notation, can be rewritten respectively in the form

$$\hat{\sigma}_{,\Omega} = \Omega(\hat{\mu} - \hat{\beta}), \tag{4.3}$$

$$\hat{\sigma}_{,\Omega} = -\frac{1}{2}\left(\Omega(\hat{\mu} - 4\hat{\beta}) + \Omega^2(\hat{\mu} - 2\hat{\beta})_{,\Omega}\right). \tag{4.4}$$

We now assume that a strain energy function supports a state of generalized azimuthal shear (1.1); it follows that it will support *both* a state of pure torsion (and therefore satisfy (4.3)) *and* a state of pure azimuthal shear (and therefore will satisfy (4.4)). Combining the two expressions (4.3) and (4.4) we arrive at the equivalent system

$$3\hat{\mu} + \Omega\hat{\mu}_{,\Omega} = 6\hat{\beta} + 2\Omega\hat{\beta}_{,\Omega}. \tag{4.5}$$

Multiply both sides by Ω^2 and integrate to find

$$\hat{\beta} = \frac{1}{2}\hat{\mu}, \tag{4.6}$$

where the integration constant is taken to be zero for $\Omega \to 0$. Substitute from the above into (4.3) to recover $(4.1)_1$. This concludes the proof of necessity. ∎

4.1. Sufficient conditions on W to support deformations that locally reduce to simple shear. We have already recorded the necessary and sufficient conditions on the strain energy function to support a state of pure torsion in (4.3) (cf. [18]) and a state of pure azimuthal shear in (4.4) (cf. [2, 20]). We augment this set by the necessary and sufficient conditions to support a state of axisymmetric antiplane shear [14, 19] and a state of antiplane shear [17], which, in our notation, can be written, respectively, in the form

$$\hat{\mu}(\hat{\sigma}_{,\Omega} - \hat{\beta}\Omega) = (\hat{\mu}\hat{\beta}_{,\Omega} - \hat{\beta}\hat{\mu}_{,\Omega})\Omega \tag{4.7}$$

and

$$\hat{\sigma}_{,\Omega} = \hat{\beta}\Omega, \quad \hat{\beta} = b\hat{\mu}, \tag{4.8}$$

where b is a constant. It is a straightforward task to deduce then that the conditions (4.1) that support a state of generalized azimuthal shear are *sufficient* to support all deformations above. Therefore, strain energy functions that satisfy conditions (4.1) form a distinctive class of materials to which more attention is devoted in the following two paragraphs.

4.2. Strain energies and necessary and sufficient conditions in terms of I_1, I_2, I_3. In terms of the principal invariants of the left (or right) Cauchy–Green strain tensors the conditions (4.1) take the form

$$\bar{W}_1 = \bar{W}_2, \tag{4.9}$$

$$\bar{W}_1 + 2I(\bar{W}_{11} + \bar{W}_{12}) + 2\bar{W}_{13} + 2\bar{W}_{23} = 0 \tag{4.10}$$

for $I_1 = I_2 = I, I_3 = 1$. The second of these conditions involve second derivatives of the strain energy with respect to the principal invariants. However, it was shown in [15] that the second of these conditions is equivalent to

$$4I\bar{W}_1 + 4\bar{W}_3 - \bar{W} = 0. \tag{4.11}$$

In summary, the following two equations:

$$\bar{W}_1(I, I, 1) = \bar{W}_2(I, I, 1), \quad 4I\bar{W}_1(I, I, 1) + 4\bar{W}_3(I, I, 1) - \bar{W}(I, I, 1) = 0 \tag{4.12}$$

constitute the necessary and sufficient conditions for the strain energy function to admit a state of general azimuthal shear, in terms of the principal invariants of the Cauchy–Green strain tensors.

We may now attempt to derive some possible forms of strain energy functions that satisfy the conditions (4.12) and ideally support a generalized azimuthal shear deformation (1.1). We begin by considering a strain energy function in the form

$$\bar{W}(I_1, I_2, I_3) = f_1(I_1)h_1(I_3) + f_2(I_2)h_2(I_3) + h_3(I_3), \tag{4.13}$$

where f_1 and f_2 are functions to be determined, while h_1, h_2, h_3 have to be compatible with (2.13) and (2.14). Without loss of generality we may set

$$h_1(1) = h_2(1) = 1. \tag{4.14}$$

It was shown in [15], that a strain energy function of the form (4.13) that satisfies the necessary and sufficient conditions (4.12), necessarily has the form

$$\bar{W}(I_1, I_2, I_3) = \frac{3\mu}{4k3^k}\left[I_1^k h_1(I_3) + I_2^k h_2(I_3)\right] + h_3(I_3) \tag{4.15}$$

for $k \neq 0$ and

$$\bar{W}(I_1, I_2, I_3) = \frac{3\mu}{4}\left[\log I_1 h_1(I_3) + \log I_2 h_2(I_3)\right] + h_3(I_3) \tag{4.16}$$

for $k = 0$. A widely used constitutive assumption, being a special case of (4.15) for the value of the constant $k = 1$, is the generalized Hadamard material [1, 14, 18, 19, 20]. Further discussion of this material in the context of the solution of boundary value problems is included in section 5.

4.3. Strain energies and necessary and sufficient conditions in terms of i_1, i_2, i_3. In this subsection we recast the conditions (4.12) as equivalent conditions in terms of the principal invariants of the stretch tensor $\boldsymbol{U}$ arising in the polar decomposition $\boldsymbol{F} = \boldsymbol{RU}$ of the deformation gradient. These are related to the invariants of the left Cauchy–Green strain tensor through

$$I_1 = i_1^2 - 2i_2, \quad I_2 = i_2^2 - 2i_1 i_3, \quad I_3 = i_3^2 \tag{4.17}$$

and are expressed in terms of the principal stretches as

$$i_1 = \lambda_1 + \lambda_2 + \lambda_3, \quad i_2 = \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3, \quad i_3 = \lambda_1\lambda_2\lambda_3. \tag{4.18}$$

Then, by using the notation $\tilde{W}(i_1, i_2, i_3)$ to represent the strain energy when regarded as a function of i_1, i_2, i_3, it is straightforward to show that (4.12) are re-expressed as

$$\tilde{W}_1(i, i, 1) = \tilde{W}_2(i, i, 1), \quad 2i\tilde{W}_1(i, i, 1) + 2\tilde{W}_3(i, i, 1) - \tilde{W}(i, i, 1) = 0, \tag{4.19}$$

where the subscripts denote partial derivatives with respect to i_1, i_2, i_3 and we have considered the first two principal directions to lie on the plane of shear. We consider the strain energy function

$$\tilde{W}(i_1, i_2, i_3) = f(i_1)h_1(i_3) + f(i_2)h_2(i_3) + h_3(i_3), \tag{4.20}$$

which is the counterpart of (4.13), so that $(4.19)_1$ is satisfied provided $h_1(1) = h_2(1)$. We therefore set

$$h_1(1) = h_2(1) = 1. \tag{4.21}$$

Equation $(4.19)_2$ then leads to

$$\tilde{W}(i_1, i_2, i_3) = \frac{3\mu}{k3^k}\left[i_1^k h_1(i_3) + i_2^k h_2(i_3)\right] + h_3(i_3), \tag{4.22}$$

where, in addition to (4.21),

$$h_1'(1) + h_2'(1) = 1 - k, \quad h_3(1) = -\frac{6\mu}{k}, \quad h_3'(1) = -\frac{3\mu}{k} \tag{4.23}$$

for all nonzero values of k, in a manner analogous to [15]. For $k = 1$, the strain energy function has the form

$$\tilde{W}(i_1, i_2, i_3) = \mu\left[i_1 h_1(i_3) + i_2 h_2(i_3)\right] + h_3(i_3). \tag{4.24}$$

For the special case $h_1'(i_3) = h_2'(i_3) \equiv 0$ this energy function reduces to a (compressible) Varga material that was introduced by Carroll [6] (cf. [12]).

4.4. Strain energies and necessary and sufficient conditions in terms of the principal stretches, $\boldsymbol{\lambda_1, \lambda_2, \lambda_3}$. Since the conditions (4.12) are also necessary and sufficient to support a state of helical shear, they can be described in terms of the principal stretches, with $W = W(\lambda_1, \lambda_2, \lambda_3)$, in the form

$$\lambda W_1 + \lambda^{-1} W_2 - W = 0, \quad \lambda W_1 + \lambda^{-1} W_2 - 2W_3 = 0, \tag{4.25}$$

where subscripts denote differentiation with respect to the principal stretches and the strain energy is evaluated at $\lambda_1 = \lambda, \lambda_2 = \lambda^{-1}, \lambda_3 = 1$. The reader is referred to [15] for the details of this derivation. Furthermore, it was shown in [15] that the strain energy function

$$W = f(I_{(\alpha)})h_1(J) + f(I_{(-\alpha)})h_2(J) + h_3(J), \tag{4.26}$$

with $I_{(\alpha)} = \lambda^\alpha + \lambda^{-\alpha} + 1$ and $J = \lambda_1 + \lambda_2 + \lambda_3$, satisfies the conditions (4.25), for any function f and $\alpha \neq 0$, subject to some conditions on the functions h_1, h_2, h_3, which for brevity are not included here.

5. Boundary value problems for generalized azimuthal shear deformations. In this section we intend to investigate specific solutions of the equilibrium equations for materials undergoing generalized azimuthal deformations. To find the general solutions to the equilibrium equations, we could have started, in principle, from the equilibrium equations (3.22)–(3.24) in the strain space and calculated the desired unknown displacement function. This would be lengthy though, if feasible at all, since the transformed equilibrium equations are not linear as was the case for the corresponding equations formulated in strain space coordinates for the antiplane shear deformations [23]. Instead, we concentrate on the equilibrium equations in their configuration space form (2.34)–(2.36). To this end, we substitute the first of the necessary and sufficient conditions,

$$\bar{W}_1 = \bar{W}_2, \tag{5.1}$$

into the equilibrium equations (2.34)–(2.36), which reduce to

$$\frac{\partial}{\partial R}\left((\omega_z^2 + 3)\bar{W}_1 + \bar{W}_3\right) + \frac{\partial}{\partial Z}\left(-\bar{W}_1\omega_r\omega_z\right) + \frac{1}{R}\left(\bar{W}_1(\omega_z^2 - 2I + 6)\right) = 0, \tag{5.2}$$

$$\frac{\partial}{\partial R}\left(\omega_r\bar{W}_1\right) + \frac{\partial}{\partial Z}\left(\bar{W}_1\omega_z\right) + \frac{2}{R}\left(\omega_r\bar{W}_1\right) = 0, \tag{5.3}$$

$$\frac{\partial}{\partial R}\left(-\bar{W}_1\omega_r\omega_z\right) + \frac{\partial}{\partial Z}\left((\omega_r^2 + 3)\bar{W}_1 + \bar{W}_3\right) + \frac{1}{R}\left(-\bar{W}_1\omega_z\omega_r\right) = 0. \tag{5.4}$$

After some lengthy but straightforward calculations, taking into account the identity $\omega_{z,R} - \omega_{r,Z} - \omega_z/R = 0$ and relation (5.1), equations (5.2), (5.3), and (5.4) reduce to

$$\bar{W}_1 + 2I(\bar{W}_{11} + \bar{W}_{12}) + 2\bar{W}_{13} + 2\bar{W}_{23} = 0 \tag{5.5}$$

and

$$\frac{\partial}{\partial R}\left(R^2\gamma_r\bar{W}_1\right) + \frac{\partial}{\partial Z}\left(R^2\gamma_z\bar{W}_1\right) = 0. \tag{5.6}$$

The first of these relations corresponds to the second order necessary and sufficient condition (4.10) required to support a state of generalized azimuthal shear which, furthermore, is equivalent to its first order counterpart

$$4I\bar{W}_1 + 4\bar{W}_3 - \bar{W} = 0. \tag{5.7}$$

Equation (5.6), which equivalently can be rewritten as

$$\nabla \cdot \left(R^3 \bar{W}_1 \nabla g\right) = 0, \tag{5.8}$$

will provide the form of the azimuthal displacement $g(R, Z)$ given a specific form of the strain energy function $\bar{W}$ (that satisfies the conditions (4.12)). Note that the above discussion constitutes an alternative way to show that conditions (4.1) indeed form a set of *sufficient* expressions a strain energy has to satisfy in order to support a state of generalized azimuthal shear (1.1). For the strain energy (4.15) we obtain

$$\hat{W}(\omega) = \frac{3\mu}{2k3^k}(3 + \omega^2)^k - \frac{3\mu}{2k}, \tag{5.9}$$

and hence

$$\hat{W}'(\omega) = \frac{\mu\omega}{3^{k-1}}(3 + \omega^2)^{k-1}, \tag{5.10}$$

from which the inequality $(2.32)_2$ follows. It is then easy to show that $(2.32)_1$ holds for all $\gamma \geq 0$ if and only if

$$k \geq \frac{1}{2}. \tag{5.11}$$

Henceforth, we consider only members of the class (4.15) for which (5.11) holds, and we note that (4.16) is therefore ruled out.

Substitution of (4.15) into the governing equation (5.8) leads to a quasi-linear PDE for $g(R, Z)$ in general. A special form of (4.15) with $k = 1$, the (general) Hadamard material, has been used widely in the literature [1, 11, 14, 18]. It is given by the expression

$$\bar{W}(I_1, I_2, I_3) = \frac{\mu}{4}\left[I_1 h_1(I_3) + I_2 h_2(I_3)\right] + h_3(I_3), \tag{5.12}$$

where

$$h_3(1) = -\frac{3\mu}{2}, \quad h_3'(1) = -\frac{3\mu}{8}, \tag{5.13}$$

while (5.8) reduces to the following linear (and elliptic) equation:

$$g_{,RR} + g_{,ZZ} + \frac{3}{R} g_{,R} = 0. \tag{5.14}$$

The form of the solution will be determined by the boundary conditions. This issue is dealt within the following paragraphs.

5.1. Boundary value problems involving the general solutions.

5.1.1. Example 1. First, the inner surface of the tube at $r = A$ is held fixed, bonded on a rigid cylinder, while the outer surface at $r = B$ is subjected to a given azimuthal displacement $\varphi(Z)$,

$$g(A, Z) = 0, \quad g(B, Z) = \varphi(Z). \tag{5.15}$$

Furthermore, we consider the following additional conditions on the boundary $g(R, 0) = g(R, L) = 0$, $\varphi(0) = \varphi(L) = 0$. We seek separable solutions for the azimuthal displacement of the form

$$g(R, Z) = f_1(R) f_2(Z). \tag{5.16}$$

On substitution of this separable form into (5.14), the boundary conditions are satisfied given the separation constant is chosen to be positive as l^2, i.e., the axial dependence is taken to vary sinusoidally. With the transformation $f_1(R) = x^{-1}y(x)$, $x = lR$, the ODE involving the radial coordinate reduces to the modified Bessel equation for y of order one. Since the configuration involves a tube of finite radius, both solutions of this equation, the modified Bessel functions of the first kind (I_1) and second kind (K_1) are involved. The final result for the l-mode of $g(R, Z)$, g_l is now given by

$$g_l(R, Z) = \frac{1}{lR} \left[C_l I_1(lR) + D_l K_1(lR) \right] \left[A_l \cos lZ + B_l \sin lZ \right], \tag{5.17}$$

where A_l, B_l, C_l, D_l are constants to be determined by the boundary conditions. Employing the boundary conditions (5.15), the azimuthal displacement can now be written as

$$g(R, Z) = \sum_{n=1}^{\infty} \frac{1}{l_n R} \left[I_1(l_n R) K_1(l_n A) - K_1(l_n R) I_1(l_n A) \right] C_n \sin l_n Z, \tag{5.18}$$

where $l_n = \pi n / L$, for some positive integer n, and

$$C_n = \frac{2}{L E_n} \int_0^L \varphi(Z) \sin l_n Z dZ, \qquad n = 1, \ldots, \infty, \tag{5.19}$$

$$E_n = \frac{1}{l_n B} \left[I_1(l_n B) K_1(l_n A) - K_1(l_n B) I_1(l_n A) \right], \qquad n = 1, \ldots, \infty. \tag{5.20}$$

5.1.2. Example 2. The second set of boundary conditions involves our elastic tube bonded between two concentric solid cylinders which do not rotate relative to each other, the lower end of the elastic tube is kept undeformed. These conditions are given by

$$g(A, Z) = 0, \quad g(B, Z) = 0, \quad g(R, 0) = 0. \tag{5.21}$$

In order to maintain the deformation (1.1), an axial load is required on the end of the tube together with a torsional couple, but these expressions will not be needed here. Seeking separable solutions of the governing equation (5.14), the boundary conditions are satisfied with a negative separation constant, $-l^2$, i.e., the axial dependence is taken to vary exponentially. With the same transformation as before, the ODE involving the radial independent variable reduces to the Bessel equation of order one.

Since the configuration here is that of a tube with a finite radius, both solutions of this equation, the Bessel functions of the first kind (J_1) and second kind (Y_1) are involved. The l-mode solution for this problem is now given by

$$g_l(R,Z) = \frac{1}{lR}\left[C_l J_1(lR) + D_l Y_1(lR)\right]\left[A_l e^{lZ} + B_l e^{-lZ}\right], \tag{5.22}$$

where A_l, B_l, C_l, D_l are constants to be determined by the boundary conditions. We note that, for the boundary conditions (5.21), the permissible values of l are to be determined by the Bessel equation as solutions of an eigenvalue problem. The solution of the problem is now given by

$$g(R,Z) = \sum_{n=1}^{\infty} \frac{1}{l_n R}\left[J_1(l_n R)Y_1(l_n A) - Y_1(l_n R)J_1(l_n A)\right]\sinh l_n Z, \tag{5.23}$$

where the l_n are the (positive) roots of the transcendental equation [5]

$$J_1(l_n A)Y_1(l_n B) = J_1(l_n B)Y_1(l_n A). \tag{5.24}$$

5.2. Some further solutions for the generalized azimuthal displacement. In this subsection it would be desirable to compare the form of the general deformation component $g = g(R,Z)$ with the special cases of pure azimuthal shear ($g = g(R)$) and pure torsion ($g = \tau Z$, τ = constant) that have been investigated in the current literature without referring to any choice of boundary conditions. Furthermore, we wish to derive some Riemann type similarity solutions to the governing differential equation (5.14), whose simple form can be used to verify the validity of the necessary and sufficient conditions (4.9) and (4.10). To this end, for the generalized Hadamard material (5.12), we examine the following three cases.

(i) We seek an additively separable solution of (5.14) of the form

$$g(R,Z) = f_1(R) + f_2(Z). \tag{5.25}$$

Substituting this relation into the governing differential equation (5.14), we obtain

$$g(R,Z) = -\frac{2C_1}{R^2} + \frac{lR^2}{8} - \frac{lZ^2}{2} + C_2 Z, \tag{5.26}$$

where C_1, C_2 are integration constants and l is a separation constant. We immediately see that the first term on the left-hand side is a solution of the *pure azimuthal shear* problem for a related material, obtained by Beatty and Jiang [2]. The last term on the right-hand side of the same formula is a deformation known as *pure torsion*, where C_2 is the angle of twist per unit undeformed length. It was shown by Polignone and Horgan [18] that pure torsion is supported for the general Hadamard materials. The simultaneous dependence of the azimuthal displacement on the radial and axial coordinate introduces the coupling constant l and generalizes the previously mentioned individual deformations.

(ii) We seek a Riemann type similarity solution of the form

$$g(R,Z) = f(\xi), \quad \xi = R^n Z^m,\ n, m \text{ being integers.} \tag{5.27}$$

Substitution of this form into (5.14) leads to the values $n = 2, m = -2$ for the two integer powers. Solving the corresponding ODE, the azimuthal displacement obtains the form

$$g(R,Z) = C_1\left(\frac{\sqrt{\xi+1}}{\xi} + \log\frac{1+\sqrt{\xi+1}}{\sqrt{\xi}}\right), \quad \xi = \frac{R^2}{Z^2}, \tag{5.28}$$

where C_1 is an integration constant. Therefore, if suitable boundary conditions can be chosen, the displacement profile will remain constant on the cones $R/Z = \text{constant}$.

(iii) The paragraph above suggests the existence of solutions in the form

$$g(R, Z) = f(\zeta), \quad \zeta = \alpha R^2 + \beta Z^2, \ \alpha, \beta \text{ are constants.} \tag{5.29}$$

Substitution into (5.14) leads to $\alpha = \beta = 1$. The corresponding solution for the azimuthal displacement is then given in the form

$$g(R, Z) = C_1 \zeta^{-\frac{3}{2}}, \quad \zeta = R^2 + Z^2, \tag{5.30}$$

where C_1 is an integration constant. Again, if suitable boundary conditions can be employed, the azimuthal profile remains constant along the spherical surfaces of revolution $R^2 + Z^2 = \text{constant}$. The simple form of this solution can be used to verify the validity of the necessary and sufficient conditions in a straightforward manner. It should be noted here that Hill [9] has provided examples of similarity solutions to a variety of deformations for incompressible elastic materials, a subject closely related to the paragraphs (ii) and (iii) above.

5.3. The effect of torsion on the pure azimuthal shear. In this section we consider the azimuthal displacement with the predetermined form

$$g(R, Z) = \rho(R) + \tau Z, \quad \tau = \text{constant}, \tag{5.31}$$

which will account for the simultaneous azimuthal shearing and torsion and will enable us to compare the unknown function $\rho(R)$, with results from the pure azimuthal deformation $g = g(R)$ in [20, 2]. We start by noting that the strain invariants depend only on R,

$$I = I_1 = I_2 = (R\rho'(R))^2 + (R\tau)^2 + 3; \tag{5.32}$$

hence the azimuthal equilibrium equation (5.3) can be integrated once to

$$4R^3 \rho'(R) \bar{W}_1 = B^2 \bar{\sigma}_\theta, \tag{5.33}$$

where the constant $\bar{\sigma}_\theta$ is the value of the azimuthal shear stress $\sigma_{r\theta}$ on the outer boundary of the cylinder $R = B$. Expression (5.33) is valid for both pure shear and azimuthal shear combined with pure torsion. However, in the case of pure shear, the strain invariants are given by

$$I = I_1 = I_2 = (R\rho'(R))^2 + 3. \tag{5.34}$$

We now consider the following special case of the material (4.15).

(i) $k = 1$. The strain energy function in this case is the generalized Hadamard material (5.12). For the deformation of shear coupled with torsion (5.33) can be integrated once to give

$$\rho(R) = -\frac{B^2 \bar{\sigma}_\theta}{2\mu} \frac{1}{R^2} + C, \tag{5.35}$$

where C is an integration constant and we note that this solution can be obtained directly from the results of the previous section with the value of the separation constant set equal to zero. It can be shown that this is the solution of the pure

azimuthal shear problem for the Hadamard material (5.12) in accordance with the work of Beatty and Jiang [2]. As in the incompressible case [22] for the neo-Hookean material, we deduce that torsion does not affect the shearing for the material under study.

(ii) $k = 1/2$. In this case the strain energy function obtains the form

$$\bar{W}(I_1, I_2, I_3) = \frac{\sqrt{3}\mu}{2}\left[\sqrt{I_1}h_1(I_3) + \sqrt{I_2}h_2(I_3)\right] + h_3(I_3), \tag{5.36}$$

where

$$h_3(1) = -3\mu, \quad h_3'(1) = -\frac{3\mu}{4}. \tag{5.37}$$

Because of the inclusion of the square root of the first principal invariant, this model is reminiscent of the recently introduced strain energy by Bischoff, Arruda, and Grosh [4] for the accurate description of the nonlinear pressure-volume response of rubber-like solids in hydrostatic compression. For shear with torsion substituting (5.36) into (5.33) and considering the invariants (5.32), the function $\rho(R)$ is given in terms of elliptic function in principle. However, requiring the stress value σ_θ and the twist τ to satisfy the relation

$$\frac{\sqrt{3}\mu}{B^2\bar{\sigma}_\theta} = \frac{\tau^2}{3}, \tag{5.38}$$

we obtain the following form for the azimuthal function:

$$\rho(R) = -\sqrt{3}\arcsin\left(\frac{\sqrt{3}}{R\tau}\right) + C, \tag{5.39}$$

where C is an integration constant and $R\tau \geq \sqrt{3}$.

For pure azimuthal shear, substituting (5.36) into (5.33) and taking into account the invariants (5.34), we obtain the solution (ρ_p for pure azimuthal shear),

$$\rho_p(R) = \frac{\sqrt{3}}{2}\arcsin\left(\frac{B^2\bar{\sigma}_\theta}{\sqrt{3}\mu R^2}\right) + C. \tag{5.40}$$

We notice that the different structure between (5.39) and (5.40) is due to the presence of the pure torsion term and is quite profound even for the special case under consideration. Similar conclusions were drawn in [22], for the power neo-Hookean material and the value of the power $n = 1/2$.

6. Concluding remarks. In this article we derived a set of necessary and sufficient conditions (4.1), a compressible and isotropic strain energy has to satisfy, in order to support the generalized azimuthal deformation (1.1). This is a deformation that locally reduces to simple shear. The same conditions remain sufficient for a strain energy to support the special cases of pure azimuthal shear and pure torsion. However, a close examination of the current literature reveals that the same conditions are also sufficient for a strain energy to support antiplane shear deformations [17], helical shear deformations [15], and, of course, their special cases such as axisymmetric antiplane shear. Therefore, if a strain energy function satisfies these conditions, the material is able to support a large class of deformations that locally reduce to simple

shear. It remains open whether such materials will support *all* isochoric deformations that locally reduce to simple shear.

Appendix. Derivatives of Eulerian strain-space quantities. Derivatives of the Eulerian strain-space coordinates and basis vectors in terms of Lagrangian strain-space coordinates are as follows:

$$\frac{\partial\theta}{\partial\Gamma} = \Gamma G_{,\Gamma\Gamma} \equiv \tilde{A}, \quad \frac{\partial\theta}{\partial\Phi} = \Gamma^2 \left(\frac{1}{\Gamma} G_{\Phi}\right)_{,\Gamma} \equiv \tilde{B}, \tag{A.1}$$

$$\frac{\partial \boldsymbol{e}_r}{\partial\Gamma} = \tilde{A}\boldsymbol{e}_\theta, \quad \frac{\partial \boldsymbol{e}_r}{\partial\Phi} = \tilde{B}\boldsymbol{e}_\theta, \tag{A.2}$$

$$\frac{\partial \boldsymbol{e}_\theta}{\partial\Gamma} = -\tilde{A}\boldsymbol{e}_r, \quad \frac{\partial \boldsymbol{e}_\theta}{\partial\Phi} = -\tilde{B}\boldsymbol{e}_r, \tag{A.3}$$

$$\frac{\partial \boldsymbol{e}_\gamma}{\partial\Gamma} = \sin\Phi\tilde{A}\boldsymbol{e}_\theta, \quad \frac{\partial \boldsymbol{e}_\gamma}{\partial\Phi} = \sin\Phi\tilde{B}\boldsymbol{e}_\theta + \boldsymbol{e}_\phi, \tag{A.4}$$

$$\frac{\partial \boldsymbol{e}_\phi}{\partial\Gamma} = \cos\Phi\tilde{A}\boldsymbol{e}_\theta, \quad \frac{\partial \boldsymbol{e}_\phi}{\partial\Phi} = \cos\Phi\tilde{B}\boldsymbol{e}_\theta - \boldsymbol{e}_\gamma. \tag{A.5}$$

Acknowledgments. The authors would like to thank the reviewers for their valuable comments. In addition, the authors are grateful to Professor T. J. Pence at Michigan State University, who invited the first author and provided the facilities during his visit that made this collaboration possible.

REFERENCES

[1] V. Agarwal, *On finite anti-plane shear for compressible elastic circular tube*, J. Elasticity, 9 (1979), pp. 311–319.

[2] M. F. Beatty and Q. Jiang, *On compressible materials capable of sustaining axisymmetric shear deformations.* II. *Rotational shear of isotropic hyperelastic materials*, Quart. J. Mech. Appl. Math., 50 (1997), pp. 211–237.

[3] M. F. Beatty and Q. Jiang, *On compressible materials capable of sustaining axisymmetric shear deformations.* III. *Helical shear of isotropic hyperelastic materials*, Quart. Appl. Math., 57 (1999), pp. 681–697.

[4] J. E. Bischoff, E. M. Arruda, and K. Grosh, *A new constitutive model for the compressibility of elastomers at finite deformations*, Rubber Chem. Technol., 74 (2001), pp. 541–559.

[5] E. Butkov, *Mathematical Physics*, Series in Advanced Physics, Addison-Wesley, Reading, MA, 1968.

[6] M. M. Carroll, *Finite strain solutions in compressible isotropic elasticity*, J. Elasticity, 20 (1988), pp. 65–92.

[7] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. II, *Partial Differential Equations*, Wiley-Interscience, New York, London, 1962.

[8] J. L. Ericksen, *Deformations possible in every compressible, isotropic, perfectly elastic material*, J. Math. Phys., 34 (1955), pp. 126–128.

[9] J. M. Hill, *Exact integrals and solutions for finite deformations of the incompressible Varga elastic material*, in Nonlinear Elasticity: Theory and Applications, R. W. Ogden and Y. B. Fu, eds., London Math. Soc. Lecture Note Ser. 283, Cambridge University Press, Cambridge, UK, 2001.

[10] C. O. Horgan, *Anti-plane shear deformations in linear and nonlinear solid mechanics*, SIAM Rev., 37 (1995), pp. 53–81.

[11] C. O. Horgan, *On axisymmetric solutions for compressible nonlinearly elastic solids*, Z. Angew. Math. Phys., 46 (1995), pp. S107–S125.

[12] C. O. Horgan, *Equilibrium Solutions for Compressible Nonlinear Elasticity*, in Nonlinear Elasticity: Theory and Applications, R. W. Ogden and Y. B. Fu, eds., London Math. Soc. 283, Lecture Note Ser., Cambridge University Press, Cambridge, UK, 2001.

[13] C. O. Horgan and G. Saccomandi, *Pure azimuthal shear of isotropic, incompressible hyperelastic materials with limiting chain extensibility*, Internat. J. Nonlinear Mechanics, 36 (2001), pp. 465–475.

[14] Q. Jiang and M. F. Beatty, *On compressible materials capable of sustaining axisymmetric shear deformations*. I. *Anti-plane shear of isotropic hyperelastic materials*, J. Elasticity, 39 (1995), pp. 75–95.

[15] E. Kirkinis and R. W. Ogden, *On helical shear of a compressible elastic circular cylindrical tube*, Quart. J. Mech. Appl. Math., 56 (2003), pp. 105–122.

[16] J. K. Knowles, *On finite anti-plane shear for incompressible elastic materials*, J. Austral. Math. Soc. Ser. B, 19 (1975/76), pp. 400–415.

[17] J. K. Knowles, *A note on anti-plane shear for compressible materials in finite elastostatics*, J. Austral. Math. Soc. Ser. B, 20 (1977/78), pp. 1–7.

[18] D. Polignone and C. O. Horgan, *Pure torsion of compressible nonlinearly elastic circular cylinders*, Quart. Appl. Math., 49 (1991), pp. 591–607.

[19] D. Polignone and C. O. Horgan, *Axisymmetric finite anti-plane shear of compressible nonlinearly elastic circular tubes*, Quart. Appl. Math., 50 (1992), pp. 323–341.

[20] D. Polignone and C. O. Horgan, *Pure azimuthal shear of compressible nonlinearly elastic circular tubes*, Quart. Appl. Math., 52 (1994), pp. 113–131.

[21] D. Polignone Warne and P. G. Warne, *Plane deformations in incompressible nonlinear elasticity*, J. Elasticity, 52 (1999), pp. 129–158.

[22] L. Tao, K. R. Rajagopal, and A. S. Wineman, *Circular shearing and torsion of generalized neo-Hookean materials*, IMA J. Appl. Math., 48 (1992), pp. 23–37.

[23] H. Tsai and X. Fan, *Anti-plane shear deformations in compressible transversely isotropic materials*, J. Elasticity, 54 (1999), pp. 73–88.

SIAM J. APPL. MATH.
Vol. 65, No. 4, pp. 1101–1129

MODELING OF WAVE RESONANCES IN LOW-CONTRAST PHOTONIC CRYSTALS*

DMITRI AGUEEV† AND DMITRY PELINOVSKY†

Abstract. Coupled-mode equations are derived from Maxwell equations for modeling of low-contrast cubic-lattice photonic crystals in three spatial dimensions. Coupled-mode equations describe resonantly interacting Bloch waves in stop bands of the photonic crystal. We study the linear boundary-value problem for stationary transmission of four counter-propagating and two oblique waves on the plane. Well-posedness of the boundary-value problem is proved by using the method of separation of variables and generalized Fourier series. For applications in photonic optics, we compute integral invariants for transmission, reflection, and diffraction of resonant waves.

Key words. photonic crystals, coupled-mode equations, wave resonances, stationary transmission boundary-value problems

AMS subject classifications. 35P10, 35P20, 35Q60, 78M35

DOI. 10.1137/040606053

1. Introduction.

Photonic band-gap crystals are periodic optical materials, the spectrum of which consists of bands separated by band gaps [13]. Linear periodic properties of the isotropic photonic crystals are modeled with the Maxwell equations

$$\nabla^2 \mathbf{E} - \frac{n^2}{c^2}\frac{\partial^2 \mathbf{E}}{\partial t^2} = \nabla(\nabla \cdot \mathbf{E}), \qquad \nabla \cdot \left(n^2 \mathbf{E}\right) = 0, \tag{1.1}$$

where $n = n(\mathbf{x})$ is the periodic refractive index, $\mathbf{E} = (E_x, E_y, E_z)$ is the electric field vector, $\mathbf{x} = (x, y, z)$ is the physical space, t is the time variable, $\nabla = (\partial_x, \partial_y, \partial_z)$ is the gradient vector, and c is the speed of light. Components of the magnetic field vector are eliminated from the Maxwell equations (1.1) [13].

The Maxwell equations (1.1) in one dimension can be simplified for a linearly polarized light, such that $\mathbf{E} = (E, 0, 0)$, where $E = E(z, t)$ and $n = n(z)$. The scalar component $E(z, t)$ solves the wave equation with the periodic speed variations

$$\frac{\partial^2 E}{\partial z^2} - \frac{n^2(z)}{c^2}\frac{\partial^2 E}{\partial t^2} = 0. \tag{1.2}$$

If the refractive index $n(z)$ is a periodic function with period z_0, the linear spectrum of the wave equation (1.2) reduces to the Mathieu equation for $E(z, t) = \psi(z)e^{-i\omega t}$, where ω is the eigenvalue and $\psi(z)$ is the eigenfunction of the spectral problem

$$\psi'' + \frac{\omega^2}{c^2} n^2(z)\psi = 0. \tag{1.3}$$

According to the Floquet theory [12], solutions of the Mathieu equation (1.3) take the form $\psi(z) = \Psi(z)e^{ik(\omega)z}$, where $\Psi(z + z_0) = \Psi(z)$ and $k = k(\omega)$ is the propagation constant. For a general class of periodic potentials $n^2(z)$, there exist infinitely many

*Received by the editors March 31, 2004; accepted for publication (in revised form) August 20, 2004; published electronically April 14, 2005.
http://www.siam.org/journals/siap/65-4/60605.html
†Department of Mathematics, McMaster University, 1280 Main Street West, Hamilton, ON, Canada, L8S 4K1 (agueevd@univmail.cis.mcmaster.ca, dmpeli@math.mcmaster.ca).

intervals of ω, called *band gaps*, where the propagation constant $k(\omega)$ is purely imaginary and the Bloch function $\psi(z)$ is unbounded in z. The band gaps are supported by the low-contrast photonic crystal with the refractive index $n(z) = n_0 + \epsilon n_1(z)$, where n_0 is constant and ϵ is small parameter.

The linear Maxwell equations (1.1) in two and three dimensions can also be reduced to a spectral problem for $\mathbf{E}(\mathbf{x}, t) = \boldsymbol{\psi}(\mathbf{x})e^{-i\omega t}$, where ω is the eigenvalue and $\boldsymbol{\psi}(\mathbf{x})$ is the eigenvector. When $n(\mathbf{x})$ is a periodic function in x, y, z with periods x_0, y_0, z_0, respectively, the eigenvector $\boldsymbol{\psi}(\mathbf{x})$ satisfies the Floquet theorem [12] and has the form of the Bloch wave: $\boldsymbol{\psi}(\mathbf{x}) = \boldsymbol{\Psi}(\mathbf{x})e^{i(k_x x + k_y y + k_z z)}$, where $\boldsymbol{\Psi}(\mathbf{x})$ is periodic in x, y, and z with periods x_0, y_0, and z_0, and $\omega = \omega(k_x, k_y, k_z)$. No band gaps exist in the linear spectrum for low-contrast photonic crystals. As a result, the bounded Bloch functions $\boldsymbol{\psi}(\mathbf{x})$ may exist for any value of $\omega \in \mathbb{R}$. High-contrast photonic crystals may, however, exhibit band gaps for some configurations of the refractive index $n(\mathbf{x})$ [13].

Modeling of time-dependent responses of photonic crystals in three spatial dimensions can be computationally difficult in the framework of the Maxwell equations, especially if the nonlinear and nonlocal dispersive terms are taken into account. A more efficient method is based on reduction of Maxwell equations (1.1) to the coupled-mode equations [23]. For instance, shock wave singularities may occur in the nonlinear Maxwell equations but they do not occur in the nonlinear coupled-mode equations [8]. Coupled-mode equations are typically derived in the first band gap of the Bragg resonance between two counter-propagating waves in one spatial dimension [20, 21]. More complicated coupled-mode equations are considered for three-dimensional nonlinear photonic crystals [1, 2, 3, 6]. Recent reviews [4, 5] also include classification of different resonances of Bloch waves in photonic crystals with quadratic nonlinearities.

In this paper, we classify wave resonances and coupled-mode equations for low-contrast cubic-lattice photonic crystals in three spatial dimensions. Since low-contrast crystals do not support band gaps beyond one dimension [12, 13], resonances are considered in stop bands of the linear spectrum [10]. Stop bands occur between resonant counter-propagating waves, which could be coupled resonantly with other oblique Bloch waves. The number of resonant Bloch waves depends on the geometric configuration of the incident wave with respect to the cubic lattice. When the Maxwell equations are truncated with the perturbation series expansions, coupled-mode equations for the lowest-order Bragg resonances are derived and studied in bounded domains, subject to the radiation boundary conditions. The radiation boundary conditions describe transmission of the incident Bloch waves which generate resonantly reflected and diffracted Bloch waves in the photonic crystals.

We study here the linear coupled-mode equations for four counter-propagating and two oblique Bloch waves on the plane. It is not a priori clear why the stationary boundary-value problem with radiation boundary conditions is well posed, since it is specified by non–self-adjoint operators on the bounded domains. We prove, however, the well-posedness of the linear stationary problem by using separation of variables and generalized Fourier series [24]. Eigenfunction expansions and convergence of generalized Fourier series follow from the general theory [7]. As a result, we construct explicit analytical expressions for stationary transmission, reflection, and diffraction of resonant Bloch waves, which are used in modeling of the low-contrast photonic crystals.

Other applications of optical photonic structures include nonlinear phenomena, such as bistable stationary transmission and gap soliton propagation [6, 14, 15, 22]. Very little is known about the persistence of such phenomena in two and three spatial

dimensions, especially given that no band gap exists in low-contrast three-dimensional photonic structures. The coupled-mode equations can be generalized to include the weakly nonlinear (cubic) terms and to extend the time-dependent problems to the nonlinear coupled-mode equations [18, 19]. Well-posedness of the nonlinear stationary problems is beyond the scope of this manuscript, which only presents solutions of the linear stationary problems. Nevertheless, linear analysis opens the road to nonlinear analysis of the corresponding boundary-value problems.

The paper is organized as follows. Classification of resonances in low-contrast cubic-lattice crystals is given in section 2. Derivation of coupled-mode equations for lowest-order resonances is described in section 3. The linear stationary boundary-value problems for four counter-propagating and two oblique resonant Bloch waves are analyzed in section 4. Section 5 concludes the paper. Appendix A gives derivation and explicit forms of the nonlinear coupled-mode equations with cubic (Kerr) nonlinearities.

2. Classification of resonances. When the optical material is homogeneous, such that $n(\mathbf{x}) = n_0$ is constant, the linear spectrum of the Maxwell equations (1.1) is defined by the free transverse waves,

$$\mathbf{E}(\mathbf{x}, t) = \mathbf{e}_{\mathbf{k}} e^{i(\mathbf{k}\cdot\mathbf{x} - \omega t)}, \tag{2.1}$$

where $\mathbf{e}_{\mathbf{k}}$ is the polarization vector, $\mathbf{k} = (k_x, k_y, k_z)$ is the wave vector, and $\omega = \omega(\mathbf{k})$ is the wave frequency. It follows from system (1.1) that

$$\mathbf{k} \cdot \mathbf{e}_{\mathbf{k}} = 0, \qquad \omega^2 = \frac{c^2}{n_0^2}(k_x^2 + k_y^2 + k_z^2). \tag{2.2}$$

For each wave vector $\mathbf{k}$ there exist two independent polarizations $\mathbf{e}_{\mathbf{k}}^{(1)}$ and $\mathbf{e}_{\mathbf{k}}^{(2)}$ such that $\mathbf{e}_{\mathbf{k}}^{(1)} \cdot \mathbf{e}_{\mathbf{k}}^{(2)} = 0$. This degeneracy in the polarization vector is neglected here by the assumption that the incident wave is linearly polarized.

When the optical material is periodic such that $n(\mathbf{x} + \mathbf{x}_0) = n(\mathbf{x}_0)$, the linear spectrum of the Maxwell equations (1.1) is defined by the Bloch waves:

$$\mathbf{E}(\mathbf{x}, t) = \mathbf{\Psi}(\mathbf{x}) e^{i(\mathbf{k}\cdot\mathbf{x} - \omega t)}, \tag{2.3}$$

where $\mathbf{\Psi}(\mathbf{x} + \mathbf{x}_0) = \mathbf{\Psi}(\mathbf{x})$ is the periodic envelope, $\mathbf{k} = (k_x, k_y, k_z)$ is the wave vector, and $\omega = \omega(\mathbf{k})$ is the wave frequency. Existence of the Bloch waves (2.3) for the Maxwell equations (1.1) is proved in [12]. The geometric configuration of the photonic crystal is defined by the fundamental (linearly independent) lattice vectors $\mathbf{x}_{1,2,3}$ and fundamental reciprocal lattice vectors $\mathbf{k}_{1,2,3}$ such that $\mathbf{k}_i \cdot \mathbf{x}_j = 2\pi\delta_{i,j}$, where $1 \le i, j \le 3$ (see [10]). Therefore, the linear refractive index $n(\mathbf{x})$ can be expanded into a triple Fourier series:

$$n(\mathbf{x}) = n_0 \sum_{(n,m,l)\in\mathbb{Z}^3} \alpha_{n,m,l} e^{i(n\mathbf{k}_1 + m\mathbf{k}_2 + l\mathbf{k}_3)\cdot\mathbf{x}}, \tag{2.4}$$

where the factor n_0 is included for convenience. If n_0 is the mean value of $n(\mathbf{x})$, then $\alpha_{0,0,0} = 1$. Let the wave vector $\mathbf{k}$ in the incident Bloch wave (2.3) be chosen as $\mathbf{k} = \mathbf{k}_{\text{in}}$. The incident wave vector $\mathbf{k}_{\text{in}}$ is expanded in terms of the lattice vectors:

$$\mathbf{k}_{\text{in}} = \frac{1}{2}(p\mathbf{k}_1 + q\mathbf{k}_2 + r\mathbf{k}_3), \tag{2.5}$$

where $(p,q,r) \in \mathbb{R}^3$ are parameters. The Bloch wave (2.3) is represented by triple Fourier series for $\mathbf{\Psi}(\mathbf{x})$, such that $\mathbf{E}(\mathbf{x},t)$ consists of an infinite superposition of free transverse waves with the wave vectors $\mathbf{k}_{\rm out}^{(n,m,l)}$:

$$\mathbf{k}_{\rm out}^{(n,m,l)} = \mathbf{k}_{\rm in} + n\mathbf{k}_1 + m\mathbf{k}_2 + l\mathbf{k}_3, \quad (n,m,l) \in \mathbb{Z}^3. \tag{2.6}$$

The wave vector $\mathbf{k}_{\rm out}^{(n,m,l)}$ with a nonempty triple (n,m,l) is said to be *resonant* with the wave vector $\mathbf{k}_{\rm in}$ if $|\mathbf{k}_{\rm out}^{(n,m,l)}| = |\mathbf{k}_{\rm in}|$ such that $|\omega(\mathbf{k}_{\rm out}^{(n,m,l)})| = |\omega(\mathbf{k}_{\rm in})|$.

We consider here a simple cubic crystal, where the fundamental lattice vectors and reciprocal lattice vectors are all orthogonal [10]:

$$\mathbf{x}_{1,2,3} = a\mathbf{e}_{1,2,3}, \qquad \mathbf{k}_{1,2,3} = k_0\mathbf{e}_{1,2,3}, \quad k_0 = \frac{2\pi}{a}, \tag{2.7}$$

where $\mathbf{e}_{1,2,3}$ are unit vectors in $\mathbb{R}^3$. The coordinate axes (x,y,z) are oriented along the axes of the simple cubic crystal, while the incident wave vector $\mathbf{k}_{\rm in}$ is directed according to the spherical angles (θ,φ) as follows:

$$\mathbf{k}_{\rm in} = k(\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta), \quad k \in \mathbb{R},\ 0 \le \theta \le \pi,\ 0 \le \varphi \le 2\pi, \tag{2.8}$$

where $k = |\mathbf{k}_{\rm in}|$. When $\theta = 0$, the wave vector $\mathbf{k}_{\rm in}$ is perpendicular to the (x,y) crystal plane. For the simple cubic crystal, the set of resonant Bloch waves is given by the set of triples

$$\mathcal{S} = \{(n,m,l) \in \mathbb{Z}^3 : n(n+p) + m(m+q) + l(l+r) = 0\}, \tag{2.9}$$

where

$$p = \frac{2k}{k_0}\sin\theta\cos\varphi, \qquad q = \frac{2k}{k_0}\sin\theta\sin\varphi, \qquad r = \frac{2k}{k_0}\cos\theta. \tag{2.10}$$

The set $\mathcal{S}$ always has a zero solution: $(n,m,l) = (0,0,0)$. When $(p,q,r) \in \mathbb{Z}^3$ and $|p|+|q|+|r| \neq 0$, the set $\mathcal{S}$ has at least one nonzero solution: $(n,m,l) = (-p,-q,-r)$. The set $\mathcal{S}$ is also bounded, since (n,m,l) are integer solutions inside the sphere:

$$\left(n + \frac{p}{2}\right)^2 + \left(m + \frac{q}{2}\right)^2 + \left(l + \frac{r}{2}\right)^2 = \left(\frac{k}{k_0}\right)^2 < \infty. \tag{2.11}$$

When $(p,q,r) \in \mathbb{Z}^3$, resonant triples (n,m,l) can all be classified analytically. However, when $(p,q,r) \notin \mathbb{Z}^3$, additional resonant triples may also exist. In solid state physics [10], a geometric solution for the resonant triples (n,m,l) is constructed from the condition that the vector $\mathbf{G}^{(n,m,l)} = \mathbf{k}_{\rm out}^{(n,m,l)} - \mathbf{k}_{\rm in}$ lies on the edge of sectors of the reciprocal lattice. Here we review particular resonant sets $\mathcal{S}$ for integer and noninteger values of (p,q,r).

2.1. A family of one-dimensional resonances. The one-dimensional Bragg resonance occurs when the incident wave is coupled with the counter-propagating reflected wave such that the set $\mathcal{S}$ has at least one nonzero solution: $(n,m,l) = (0,0,-r)$, where $r \in \mathbb{Z}_+$. The values of p and q are not defined for the Bragg resonance when $n = m = 0$. As a result, spherical angles θ and φ in the parametrization (2.8) are arbitrary, while the wave number k satisfies the Bragg resonance condition [10]:

$$rk_0 = 2k\cos\theta \tag{2.12}$$

such that $r\lambda = 2a\cos\theta$, where λ is the wavelength. The one-dimensional Bragg resonance is generalized in three dimensions for $p = q = 0$ and $r \in \mathbb{Z}_+$, when the geometric configuration for the Bragg resonance (2.12) is fixed at the specific value $\theta = 0$, and

$$\mathbf{k}_{\text{in}} = \frac{\pi}{a}(0,0,r), \qquad \mathbf{k}_{\text{out}}^{(0,0,-r)} = \frac{\pi}{a}(0,0,-r). \tag{2.13}$$

The incident wave is directed to the z-axis of the cubic lattice crystal, and the wavelength is $\lambda = 2a/r$. The family of Bragg resonances with $p = q = 0$ and $r \in \mathbb{Z}_+$ may include not only the two counter-propagating waves (2.13) but also other Bloch waves in three-dimensional photonic crystals. The lowest-order resonant sets $\mathcal{S}$ for $p = q = 0$ and $r \in \mathbb{Z}_+$ are listed below:

$$\begin{aligned}
&r = 1 : \mathcal{S} = \{(0,0,0),(0,0,-1)\},\\
&r = 2 : \mathcal{S} = \{(0,0,0),(1,0,-1),(-1,0,-1),(0,1,-1),(0,-1,-1),(0,0,-2)\},\\
&r = 3 : \mathcal{S} = \{(0,0,0),(1,1,-1),(-1,1,-1),(1,-1,-1),(-1,-1,-1),\\
&\qquad\qquad (1,1,-2),(-1,1,-2),(1,-1,-2),(-1,-1,-2),(0,0,-3)\}.
\end{aligned}$$

The dimension of $\mathcal{S}$ depends on the total number of all possible integer solutions for (n,m,l). The sets $\mathcal{S}$ for higher-order resonances with $r \in \mathbb{Z}_+$ can be found algorithmically, with symbolic computing software.

2.2. A family of two-dimensional resonances. Two-dimensional Bragg resonances occur when the incident wave vector $\mathbf{k}_{\text{in}}$ is resonant to the counter-propagating reflected wave vector $\mathbf{k}_{\text{out}}^{(-p,-q,0)}$, as well as to two other diffracted wave vectors $\mathbf{k}_{\text{out}}^{(0,-q,0)}$ and $\mathbf{k}_{\text{out}}^{(-p,0,0)}$, where $(p,q) \in \mathbb{Z}_+^2$. The value of r is not defined for the two-dimensional resonance, such that the angle θ in the parametrization (2.8) is arbitrary, while k and φ satisfy the resonance conditions

$$\varphi = \arctan\left(\frac{q}{p}\right), \qquad \sqrt{p^2+q^2}k_0 = 2k\sin\theta. \tag{2.14}$$

The two-dimensional Bragg resonances are generalized in three dimensions for $(p,q) \in \mathbb{Z}_+^2$ and $r = 0$, when the geometric configuration for the Bragg resonance (2.14) is fixed at the specific value $\theta = \frac{\pi}{2}$, and

$$\begin{aligned}
\mathbf{k}_{\text{in}} &= \frac{\pi}{a}(p,q,0), \qquad \mathbf{k}_{\text{out}}^{(-p,-q,0)} = \frac{\pi}{a}(-p,-q,0),\\
\mathbf{k}_{\text{out}}^{(0,-q,0)} &= \frac{\pi}{a}(p,-q,0), \qquad \mathbf{k}_{\text{out}}^{(-p,0,0)} = \frac{\pi}{a}(-p,q,0).
\end{aligned} \tag{2.15}$$

The incident wave $\mathbf{k}_{\text{in}}$ is directed along the diagonal of the (px,qy)-cell of the cubic lattice crystal, and the wavelength is $\lambda = 2a/\sqrt{p^2+q^2}$.

The families of Bragg resonances with $(p,q) \in \mathbb{Z}_+^2$ and $r = 0$ may include not only the four resonant waves (2.15) but also other Bloch waves in three-dimensional photonic crystals. The lowest-order resonant sets $\mathcal{S}$ for $(p,q) \in \mathbb{Z}_+^2$ and $r = 0$ are listed below:

$$\begin{aligned}
&p{=}1,\ q{=}1 : \mathcal{S} = \{(0,0,0),(-1,0,0),(0,-1,0),(-1,-1,0)\},\\
&p{=}2,\ q{=}1 : \mathcal{S} = \{(0,0,0),(0,-1,0),(-1,0,1),(-1,0,-1),(-1,-1,1),(-1,-1,-1),\\
&\qquad\qquad (-2,0,0),(-2,-1,0)\},
\end{aligned}$$

$$p=2,\ q=2 : \mathcal{S} = \{(0,0,0),(0,-1,1),(0,-1,-1),(0,-2,0),(-1,0,1),(-1,0,-1),$$
$$(-1,-2,1),(-1,-2,-1),(-2,0,0),(-2,-1,1),(-2,-1,-1),$$
$$(-2,-2,0)\}.$$

2.3. Two-dimensional resonances of oblique waves. The resonant set $\mathcal{S}$ can be nonempty for $(p,q,r) \notin \mathbb{Z}^3$, which correspond to oblique Bloch waves. For instance, two oblique waves can be resonant on the (x,y)-plane if

$$\mathbf{k}_{\rm in} = \frac{\pi}{a}(p,q,0), \qquad \mathbf{k}_{\rm out}^{(n,m,0)} = \frac{\pi}{a}(p+2n, q+2m, 0), \tag{2.16}$$

where $(n,m) \in \mathbb{Z}^2$ are arbitrary and $(p,q) \in \mathbb{R}^2$ are taken on the straight line:

$$np + mq = -(n^2+m^2). \tag{2.17}$$

Similarly, three oblique waves can be resonant on the (x,y)-plane if

$$\begin{aligned} \mathbf{k}_{\rm in} &= \frac{\pi}{a}(p,q,0), \\ \mathbf{k}_{\rm out}^{(n_1,m_1,0)} &= \frac{\pi}{a}(p+2n_1, q+2m_1, 0), \\ \mathbf{k}_{\rm out}^{(n_2,m_2,0)} &= \frac{\pi}{a}(p+2n_2, q+2m_2, 0), \end{aligned} \tag{2.18}$$

where $(n_1,m_1) \in \mathbb{Z}^2$ and $(n_2,m_2) \in \mathbb{Z}^2$ are arbitrary subject to the constraint $m_1 n_2 \neq m_2 n_1$, while (p,q) take rational values

$$p = \frac{m_1(n_2^2+m_2^2) - m_2(n_1^2+m_1^2)}{m_2 n_1 - m_1 n_2}, \qquad q = \frac{n_1(n_2^2+m_2^2) - n_2(n_1^2+m_1^2)}{n_2 m_1 - n_1 m_2}. \tag{2.19}$$

In the general case, two oblique waves (2.16) or three oblique waves (2.18) may have resonances with other Bloch waves in three-dimensional photonic crystals.

2.4. A family of three-dimensional resonances. When $(p,q,r) \in \mathbb{Z}_+^3$, the resonant sets $\mathcal{S}$ include eight coupled waves for fully three-dimensional Bragg resonance:

$$\begin{aligned} \mathbf{k}_{\rm in} &= \frac{\pi}{a}(p,q,r), \qquad \mathbf{k}_{\rm out}^{(-p,-q,-r)} = \frac{\pi}{a}(-p,-q,-r), \\ \mathbf{k}_{\rm out}^{(-p,0,0)} &= \frac{\pi}{a}(-p,q,r), \qquad \mathbf{k}_{\rm out}^{(0,-q,0)} = \frac{\pi}{a}(p,-q,r), \\ \mathbf{k}_{\rm out}^{(0,0,-r)} &= \frac{\pi}{a}(p,q,-r), \qquad \mathbf{k}_{\rm out}^{(-p,-q,0)} = \frac{\pi}{a}(-p,-q,r), \\ \mathbf{k}_{\rm out}^{(-p,0,-r)} &= \frac{\pi}{a}(-p,q,-r), \qquad \mathbf{k}_{\rm out}^{(0,-q,-r)} = \frac{\pi}{a}(p,-q,-r). \end{aligned} \tag{2.20}$$

The resonance condition for the three-dimensional Bragg resonance takes the form

$$\varphi = \arctan\left(\frac{q}{p}\right), \qquad \theta = \arctan\left(\frac{\sqrt{p^2+q^2}}{r}\right), \qquad \sqrt{p^2+q^2+r^2}k_0 = 2k. \tag{2.21}$$

The incident wave $\mathbf{k}_{\rm in}$ is directed along the diagonal of the (px, qy, rz)-cell of the cubic lattice crystal, and the wavelength is $\lambda = 2a/\sqrt{p^2+q^2+r^2}$. The eight waves (2.20) can be coupled with some other resonant waves such that $\dim(\mathcal{S}) \geq 8$ for $(p,q,r) \in \mathbb{Z}_+^3$. For instance, $\dim(\mathcal{S}) = 8$ for $(p,q,r) = (1,1,1)$ and $(p,q,r) = (2,1,1)$, but $\dim(\mathcal{S}) = 10$ for $(p,q,r) = (2,2,1)$ and $\dim(\mathcal{S}) = 16$ for $(p,q,r) = (3,2,1)$.

3. Derivation of coupled-mode equations. The dispersion surface $\omega = \omega(\mathbf{k})$ for the Bloch waves (2.3) in the periodic photonic crystal is defined by the profile of the refractive index $n(\mathbf{x})$. We shall consider the asymptotic approximation of the dispersion surface $\omega = \omega(\mathbf{k})$ in the limit when the photonic crystal is low-contrast, such that the refractive index $n(\mathbf{x})$ is given by

$$n(\mathbf{x}) = n_0 + \epsilon n_1(\mathbf{x}), \tag{3.1}$$

where n_0 is a constant and ϵ is small parameter. It is proved in [12] that the Bloch waves (2.3) are smooth functions of ϵ, such that the asymptotic solution of the Maxwell equations (1.1) as $\epsilon \to 0$ takes the form of the perturbation series expansions:

$$\mathbf{E}(\mathbf{x}, t) = \mathbf{E}_0(\mathbf{x}, t) + \epsilon \mathbf{E}_1(\mathbf{x}, t) + \mathrm{O}(\epsilon^2). \tag{3.2}$$

The leading-order term $\mathbf{E}_0(\mathbf{x}, t)$ consists of free transverse waves (2.1) with wave vectors $\mathbf{k}_{\text{out}}^{(n,m,l)}$, given by (2.6), such that the asymptotic form (3.2) represents the Bloch wave (2.3) as $\epsilon \neq 0$.

Coupled-mode equations are derived by separating resonant free waves from nonresonant free waves in the Bloch wave (2.3), where the resonant set $\mathcal{S}$ with $N = \dim(\mathcal{S}) < \infty$ is defined by (2.9). Let $\mathbf{E}_0(\mathbf{x}, t)$ be a linear superposition of N resonant waves with wave vectors $\mathbf{k}_j$ at the same frequency ω:

$$\mathbf{E}_0(\mathbf{x}, t) = \sum_{j=1}^{N} A_j(\mathbf{X}, T) \mathbf{e}_{\mathbf{k}_j} e^{i(\mathbf{k}_j \mathbf{x} - \omega t)}, \quad \mathbf{X} = \frac{\epsilon \mathbf{x}}{k}, \; T = \frac{\epsilon t}{\omega}, \tag{3.3}$$

where ω and $\mathbf{k}_j$ are related by the same dispersion equation (2.2), $A_j(\mathbf{X}, T)$ is the envelope amplitude of the jth resonant wave (2.1), and $(\mathbf{X}, T)$ are slow variables. The slow variables represent a deformation of the dispersion surface $\omega = \omega(\mathbf{k}_j)$ due to the low-contrast periodic photonic crystal. The degeneracy in the polarization vector is neglected by the assumption that the incident wave is linearly polarized with the polarization vector $\mathbf{e}_{\text{in}} = \mathbf{e}_{\mathbf{k}_{\text{in}}}$. The triple Fourier series (2.4) for the cubic-lattice crystal (2.7) is simplified as follows:

$$n_1(\mathbf{x}) = n_0 \sum_{(n,m,l) \in \mathbb{Z}^3} \alpha_{n,m,l} e^{ik_0(nx+my+lz)}, \tag{3.4}$$

where $\alpha_{0,0,0} = 0$. The Fourier coefficients $\alpha_{n,m,l}$ satisfy the constraints

$$\alpha_{n,m,l} = \bar{\alpha}_{-n,-m,-l}, \tag{3.5}$$

due to the reality of $n_1(\mathbf{x})$;

$$\alpha_{n,m,l} = \alpha_{m,n,l} = \alpha_{n,l,m} = \alpha_{l,m,n}, \tag{3.6}$$

due to the crystal isotropy in the directions of x, y, z-axes; and

$$\alpha_{-n,m,l} = \alpha_{n,m,l}, \quad \alpha_{n,-m,l} = \alpha_{n,m,l}, \quad \alpha_{n,m,-l} = \alpha_{n,m,l}, \tag{3.7}$$

due to the crystal symmetry with respect to the origin $(0, 0, 0)$. (The latter property can be achieved by a simple shift of (x, y, z).) It follows from constraints (3.5) and (3.7) that all coefficients $\alpha_{n,m,l}$ for $(n, m, l) \in \mathbb{Z}^3$ are real-valued.

It follows from (1.1), (3.1), and (3.2) that the first-order correction term $\mathbf{E}_1(\mathbf{x},t)$ solves the nonhomogeneous linear problem

$$\begin{aligned}\nabla^2\mathbf{E}_1 - \frac{n_0^2}{c^2}\frac{\partial^2\mathbf{E}_1}{\partial t^2} &= 2\frac{n_0^2\omega}{c^2}\frac{\partial^2\mathbf{E}_0}{\partial T\partial t} - 2k\left(\nabla\cdot\nabla_X\right)\mathbf{E}_0 \\ &\quad + \frac{2n_0n_1(\mathbf{x})}{c^2}\frac{\partial^2\mathbf{E}_0}{\partial t^2} + \frac{2}{n_0}\nabla(\nabla n_1\cdot\mathbf{E}_0),\end{aligned} \tag{3.8}$$

where $\nabla_X = (\partial_X,\partial_Y,\partial_Z)$ and the second equation of (1.1) has been used. The right-hand side of the nonhomogeneous equation (3.8) has resonant terms, which are parallel to the free-wave resonant solutions of the homogeneous problem. The resonant terms lead to the secular growth of $\mathbf{E}_1(\mathbf{x},t)$ in t unless they are identically zero. The latter conditions define the coupled-mode equations for amplitudes $A_j(\mathbf{X},T)$, $j=1,\ldots,N$, in the general form

$$i\left(\frac{\partial A_j}{\partial T} + \left(\frac{\mathbf{k}_j}{k}\cdot\nabla_X\right)A_j\right) + \sum_{k\neq j}\hat{\alpha}_{j,k}A_k = 0, \quad j=1,\ldots,N, \tag{3.9}$$

where the elements $\{\hat{\alpha}_{j,k}\}_{1\leq j,k\leq N}$ are related to the Fourier coefficients of the resonant waves $\{\alpha_{n,m,l}\}_{(n,m,l)\in\mathcal{S}}$. The explicit forms of the coupled-mode equations (3.9) are given for two and four counter-propagating and two oblique resonant Bloch waves.

3.1. Coupled-mode equations for two counter-propagating waves. The lowest-order Bragg resonance for two counter-propagating waves (2.13) occurs for $r=1$, when

$$\mathbf{k}_1 = \frac{\pi}{a}(0,0,1), \qquad \mathbf{k}_2 = \frac{\pi}{a}(0,0,-1). \tag{3.10}$$

Let $A_1 = A_+(Z,T)$ and $A_2 = A_-(Z,T)$ be the amplitudes of the right (forward) and left (backward) propagating waves, respectively. The envelope amplitudes are not modulated across the (X,Y)-plane, since the coupled-mode equations for $A_\pm$ are essentially one-dimensional. The polarization vectors are chosen in the x-direction such that $\mathbf{e}_{\mathbf{k}_1} = \mathbf{e}_{\mathbf{k}_2} = (1,0,0)$ and $\mathbf{E}_0 = (E_{0,x}(z,Z,T)e^{-i\omega t},0,0)$. The nonhomogeneous equation (3.8) at the x-component of the solution $\mathbf{E}_1$ at $e^{-i\omega t}$ takes the form

$$\begin{aligned}\nabla^2E_{1,x} + k^2E_{1,x} &= -2ik^2\frac{\partial}{\partial T}E_{0,x} - 2k\frac{\partial^2}{\partial Z\partial z}E_{0,x} \\ &\quad - \frac{2k^2n_1(\mathbf{x})}{n_0}E_{0,x} + \frac{2}{n_0}\frac{\partial^2n_1(\mathbf{x})}{\partial x^2}E_{0,x}.\end{aligned} \tag{3.11}$$

By removing the resonant terms at $e^{\pm ikz}$, the coupled-mode equations for amplitudes $A_\pm(Z,T)$ take the form

$$i\left(\frac{\partial A_+}{\partial T} + \frac{\partial A_+}{\partial Z}\right) + \alpha A_- = 0, \tag{3.12}$$

$$i\left(\frac{\partial A_-}{\partial T} - \frac{\partial A_-}{\partial Z}\right) + \alpha A_+ = 0, \tag{3.13}$$

where $\alpha = \alpha_{0,0,1} = \alpha_{0,0,-1}$. The coupled-mode equations (3.12)–(3.13) can be defined on the interval $0\leq Z\leq L_z$ for $T\geq 0$, where the end points at $Z=0$ and $Z=L_z$ are the left and right (x,y)-planes, which cut a slice of the photonic crystal. The linear system (3.12)–(3.13) is reviewed in [23]. The nonlinear coupled-mode equations are derived in [6, 22] and analyzed recently in [8, 14, 15].

3.2. Coupled-mode equations for four counter-propagating waves. The lowest-order resonance for four counter-propagating waves (2.15) occurs for $p = q = 1$, when

$$\mathbf{k}_1 = \frac{\pi}{a}(1,1,0), \quad \mathbf{k}_2 = \frac{\pi}{a}(1,-1,0), \quad \mathbf{k}_3 = \frac{\pi}{a}(-1,1,0), \quad \mathbf{k}_4 = \frac{\pi}{a}(-1,-1,0). \tag{3.14}$$

Let $A_1 = A_+(X,Y,T)$ and $A_4 = A_-(X,Y,T)$ be the amplitudes of the counter-propagating waves along the main diagonal of the (x,y) plane, while $A_2 = B_+(X,Y,T)$ and $A_3 = B_-(X,Y,T)$ are the amplitudes of the counter-propagating waves along the antidiagonal of the (x,y)-plane. The envelope amplitudes are not modulated in the Z-direction, since the coupled-mode equations for $A_\pm$ and $B_\pm$ are essentially two-dimensional. The polarization vectors are chosen in the z-direction such that $\mathbf{e}_{\mathbf{k}_j} = (0,0,1)$, $1 \le j \le 4$, and $\mathbf{E}_0 = (0,0,E_{0,z}(x,y,X,Y,T)e^{-i\omega t})$. The nonhomogeneous equation (3.8) at the z-component of the solution $\mathbf{E}_1$ at $e^{-i\omega t}$ takes the form

$$\begin{aligned}\nabla^2 E_{1,z} + k^2 E_{1,z} = &-2ik^2 \frac{\partial}{\partial T} E_{0,z} - 2k \frac{\partial^2}{\partial X \partial x} E_{0,z} - 2k \frac{\partial^2}{\partial Y \partial y} E_{0,z} \\ &- \frac{2k^2 n_1(\mathbf{x})}{n_0} E_{0,z} + \frac{2}{n_0} \frac{\partial^2 n_1(\mathbf{x})}{\partial z^2} E_{0,z}.\end{aligned} \tag{3.15}$$

By removing the resonant terms at $e^{\frac{i}{\sqrt{2}}(\pm kx \pm ky)}$, the coupled-mode equations for amplitudes $A_\pm(X,Y,T)$ and $B_\pm(X,Y,T)$ take the form

$$i\left(\frac{\partial A_+}{\partial T} + \frac{\partial A_+}{\partial X} + \frac{\partial A_+}{\partial Y}\right) + \alpha A_- + \beta\,(B_+ + B_-) = 0, \tag{3.16}$$

$$i\left(\frac{\partial A_-}{\partial T} - \frac{\partial A_-}{\partial X} - \frac{\partial A_-}{\partial Y}\right) + \alpha A_+ + \beta\,(B_+ + B_-) = 0, \tag{3.17}$$

$$i\left(\frac{\partial B_+}{\partial T} + \frac{\partial B_+}{\partial X} - \frac{\partial B_+}{\partial Y}\right) + \beta\,(A_+ + A_-) + \alpha B_- = 0, \tag{3.18}$$

$$i\left(\frac{\partial B_-}{\partial T} - \frac{\partial B_-}{\partial X} + \frac{\partial B_-}{\partial Y}\right) + \beta\,(A_+ + A_-) + \alpha B_+ = 0, \tag{3.19}$$

where $\alpha = \alpha_{1,1,0} = \alpha_{-1,-1,0} = \alpha_{1,-1,0} = \alpha_{-1,1,0}$ and $\beta = \alpha_{0,1,0} = \alpha_{1,0,0} = \alpha_{0,-1,0} = \alpha_{-1,0,0}$. The coupled-mode equations (3.16)–(3.19) can be defined in the domain $(X,Y) \in \mathcal{D}$ and $T \ge 0$, where $\mathcal{D}$ is a domain on the (x,y)-plane of the photonic crystal. The system has not been previously studied in literature, to the best of our knowledge.

3.3. Coupled-mode equations for two oblique waves. Two oblique resonant waves on the (x,y)-plane are defined by the resonant wave vectors (2.16) under the constraint (2.17). Assuming that $\mathbf{e}_1 = \mathbf{e}_2 = (0,0,1)$, the Maxwell equations can be reduced to the same form (3.15), where the resonant terms are eliminated at the wave vectors $\mathbf{k}_1 = \mathbf{k}_{\text{in}}$ and $\mathbf{k}_2 = \mathbf{k}_{\text{out}}^{(n,m,0)}$. The coupled-mode equations for amplitudes $A_{1,2}(X,Y,T)$ take the form

$$i\left(\frac{\partial A_1}{\partial T} + \frac{p}{\sqrt{p^2+q^2}}\frac{\partial A_1}{\partial X} + \frac{q}{\sqrt{p^2+q^2}}\frac{\partial A_1}{\partial Y}\right) + \alpha A_2 = 0, \tag{3.20}$$

$$i\left(\frac{\partial A_2}{\partial T} + \frac{p+2n}{\sqrt{p^2+q^2}}\frac{\partial A_2}{\partial X} + \frac{q+2m}{\sqrt{p^2+q^2}}\frac{\partial A_2}{\partial Y}\right) + \alpha A_1 = 0, \tag{3.21}$$

where $\alpha = \alpha_{n,m,0} = \alpha_{-n,-m,0}$. Coupled-mode equations (3.20)–(3.21) for two oblique waves cannot be reduced to the one-dimensional system (3.12)–(3.13), since the characteristics in the system (3.20)–(3.21) are no longer parallel.

The coupled-mode equations for three oblique resonant waves (2.18) can be derived similarly, subject to the resonance condition (2.19). Three characteristics along the wave vectors $\mathbf{k}_1 = \mathbf{k}_{\rm in}$, $\mathbf{k}_2 = \mathbf{k}_{\rm out}^{(n_1,m_1,0)}$, and $\mathbf{k}_3 = \mathbf{k}_{\rm out}^{(n_2,m_2,0)}$ belong to the same (X,Y)-plane. The stationary transmission problem for the three oblique waves is hence a boundary-value problem on the (X,Y)-plane with three (linearly dependent) characteristic coordinates. Oblique interaction of three oblique resonant Bloch waves in a hexagonal crystal was considered numerically in [18].

4. Analysis of stationary transmission. The stationary transmission problem follows from separation of variables in the coupled-mode equations (3.9):

$$A_j(\mathbf{X},T) = a_j(\mathbf{X})e^{-i\Omega T}, \quad j = 1,\ldots,N, \tag{4.1}$$

where Ω is the detuning frequency. When the boundary-value problem for $a_j(\mathbf{X})$ is well posed in a bounded domain, analytical solutions for the linear stationary coupled-mode equations can be derived by using separation of variables and generalized Fourier series [24]. Exploiting these analytical solutions, integral invariants of the stationary transmission, reflection, and diffraction of the resonant Bloch waves can be computed explicitly. We analyze here the stationary coupled-mode equations for two and four counter-propagating and two oblique resonant Bloch waves.

4.1. Transmission of two counter-propagating waves. After separation of variables (4.1), the linear coupled-mode equations (3.12)–(3.13) reduce to the following ODE system:

$$i\frac{da_+}{dZ} + \Omega a_+ + \alpha a_- = 0, \tag{4.2}$$

$$-i\frac{da_-}{dZ} + \alpha a_+ + \Omega a_- = 0. \tag{4.3}$$

The problem (4.2)–(4.3) is defined on the interval $0 \le Z \le L_Z$. When the incident wave strikes the photonic crystal from the left, the linear system (4.2)–(4.3) is completed by the boundary conditions

$$a_+(0) = \alpha_+, \qquad a_-(L_Z) = 0, \tag{4.4}$$

where α_+ is the given amplitude of the incident wave at the left (x,y)-plane of the crystal. The general solution of the ODE system (4.2)–(4.3) is given explicitly as follows:

$$\begin{pmatrix} a_+ \\ a_- \end{pmatrix} = c_+ \begin{pmatrix} \alpha \\ \Omega + i\kappa \end{pmatrix} e^{\kappa Z} + c_- \begin{pmatrix} \alpha \\ \Omega - i\kappa \end{pmatrix} e^{-\kappa Z}, \tag{4.5}$$

where $c_\pm \in \mathbb{C}$ are arbitrary and $\kappa \in \mathbb{C}$ is the root of the determinant equation

$$\kappa = \sqrt{\alpha^2 - \Omega^2}. \tag{4.6}$$

When $\kappa = iK$, $K \in \mathbb{R}$, the linear dispersion relation $\Omega = \Omega(K)$ follows from the quadratic equation

$$\Omega^2 = \alpha^2 + K^2. \tag{4.7}$$

The two branches of the dispersion relation (4.7) correspond to the two counter-propagating resonant waves. Their resonance leads to the photonic stop band, which is located in the interval $|\Omega| < |\alpha|$. Let $\Omega = 0$ for simplicity; i.e., the detuning frequency is fixed in the middle of the stop band. The unique solution of the boundary-value problem (4.2)–(4.4) follows from the general solution (4.5):

$$\begin{pmatrix} a_+ \\ a_- \end{pmatrix} = \frac{\alpha_+}{\cosh \alpha L_Z} \begin{pmatrix} \cosh \alpha (L_Z - Z) \\ -i \sinh \alpha (L_Z - Z) \end{pmatrix}. \tag{4.8}$$

The transmittance T and reflectance R are defined from the other boundary values of the solution (4.8),

$$T = \left| \frac{a_+(L_Z)}{a_+(0)} \right|^2 = \frac{1}{\cosh^2 \alpha L_Z}, \qquad R = \left| \frac{a_-(0)}{a_+(0)} \right|^2 = \frac{\sinh^2 \alpha L_Z}{\cosh^2 \alpha L_Z}, \tag{4.9}$$

such that the balance identity $T + R = 1$ is satisfied. The analytical solution (4.8) for the two counter-propagating waves is well known [23] and is reproduced here for comparison with the case of four counter-propagating and two oblique waves on the plane.

4.2. Transmission of four counter-propagating waves. The stationary transmission of four counter-propagating waves in the coupled-mode equations (3.16)–(3.19) is studied in the characteristic coordinates (ξ, η):

$$\xi = \frac{X + Y}{2}, \qquad \eta = \frac{X - Y}{2}. \tag{4.10}$$

After the separation of variables (4.1), the linear coupled-mode equations (3.16)–(3.19) reduce to the PDE system

$$i \frac{\partial a_+}{\partial \xi} + \Omega a_+ + \alpha a_- + \beta (b_+ + b_-) = 0, \tag{4.11}$$

$$-i \frac{\partial a_-}{\partial \xi} + \alpha a_+ + \Omega a_- + \beta (b_+ + b_-) = 0, \tag{4.12}$$

$$i \frac{\partial b_+}{\partial \eta} + \beta (a_+ + a_-) + \Omega b_+ + \alpha b_- = 0, \tag{4.13}$$

$$-i \frac{\partial b_-}{\partial \eta} + \beta (a_+ + a_-) + \alpha b_+ + \Omega b_- = 0. \tag{4.14}$$

The problem (4.11)–(4.14) is defined in a bounded domain on the plane (ξ, η). We consider the rectangle

$$\mathcal{D} = \{(\xi, \eta) : 0 \le \xi \le L_\xi,\ 0 \le \eta \le L_\eta\}, \tag{4.15}$$

which corresponds to a rectangle in physical coordinates (X, Y), rotated at 45° in characteristic coordinates (ξ, η). When the incident wave moves along the main diagonal in the (X, Y)-plane of the photonic crystal, the linear system (4.11)–(4.14) is completed by the boundary conditions

$$a_+(0, \eta) = \alpha_+(\eta), \qquad a_-(L_\xi, \eta) = 0, \qquad b_+(\xi, 0) = 0, \qquad b_-(\xi, L_\eta) = 0, \tag{4.16}$$

where $\alpha_+(\eta)$ is the given amplitude of the incident wave at the left boundary of the crystal. The linear dispersion relation $\Omega = \Omega(K_\xi, K_\eta)$, where (K_ξ, K_η) are Fourier

wave numbers, follows from the determinant equation of the linear PDE system (4.11)–(4.14).

LEMMA 4.1. *The linear dispersion relation $\Omega = \Omega(K_\xi, K_\eta)$ is defined by the roots of $D(\Omega, K_\xi, K_\eta)$, where*

$$D(\Omega, K_\xi, K_\eta) = (\Omega^2 - \alpha^2 - K_\xi^2)(\Omega^2 - \alpha^2 - K_\eta^2) - 4\beta^2(\Omega - \alpha)^2. \tag{4.17}$$

There exist real-valued roots of $D(0, K_\xi, K_\eta) = 0$ for $\alpha^2 \leq 4\beta^2$, while no real-valued roots exist for $\alpha^2 > 4\beta^2$.

Proof. The determinant equation follows from the PDE system (4.11)–(4.14) for the Fourier modes $e^{i(K_\xi \xi + K_\eta \eta)}$ in the explicit form

$$D(\Omega, K_\xi, K_\eta) = \begin{vmatrix} \Omega - K_\xi & \alpha & \beta & \beta \\ \alpha & \Omega + K_\xi & \beta & \beta \\ \beta & \beta & \Omega - K_\eta & \alpha \\ \beta & \beta & \alpha & \Omega + K_\eta \end{vmatrix}. \tag{4.18}$$

Although the straightforward computations of $D(\Omega, K_\xi, K_\eta)$ are involved technically, it is easy to compute that

$$\frac{\partial D}{\partial \Omega} = 2\Omega(\Omega^2 - \alpha^2 - K_\xi^2) + 2\Omega(\Omega^2 - \alpha^2 - K_\eta^2) - 8\beta^2\Omega + 8\alpha\beta^2 \tag{4.19}$$

and

$$D(0, K_\xi, K_\eta) = (\alpha^2 + K_\xi^2)(\alpha^2 + K_\eta^2) - 4\alpha^2\beta^2. \tag{4.20}$$

Integrating (4.19)–(4.20), we find that $D(\Omega, K_\xi, K_\eta)$ is given by (4.17). When $\alpha^2 > 4\beta^2$, the function $D(0, K_\xi, K_\eta)$ is positive definite on $(K_\xi, K_\eta) \in \mathbb{R}^2$ such that no real-valued roots (K_ξ, K_η) exist for $\Omega = 0$. When $\alpha^2 \leq 4\beta^2$, there exist two curves on the (K_ξ, K_η)-plane, which correspond to the real-valued roots of $D(0, K_\xi, K_\eta)$. ∎

There are four surfaces of the dispersion relations $\Omega = \Omega(K_\xi, K_\eta)$, which correspond to the four resonant counter-propagating Bloch waves. When $\alpha^2 > 4\beta^2$, the interaction of four resonant waves leads to a stop band near the zero detuning frequency $\Omega = 0$. When $\alpha^2 \leq 4\beta^2$, no stop bands occur in the interaction of the four resonant waves. We consider solutions of the system (4.11)–(4.14) at $\Omega = 0$. By separating variables [24], we reduce the PDE problem to two ODE problems as follows:

$$a_+(\xi, \eta) = u_+(\xi) w_a(\eta), \qquad a_-(\xi, \eta) = u_-(\xi) w_a(\eta), \tag{4.21}$$

$$b_+(\xi, \eta) = w_b(\xi) v_+(\eta), \qquad b_-(\xi, \eta) = w_b(\xi) v_-(\eta), \tag{4.22}$$

where

$$v_+(\eta) + v_-(\eta) = \mu w_a(\eta), \qquad u_+(\xi) + u_-(\xi) = -\lambda w_b(\xi), \tag{4.23}$$

parameters (λ, μ) are arbitrary, and vectors $(u_+, u_-)^T$ and $(v_+, v_-)^T$ solve the two uncoupled ODE systems

$$\begin{pmatrix} i\partial_\xi & \alpha \\ \alpha & -i\partial_\xi \end{pmatrix} \begin{pmatrix} u_+ \\ u_- \end{pmatrix} = \beta \Gamma^{-1} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} u_+ \\ u_- \end{pmatrix} \tag{4.24}$$

and

$$\begin{pmatrix} i\partial_\eta & \alpha \\ \alpha & -i\partial_\eta \end{pmatrix} \begin{pmatrix} v_+ \\ v_- \end{pmatrix} = \beta \Gamma \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} v_+ \\ v_- \end{pmatrix}, \tag{4.25}$$

where $\Gamma = \lambda/\mu$. The boundary conditions for (4.24)–(4.25) follow from (4.16) as follows:

$$u_+(0) = 1, \qquad u_-(L_\xi) = 0 \tag{4.26}$$

and

$$v_+(0) = v_-(L_\eta) = 0. \tag{4.27}$$

The homogeneous problem (4.25) and (4.27) defines the spectrum of Γ, while the inhomogeneous problem (4.24) and (4.26) defines a particular solution (4.21)–(4.22). The general solution of the problem (4.11)–(4.14) with the boundary values (4.16) is thought to be a linear superposition of infinitely many particular solutions, if the convergence and completeness of the decomposition formulas can be proved [24]. We first give solutions of the two problems above and then consider the orthogonality and completeness of the generalized Fourier series.

LEMMA 4.2. *All eigenvalues* Γ *of the homogeneous problem* (4.25) *and* (4.27) *are given by nonzero roots of the characteristic equation*

$$\mathcal{R} = \left\{ k \in \mathbb{C} : \left(\frac{k-\alpha}{k+\alpha} \right)^2 e^{-2ikL_\eta} = 1, \ \mathrm{Re}(k) \geq 0, \ k \neq 0 \right\} \tag{4.28}$$

such that

$$\Gamma = \frac{\alpha^2 + k^2}{2\alpha\beta}. \tag{4.29}$$

Let $\alpha > 0$. *Then the roots* $k \in \mathcal{R}$ *are all located in the first open quadrant of* $k \in \mathbb{C}$. *Moreover, all roots are simple, and there exist* $C > 0$ *and* $N \in \mathbb{Z}_+$ *such that only one root* $k \in \mathcal{R}$ *is located in each rectangle:*

$$\mathcal{D}_n^+ = \left\{ k \in \mathbb{C} : \frac{\pi(4n-1)}{2L_\eta} < k < \frac{\pi(4n+1)}{2L_\eta}, \ 0 < \mathrm{Im}(k) < C \right\}, \quad n \geq N, \tag{4.30}$$

and

$$\mathcal{D}_n^- = \left\{ k \in \mathbb{C} : \frac{\pi(4n+1)}{2L_\eta} < k < \frac{\pi(4n+3)}{2L_\eta}, \ 0 < \mathrm{Im}(k) < C \right\}, \quad n \geq N. \tag{4.31}$$

Proof. The general solution of the ODE system (4.25) with the use of (4.29) is found explicitly as follows:

$$\begin{pmatrix} v_+ \\ v_- \end{pmatrix} = c_k \begin{pmatrix} \alpha - k \\ \alpha + k \end{pmatrix} e^{ik\eta} + c_{-k} \begin{pmatrix} \alpha + k \\ \alpha - k \end{pmatrix} e^{-ik\eta}. \tag{4.32}$$

The coefficients c_k and c_{-k} satisfy the relations due to the boundary conditions (4.27):

$$\frac{c_k}{c_{-k}} = \frac{k+\alpha}{k-\alpha} = \frac{k-\alpha}{k+\alpha} e^{-2ikL_\eta}, \tag{4.33}$$

from which the characteristic equation (4.28) for roots $k \in \mathbb{C}$ follows. The symmetric roots k and $(-k)$ correspond to the same Γ and $v_\pm(\eta)$. The root $k = 0$ corresponds to the zero solution for $v_\pm(\eta)$. Therefore, the roots $k = 0$ and $\mathrm{Re}(k) < 0$ are excluded

from the definition of $\mathcal{R}$. The characteristic equation (4.28) results in the modulus equation

$$\frac{|k-\alpha|}{|k+\alpha|} = |e^{ikL_\eta}|.$$

When $\alpha > 0$, the left-hand side equals 1 at $\mathrm{Re}(k) = 0$ and is smaller than 1 for $\mathrm{Re}(k) > 0$. The right-hand side equals 1 at $\mathrm{Im}(k) = 0$ and is larger than 1 for $\mathrm{Im}(k) < 0$. Therefore, roots $k \in \mathcal{R}$ may occur only in the first open quadrant of $k \in \mathbb{C}$.

Let the roots $k \in \mathcal{R}$ be defined by the function $f(k) = (k-\alpha)^2 e^{-2ikL_\eta} - (k+\alpha)^2 = 0$. Then,

$$f'(k) = -\frac{2i(k+\alpha)}{(k-\alpha)}[(k^2-\alpha^2)L_\eta + 2i\alpha]. \tag{4.34}$$

Since the values of $k^2-\alpha^2$ for $k \in \mathcal{R}$ are located in the upper half-plane of the complex plane, $f'(k) \neq 0$ for $\alpha > 0$ such that all roots of $k \in \mathcal{R}$ are simple.

The characteristic equation (4.28) splits into two sets of roots $\mathcal{R}_+$ and $\mathcal{R}_-$ such that $\mathcal{R}_+ \cup \mathcal{R}_- = \mathcal{R}$, where

$$\mathcal{R}_\pm = \left\{ k \in \mathbb{C} : \frac{k-\alpha}{k+\alpha}\, e^{-ikL_\eta} = \pm 1,\ \mathrm{Re}(k) > 0 \right\}. \tag{4.35}$$

We consider the set $k \in \mathcal{R}_+$ and rewrite it in the form $f(k) + g(k) = 0$, where

$$f(k) = e^{ikL_\eta} - 1, \qquad g(k) = \frac{2\alpha}{k+\alpha}.$$

The function $f(k)$ has a zero at

$$k = k_n = \frac{2\pi n}{L_\eta}, \quad n \geq 1.$$

Let us consider the domain $\tilde{\mathcal{D}}_n^+$:

$$\tilde{\mathcal{D}}_n^+ = \left\{ k \in \mathbb{C} : \frac{\pi(4n-1)}{2L_\eta} < k < \frac{\pi(4n+1)}{2L_\eta},\ -C < \mathrm{Im}(k) < C \right\}, \quad n \geq N,$$

for some large $C > 0$ and $N \geq 1$, such that $\frac{\pi(4n-1)}{2L_\eta} > \alpha$. The domain $\tilde{\mathcal{D}}_n^+$ surrounds a simple zero of $f(k)$ at $k = k_n$ such that $|f(k)| > |g(k)|$ on the boundary of $\tilde{\mathcal{D}}_n^+$. By Rouche's theorem, the function $f(k) + g(k)$ has the same number of zeros inside $\tilde{\mathcal{D}}_n^+$ as $f(k)$ does, i.e., only one zero. Since the roots are located in the first open quadrant of $k \in \mathbb{C}$, the root in $\tilde{\mathcal{D}}_n^+$ is located in $\mathcal{D}_n^+$. The same analysis applies to the second set $k \in \mathcal{R}_-$ in the domain $\mathcal{D}_n^-$. □

Roots $k \in \mathcal{R}$ and $(-k) \in \mathcal{R}$ are shown in Figure 1 from the numerical solution of the characteristic equation (4.28) for $\alpha = 1$ and $L_\eta = 20$. In agreement with Lemma 4.2, all roots $k \in \mathcal{R}$ are isolated points in the first open quadrant, which accumulate to the real axis of k at infinity. The standard analysis of analytic functions at infinity leads to the asymptotic formula for distribution of large roots k in the domain $|k| > k_0 \gg 1$:

$$k_n^+ = \frac{2\pi n}{L_\eta} + \frac{i\alpha}{\pi n} + \mathrm{O}\left(\frac{1}{n^2}\right), \qquad k_n^- = \frac{\pi(1+2n)}{L_\eta} + \frac{2i\alpha}{\pi(1+2n)} + \mathrm{O}\left(\frac{1}{n^2}\right), \tag{4.36}$$

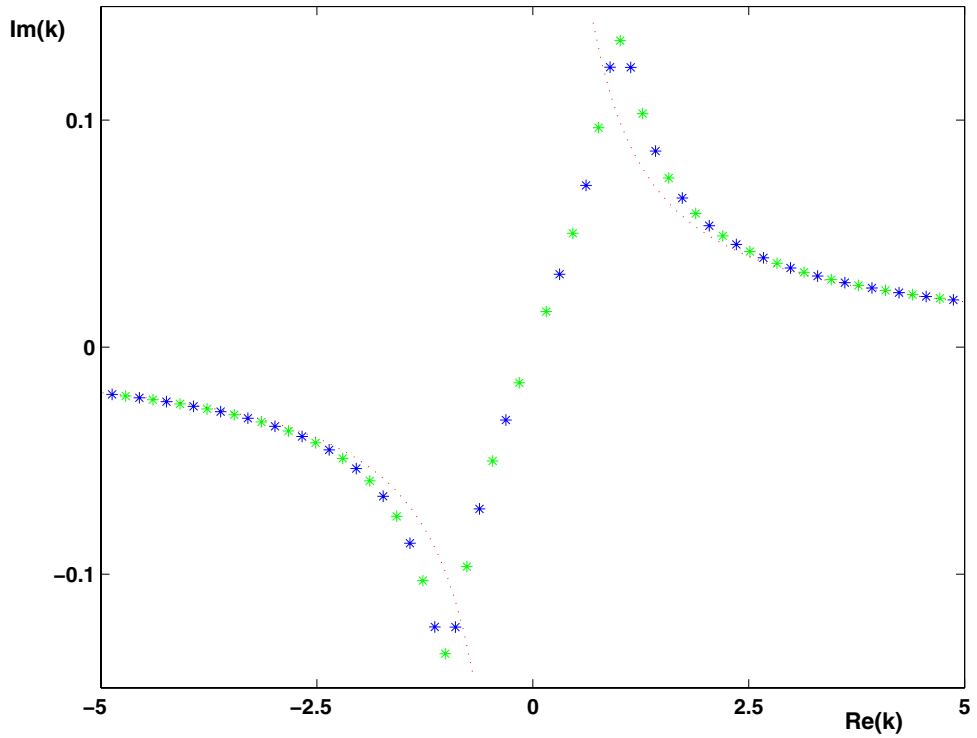


FIG. 1. *Roots $k \in \mathcal{R}$ and $(-k) \in \mathcal{R}$ of the characteristic equation* (4.28) *for $\alpha = 1$ and $L_\eta = 20$. Dark dots show roots of $\mathcal{R}_+$, and bright dots show roots of $\mathcal{R}_-$. The dotted curves show the leading-order asymptotic approximation* (4.36).

where n is a large positive integer. The leading order of the asymptotic approximation (4.36) is also shown in Figure 1 by dotted curves. The two sets in (4.36) correspond to the splitting $k \in \mathcal{R}_\pm$ in (4.35). The eigenfunction $v(\eta) = v_+(\eta) + v_-(\eta)$ is symmetric (antisymmetric) with respect to $\eta = L_\eta/2$ for $k \in \mathcal{R}_+$ ($k \in \mathcal{R}_-$). Moreover, explicit formulas for $v(\eta)$ follow from (4.32) and (4.33):

$$k \in \mathcal{R}_+ : v(\eta) = c_+ \cos k \left(\frac{L_\eta}{2} - \eta \right), \tag{4.37}$$

$$k \in \mathcal{R}_- : v(\eta) = c_- \sin k \left(\frac{L_\eta}{2} - \eta \right), \tag{4.38}$$

where (c_+, c_-) are normalization constants. Asymptotic solutions (4.36) correspond to two sets of eigenfunctions

$$\left\{ \cos(\pi n \tilde{\eta}),\ \sin\left(\frac{\pi(2n+1)\tilde{\eta}}{2} \right) \right\}, \quad \tilde{\eta} = \frac{2\eta}{L_\eta} - 1, \tag{4.39}$$

which solve the homogeneous Neumann problem on the normalized interval $-1 \le \tilde{\eta} \le 1$.

LEMMA 4.3. *Let Γ be an eigenvalue of the problem* (4.25) *and* (4.27). *There exists a unique solution of the nonhomogeneous problem* (4.24) *and* (4.26) *for this Γ.*

Proof. A general solution of the ODE system (4.24) is found explicitly as follows:

$$\begin{pmatrix} u_+ \\ u_- \end{pmatrix} = d_k \begin{pmatrix} \alpha^2+k^2-2\beta^2 \\ \lambda_k(\alpha^2+k^2)+2\beta^2 \end{pmatrix} e^{i\alpha\lambda_k\eta} + d_{-k} \begin{pmatrix} \alpha^2+k^2-2\beta^2 \\ -\lambda_k(\alpha^2+k^2)+2\beta^2 \end{pmatrix} e^{-i\alpha\lambda_k\eta}, \tag{4.40}$$

where

$$\lambda_k = \sqrt{\frac{4\beta^2}{\alpha^2+k^2} - 1}. \tag{4.41}$$

The relation (4.41) satisfies the determinant equation (4.17) such that $D(0, \alpha\lambda_k, k) = 0$. Using the boundary conditions (4.26), coefficients d_k and d_{-k} are found uniquely, under the constraint

$$u_0 = \lambda_k(\alpha^2+k^2)\cos\alpha\lambda_k L_\xi + 2i\beta^2 \sin\alpha\lambda_k L_\xi \neq 0. \tag{4.42}$$

We show that $u_0 \neq 0$. The equation $u_0 = 0$ can be rewritten in the form

$$\frac{(\lambda_k-1)^2}{(\lambda_k+1)^2} = e^{2i\alpha\lambda_k L_\xi}. \tag{4.43}$$

By analysis of Lemma 4.2, it is clear that nonzero roots of the characteristic equation (4.43) may exist only in the first and third open quadrants of $\lambda_k \in \mathbb{C}$ for $\alpha > 0$, such that the values of $\lambda_k^2 + 1$ are located in the upper half-plane of the complex plane. The zero root $\lambda_k = 0$ is located on the real axis for $\lambda_k^2 + 1$. On the other hand, the values of $4\beta^2/(\alpha^2+k^2)$ for $k \in \mathcal{R}$ are located in the lower half-plane. Therefore, the relation (4.41) leads to a contradiction, which proves that $u_0 \neq 0$. ∎

Solutions of the nonhomogeneous problem (4.24) and (4.26) with the normalization $u_+(0) = u_0 \neq 0$ can be written explicitly by eliminating d_k and d_{-k} from the implicit form (4.40):

$$\begin{pmatrix} u_+ \\ u_- \end{pmatrix} = \lambda_k(\alpha^2+k^2) \begin{pmatrix} 1 \\ 0 \end{pmatrix} \cos\alpha\lambda_k(L_\xi - \xi) + i \begin{pmatrix} 2\beta^2 \\ \alpha^2+k^2-2\beta^2 \end{pmatrix} \sin\alpha\lambda_k(L_\xi - \xi). \tag{4.44}$$

When the representation (4.21) is used for $\alpha_+(\eta) = a_+(0,\eta)$, the function $\alpha_+(\eta)$ is expanded as a series of scalar eigenfunctions $v(\eta) = v_+(\eta) + v_-(\eta)$, defined for roots $k \in \mathcal{R}$. This decomposition is possible only if the set of eigenfunctions $v(\eta)$ is orthogonal and complete.

LEMMA 4.4. *There exists a set of normalized and orthogonal eigenfunctions $v_j(\eta)$ for distinct roots $k = k_j \in \mathcal{R}$, according to the inner product*

$$\int_0^{L_\eta} v_i(\eta) v_j(\eta) d\eta = \delta_{i,j}. \tag{4.45}$$

Proof. The set of adjoint eigenvectors to the problem (4.25) and (4.27) with respect to the standard inner product in $L^2([0, L_\eta])$ is given by the vectors $(\bar{v}_-, \bar{v}_+)^T$. As a result, the scalar eigenfunctions $v_j(\eta)$ for distinct roots $k = k_j$ satisfy the orthogonality relations (4.45) with $i \neq j$. The scalar eigenfunction $v(\eta)$ is found from (4.32) and (4.33) in the explicit form

$$v(\eta) = c_0(k\cos k\eta + i\alpha \sin k\eta), \tag{4.46}$$

where c_0 is a normalization constant. Integrating $v^2(\eta)$ on $\eta \in [0, L_\eta]$, we confirm that the eigenfunctions $v_j(\eta)$ can be normalized by the inner product (4.45) with $i = j$, under the constraint

$$(k^2 - \alpha^2)L_\eta + 2i\alpha \neq 0. \tag{4.47}$$

Since the roots $k \in \mathcal{R}$ are all simple, such that $f'(k) \neq 0$ in (4.34), the constraint (4.47) is met. □

PROPOSITION 4.5. *Any continuously differentiable complex-valued function $f(\eta)$ on $0 \leq \eta \leq L_\eta$ is uniquely represented by the series of eigenfunctions*

$$f(\eta) = \sum_{\text{all } k_j \in \mathcal{R}} c_j v_j(\eta), \quad c_j = \int_0^{L_\eta} f(\eta) v_j(\eta) d\eta, \tag{4.48}$$

and the series converges to $f(\eta)$ uniformly on $0 \leq \eta \leq L_\eta$.

Proof. It follows from (4.25) and (4.27) that the scalar eigenfunction $v(\eta)$ solves the second-order boundary-value problem

$$v'' + k^2 v = 0 \tag{4.49}$$

such that

$$iv'(0) + \alpha v(0) = 0, \qquad -iv'(L_\eta) + \alpha v(L_\eta) = 0. \tag{4.50}$$

The Sommerfeld radiation boundary conditions (4.50) explain why the spectrum of the formally self-adjoint operator (4.49) is complex-valued. The statement of the proposition follows from the expansion theorem [7, p. 303], since the theorem's condition is satisfied: $A_{2,4} = 1$, where $A_{2,4}$ is the determinant of the second and fourth columns of the matrix A, associated with the boundary conditions

$$A = \begin{pmatrix} \alpha & i & 0 & 0 \\ 0 & 0 & \alpha & -i \end{pmatrix}.$$

As a result, the Fourier series of asymptotic eigenfunctions (4.39) approximates the series expansion (4.48) for large roots $k = k_n^{\pm}$ uniformly on $\eta \in [0, L_\eta]$. The uniform convergence of (4.48) follows from that of the Fourier series [24]. □

Using separation of variables and convergence of series of eigenfunctions, we summarize the existence and uniqueness results on the generalized Fourier series solutions of the linear boundary-value problem (4.11)–(4.14) and (4.16) with $\Omega = 0$.

PROPOSITION 4.6. *Let the set $\{c_j\}$ be uniquely defined by the series* (4.48) *for $f(\eta) = \alpha_+(\eta)$. There exists a unique solution of the boundary-value problem* (4.11)–(4.14) *and* (4.16) *with $\Omega = 0$ in the domain* (4.15):

$$a_+(\xi, \eta) = \sum_{\text{all } k_j \in \mathcal{R}} c_j \frac{u_{+j}(\xi)}{u_{+j}(0)} (v_{+j}(\eta) + v_{-j}(\eta)), \tag{4.51}$$

$$a_-(\xi, \eta) = \sum_{\text{all } k_j \in \mathcal{R}} c_j \frac{u_{-j}(\xi)}{u_{+j}(0)} (v_{+j}(\eta) + v_{-j}(\eta)), \tag{4.52}$$

$$b_+(\xi, \eta) = -\sum_{\text{all } k_j \in \mathcal{R}} c_j \frac{u_{+j}(\xi) + u_{-j}(\xi)}{\Gamma_j u_{+j}(0)} v_{+j}(\eta), \tag{4.53}$$

$$b_-(\xi, \eta) = -\sum_{\text{all } k_j \in \mathcal{R}} c_j \frac{u_{+j}(\xi) + u_{-j}(\xi)}{\Gamma_j u_{+j}(0)} v_{-j}(\eta). \tag{4.54}$$

We illustrate the generalized Fourier series solutions (4.51)–(4.54) with two examples: (i) a single term of the generalized Fourier series and (ii) a constant input function $\alpha_+(\eta) = \alpha_+$. For both examples, we compute the integral invariants for the incident ($\mathcal{I}_{\rm in}$), transmitted ($\mathcal{I}_{\rm out}$), reflected ($\mathcal{I}_{\rm ref}$), and diffracted ($\mathcal{I}_{\rm dif}$) waves from their definitions:

$$\mathcal{I}_{\rm in} = \int_0^{L_\eta} |a_+(0,\eta)|^2 d\eta, \qquad \mathcal{I}_{\rm out} = \int_0^{L_\eta} |a_+(L_\xi,\eta)|^2 d\eta, \tag{4.55}$$

$$\mathcal{I}_{\rm ref} = \int_0^{L_\eta} |a_-(0,\eta)|^2 d\eta, \qquad \mathcal{I}_{\rm dif} = \int_0^{L_\xi} (|b_+(\xi,L_\eta)|^2 + |b_-(\xi,0)|^2) d\xi. \tag{4.56}$$

Let the transmittance T, reflectance R, and diffractance D be defined from the relations

$$T = \frac{\mathcal{I}_{\rm out}}{\mathcal{I}_{\rm in}}, \qquad R = \frac{\mathcal{I}_{\rm ref}}{\mathcal{I}_{\rm in}}, \qquad D = \frac{\mathcal{I}_{\rm dif}}{\mathcal{I}_{\rm in}}. \tag{4.57}$$

The integral invariants satisfy the balance identity

$$R + T + D = 1, \tag{4.58}$$

which follows from integration of the balance equation

$$\frac{\partial}{\partial \xi}(|a_+|^2 - |a_-|^2) + \frac{\partial}{\partial \eta}(|b_+|^2 - |b_-|^2) = 0. \tag{4.59}$$

First, we consider a single term of the Fourier series solutions (4.51)–(4.54). The transmittance and reflectance for $k \in \mathcal{R}$ are found from (4.44) in the explicit form

$$T_k = \left| \frac{\lambda_k(\alpha^2 + k^2)}{\lambda_k(\alpha^2 + k^2)\cos \alpha\lambda_k L_\xi + 2i\beta^2 \sin \alpha\lambda_k L_\xi} \right|^2, \tag{4.60}$$

$$R_k = \left| \frac{(\alpha^2 + k^2 - 2\beta^2)\sin \alpha\lambda_k L_\xi}{\lambda_k(\alpha^2 + k^2)\cos \alpha\lambda_k L_\xi + 2i\beta^2 \sin \alpha\lambda_k L_\xi} \right|^2, \tag{4.61}$$

while the diffractance is found from the balance identity as $D_k = 1 - T_k - R_k$. These integral invariants of the stationary transmission for $\alpha = 1$ and $L_\xi = L_\eta = 20$ are shown in Figure 2 for $\beta = 0.25$ and in Figure 3 for $\beta = 0.75$. In the first case, when $\alpha^2 > 4\beta^2$, there is a stop band at $\Omega = 0$, such that all modes are fully reflected except for small losses due to diffraction. In the second case, when $\alpha^2 < 4\beta^2$, there is no stop band at $\Omega = 0$, such that transmittance and diffractance are large for smaller values of $|k|$ and become negligible for larger values of $|k|$.

Next, we consider a constant input function:

$$\alpha_+(\eta) = \alpha_+, \quad \eta \in [0, L_\eta], \tag{4.62}$$

when c_j can be found from (4.48),

$$c_j = \frac{4i\alpha\alpha_+}{k_j[L(k_j^2 - \alpha^2) + 2i\alpha]}, \quad k_j \in \mathcal{R}_+, \tag{4.63}$$

and $c_j = 0$ for $k_j \in \mathcal{R}_-$. The solution surfaces $|a_\pm(\xi,\eta)|^2$ and $|b_\pm(\xi,\eta)|^2$ in the domain (4.15) are shown for $\alpha = 1$, $L_\xi = L_\eta = 20$, and $\alpha_+ = 1$ in Figure 4 for $\beta = 0.25$ and

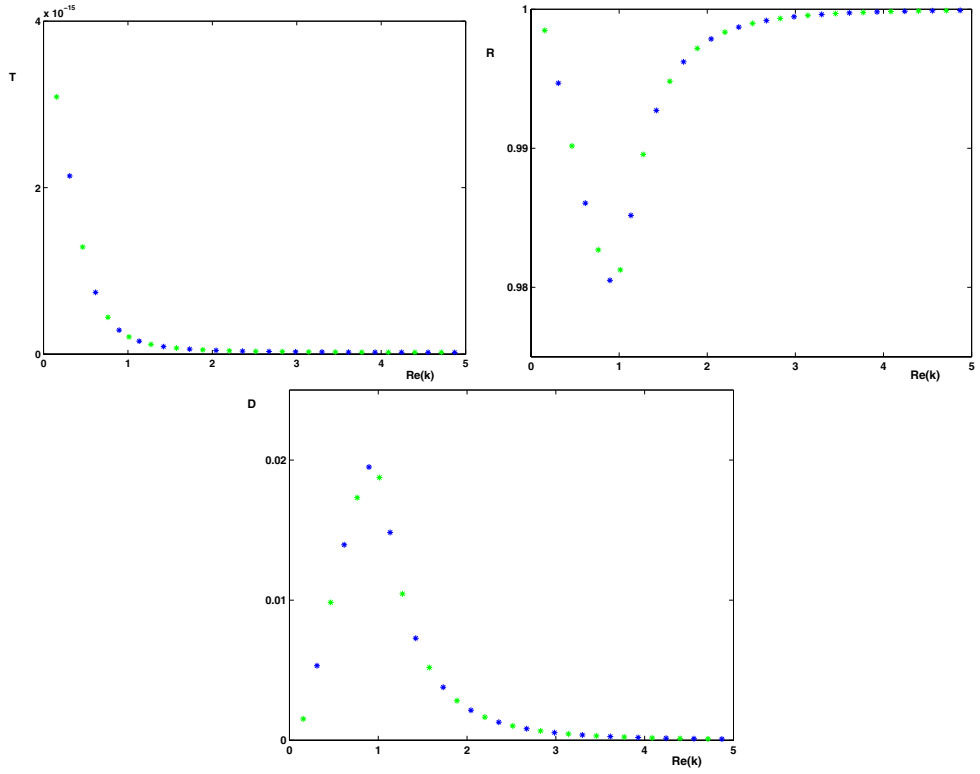


FIG. 2. *Transmittance* (T_k)*, reflectance* (R_k)*, and diffractance* (D_k) *versus* $\mathrm{Re}(k)$ *for the roots* $k \in \mathcal{R}$ *when* $\alpha = 1$*,* $\beta = 0.25$*, and* $L_\xi = L_\eta = 20$*.*

in Figure 5 for $\beta = 0.75$. We can see from the figures that the boundary conditions (4.16) are satisfied by the truncated generalized Fourier series (4.51)–(4.54) with only 30 first terms.

The Parseval identity cannot be applied to eigenfunctions $v_j(\eta)$, because the inner product (4.45) is not the standard inner product in $L^2([0, L_\eta])$. As a result, the energy spectrum of $I_{\rm out}$, $I_{\rm ref}$, and $I_{\rm dif}$ cannot be decomposed into a superposition of the squared amplitudes $|c_j|^2$. Nevertheless, the numerical values for T, R, and D can be found from numerical integration of the solution surfaces (4.55)–(4.56). The numerical values are

$$\begin{aligned} &\beta = 0.25 : T \approx 3 \times 10^{-15}, \quad R \approx 0.9853, \quad D \approx 0.0147, \\ &\beta = 0.75 : T \approx 0.7394, \quad R \approx 0.0133, \quad D \approx 0.2473, \end{aligned}$$

such that $T + R + D \approx 1$. When $\alpha^2 > 4\beta^2$, there exists a stop band at $\Omega = 0$, and the incident wave is reflected from the photonic crystal with energy loss of 1.5% due to diffraction. When $\alpha^2 < 4\beta^2$, there is no stop band at $\Omega = 0$, and the incident wave is transmitted along the photonic crystal with energy loss of 26% due to reflection and diffraction.

4.3. Transmission of two oblique waves. The stationary transmission of two oblique waves in the coupled-mode equations (3.20)–(3.21) becomes diagonal in the

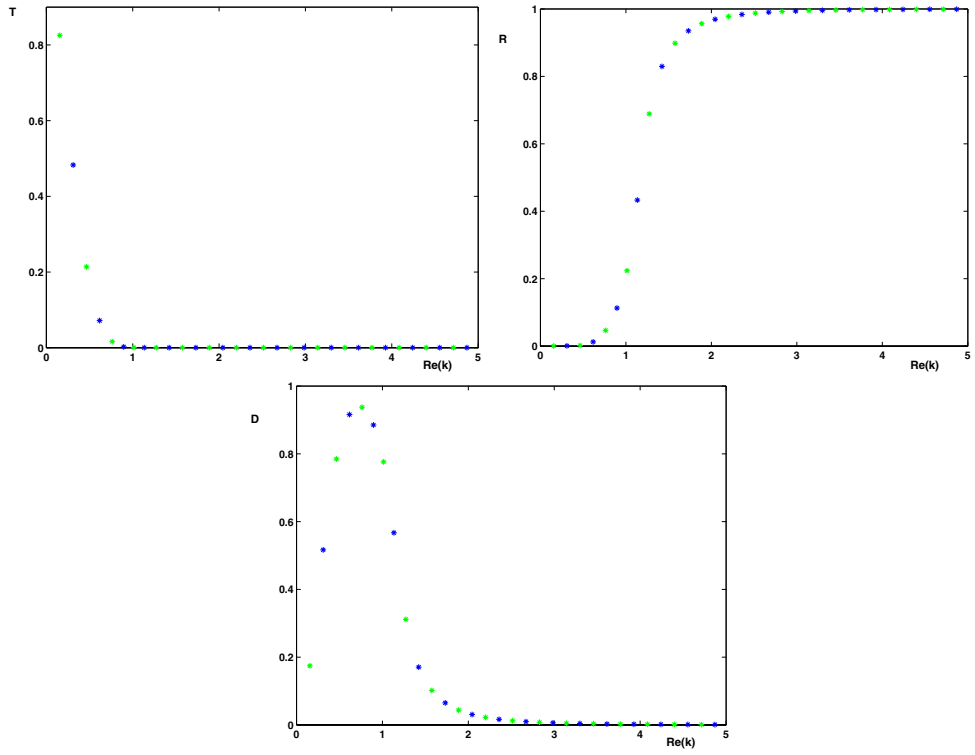


FIG. 3. *Transmittance* (T_k)*, reflectance* (R_k)*, and diffractance* (D_k) *versus* $\mathrm{Re}(k)$ *for the roots* $k \in \mathcal{R}$ *when* $\alpha = 1$*,* $\beta = 0.75$*, and* $L_\xi = L_\eta = 20$*.*

characteristic coordinates (ξ, η):

$$X = \frac{p\xi + (p+2n)\eta}{\sqrt{p^2+q^2}}, \qquad Y = \frac{q\xi + (q+2m)\eta}{\sqrt{p^2+q^2}}. \tag{4.64}$$

After the separation of variables (4.1), the linear coupled-mode equations (3.20)–(3.21) reduce to the PDE system

$$i\frac{\partial a_1}{\partial \xi} + \Omega a_1 + \alpha a_2 = 0, \tag{4.65}$$

$$i\frac{\partial a_2}{\partial \eta} + \alpha a_1 + \Omega a_2 = 0. \tag{4.66}$$

Coordinate axes (ξ, η) are parallel to the wave vectors $\mathbf{k}_1 = \mathbf{k}_{\rm in}$ and $\mathbf{k}_2 = \mathbf{k}_{\rm out}^{(n,m,0)}$, but they are no longer orthogonal. The problem (4.65)–(4.66) is defined in a bounded domain on the plane (ξ, η). We consider the same rectangle $\mathcal{D}$, defined by (4.15). When the incident wave is illuminated in the direction of the wave vector $\mathbf{k}_1$ but not in the direction of the wave vector $\mathbf{k}_2$, the linear system (4.65)–(4.66) is completed by the boundary conditions

$$a_1(0,\eta) = \alpha_1(\eta), \qquad a_2(\xi, 0) = 0. \tag{4.67}$$

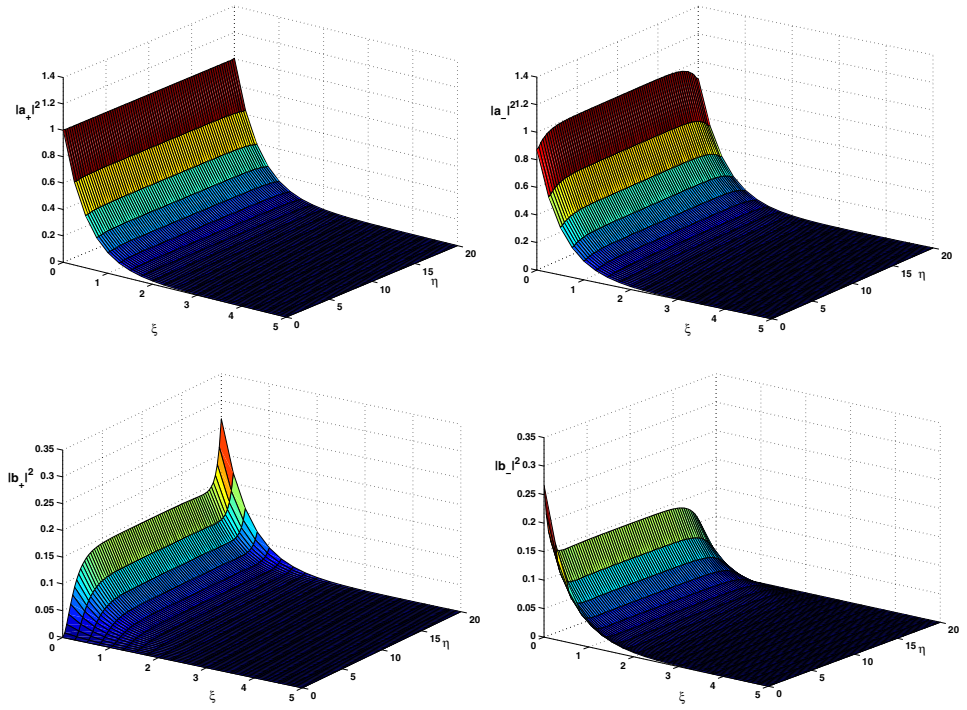


FIG. 4. *Solution surfaces* $|a_\pm|^2(\xi,\eta)$ *and* $|b_\pm|^2(\xi,\eta)$ *on the domain* (4.15) *for* $\alpha = 1$, $\beta = 0.25$, $L_\xi = L_\eta = 20$, *and* $\alpha_+ = 1$.

The linear dispersion relation $\Omega = \Omega(K_\xi, K_\eta)$, where (K_ξ, K_η) are Fourier wave numbers, is given explicitly as

$$\left(\Omega - \frac{K_\xi + K_\eta}{2}\right)^2 = \alpha^2 + \left(\frac{K_\xi - K_\eta}{2}\right)^2. \tag{4.68}$$

Two surfaces of the dispersion relation (4.68) correspond to the two oblique resonant waves. In a moving reference frame on the plane (ξ,η) there exists a stop band in the dispersion relation (4.68). We consider solutions of the system (4.65)–(4.66) at $\Omega = 0$ by using the Fourier transform

$$a_1(\xi,\eta) = \int_{-\infty}^{\infty} kc(k) e^{i\alpha(k^{-1}\xi + k\eta)} dk, \tag{4.69}$$

$$a_2(\xi,\eta) = \int_{-\infty}^{\infty} c(k) e^{i\alpha(k^{-1}\xi + k\eta)} dk. \tag{4.70}$$

It follows from the boundary conditions (4.67) that

$$kc(k) = \frac{\alpha}{2\pi} \int_0^{L_\eta} \alpha_1(\eta) e^{-i\alpha k\eta} d\eta, \quad k \in \mathbb{R}, \tag{4.71}$$

and

$$0 = \int_{-\infty}^{\infty} c(k) e^{i\alpha k^{-1}\xi} dk, \quad 0 \le \xi \le L_\xi. \tag{4.72}$$

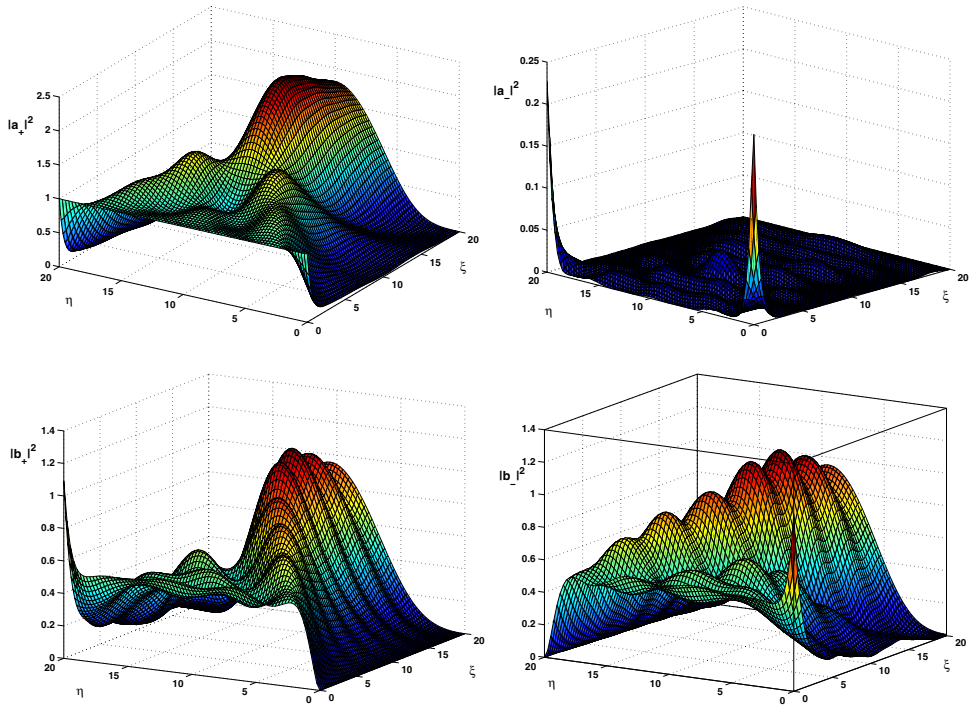


FIG. 5. *Solution surfaces* $|a_\pm|^2(\xi,\eta)$ *and* $|b_\pm|^2(\xi,\eta)$ *on the domain* (4.15) *for* $\alpha = 1$, $\beta = 0.75$, $L_\xi = L_\eta = 20$, *and* $\alpha_+ = 1$.

Interchanging integrals, we reduce the constraint (4.72) to the form

$$0 = \frac{\alpha}{2\pi i} \int_0^{L_\eta} \alpha_1(\eta) \left(\int_{-\infty}^{\infty} \frac{\sin \alpha(k\eta - k^{-1}\xi)}{k} dk \right) d\eta, \quad 0 \leq \xi \leq L_\xi. \tag{4.73}$$

The inner integral is zero for $\xi > 0$ and $\eta > 0$, due to the table integral 3.871 on p. 474 of [9]. Therefore, the constraint (4.72) is satisfied, and a unique solution of the problem (4.65)–(4.67) exists in the form (4.69)–(4.71).

We illustrate the Fourier transform solution (4.69)–(4.70) with the constant input function

$$\alpha_1(\eta) = \alpha_1, \quad \eta \in [0, L_\eta], \tag{4.74}$$

when $c(k)$ can be found from (4.71):

$$c(k) = \frac{\alpha_1}{2\pi i} \frac{1 - e^{-i\alpha k L_\eta}}{k^2}, \quad k \in \mathbb{R}. \tag{4.75}$$

Evaluating Fourier integrals (4.69)–(4.70) with the help of the table integral 3.871 on p. 474 of [9], we find the explicit solution of the stationary problem:

$$a_1(\xi,\eta) = \alpha_1 J_0(2\alpha\sqrt{\xi\eta}), \qquad a_2(\xi,\eta) = \frac{i\alpha_1\sqrt{\eta}}{\sqrt{\xi}} J_1(2\alpha\sqrt{\xi\eta}), \tag{4.76}$$

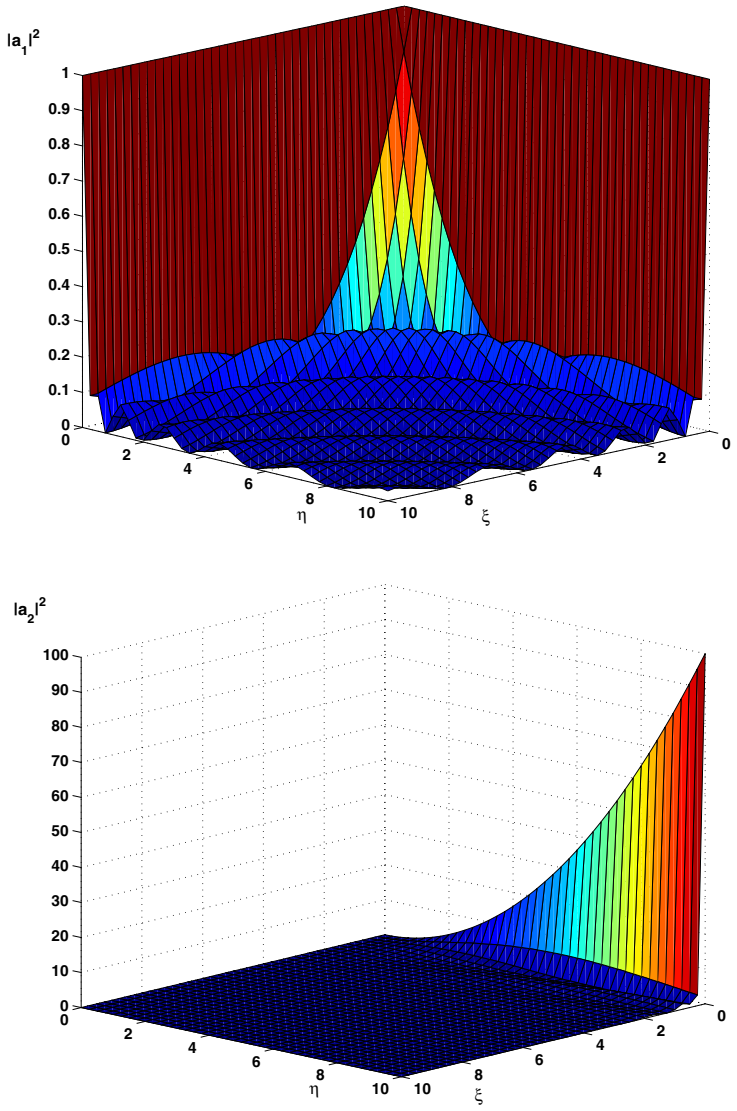


FIG. 6. *Solution surfaces* $|a_1|^2(\xi,\eta)$ *and* $|a_2|^2(\xi,\eta)$ *on the domain* (4.15) *for* $\alpha = 1$, $L_\xi = L_\eta = 10$, *and* $\alpha_1 = 1$.

where $J_{0,1}(z)$ are Bessel functions [9]. Figure 6 shows the solution surfaces $|a_1(\xi,\eta)|^2$ and $|a_2(\xi,\eta)|^2$ in the domain (4.15) for $\alpha = 1$, $L_\xi = L_\eta = 10$, and $\alpha_1 = 1$. The integral invariants for the stationary transmission follow from integration of the balance equation:

$$\frac{\partial}{\partial\xi}|a_1|^2 + \frac{\partial}{\partial\eta}|a_2|^2 = 0. \tag{4.77}$$

We define the incident ($\mathcal{I}_{\rm in}$), transmitted ($\mathcal{I}_{\rm out}$), and diffracted ($\mathcal{I}_{\rm dif}$) intensities by

$$\mathcal{I}_{\rm in} = \int_0^{L_\eta} |a_1(0,\eta)|^2 d\eta, \quad \mathcal{I}_{\rm out} = \int_0^{L_\eta} |a_1(L_\xi,\eta)|^2 d\eta, \quad \mathcal{I}_{\rm dif} = \int_0^{L_\xi} |a_2(\xi,L_\eta)|^2 d\xi. \tag{4.78}$$

The transmittance (T) and diffractance (D) are defined by the same relations (4.57), and the balance identity $T + D = 1$ follows from integration of the balance equation (4.77). The numerical values for T and D are found from numerical integration of the solution surfaces (4.78) as follows:

$$T \approx 0.032, \qquad D \approx 0.968,$$

such that $T + D \approx 1$. These values show that the incident wave is diffracted to the oblique resonance wave such that only 3.2% of the wave energy remains in the transmitted wave.

4.4. General transmission problems. A general system of coupled-mode equations (3.9) can be diagonalized in characteristic coordinates, similarly to the case of four counter-propagating and two oblique resonant waves. The characteristic coordinates are introduced from the set of resonant wave vectors as follows:

$$\frac{\partial}{\partial \xi_j} = \frac{\mathbf{k}_{j,x}}{k}\frac{\partial}{\partial X} + \frac{\mathbf{k}_{j,y}}{k}\frac{\partial}{\partial Y} + \frac{\mathbf{k}_{j,z}}{k}\frac{\partial}{\partial Z}, \quad j = 1, \dots, N, \tag{4.79}$$

such that the characteristic coordinate ξ_j extends in the direction of the wave vector $\mathbf{k}_j$. The characteristic coordinates $(\xi_1, \dots, \xi_N) \in \mathbb{R}^N$ are related to the physical coordinates $(X, Y, Z) \in \mathbb{R}^3$ as follows:

$$\mathbf{X} = \mathbf{X}_0 + \sum_{j=1}^N \xi_j \frac{\mathbf{k}_j}{k}, \tag{4.80}$$

where $\mathbf{X}_0 \in \mathbb{R}^3$ is an arbitrary point. The boundary-value problem for the linear stationary transmission with $\Omega = 0$ can be rewritten in the form

$$i\frac{\partial a_j}{\partial \xi_j} + \sum_{k \neq j} \hat{\alpha}_{j,k} a_k = 0, \quad j = 1, \dots, N. \tag{4.81}$$

We consider the domain of definition in the cone $(\xi_1, \dots, \xi_N) \in \mathbb{R}^N_+$, subject to the Goursat boundary values

$$a_j(\xi_1, \dots, \xi_{j-1}, 0, \xi_{j+1}, \dots, \xi_N) = \alpha_j(\xi_1, \dots, \xi_{j-1}, \xi_{j+1}, \dots, \xi_N), \quad j = 1, \dots, N. \tag{4.82}$$

The Goursat boundary-value problem (4.81)–(4.82) can be rewritten as the Volterra integral equations:

$$a_j(\xi_j) = \alpha_j + i\int_0^{\xi_j} \sum_{k \neq j} \hat{\alpha}_{j,k} a_k(\xi_j') d\xi_j'. \tag{4.83}$$

By the contraction mapping principle [11], there exists a unique solution of the Volterra equations (4.83) in the cone $(\xi_1, \dots, \xi_N) \in \mathbb{R}^N_+$, such that we have the following result.

THEOREM 4.7. *Let $\mathcal{D}$ be a convex domain in $\mathbb{R}^3$, which is cut by the characteristic coordinate projections $\xi_j = 0$, $j = 1, \ldots, N$. There exists a unique solution $a_j = a_j(\xi_1, \ldots, \xi_N)$, which corresponds to the boundary-value problem* (4.81)–(4.82) *and depends smoothly on the boundary values α_j, $j = 1, \ldots, N$.*

If $N = \text{rank}(\mathbf{k}_1, \ldots, \mathbf{k}_N)$, there exists only one domain $\mathcal{D} \subset \mathbb{R}^N$, which corresponds to the cone $\xi_j \geq 0$, $j = 1, \ldots, N$. The case of two oblique waves on the plane (X, Y) gives an example of this situation for $N = 2$. The Goursat problem (4.81)–(4.82) is rewritten as the PDE problem (4.65)–(4.66) with the boundary values (4.67), which is solved with the explicit Fourier transform solutions (4.69)–(4.71).

If $N > \text{rank}(\mathbf{k}_1, \ldots, \mathbf{k}_N)$, the characteristic coordinates $(\xi_1, \ldots, \xi_N)$ are linearly dependent such that there exist multiple ways to choose a convex domain $\mathcal{D} \subset \mathbb{R}^3$ that corresponds to the cone $\xi_j \geq 0$, $j = 1, \ldots, N$. The case of four counter-propagating waves on the plane (X, Y) gives an example of this situation for $N = 4$ and $\text{rank}(\mathbf{k}_1, \ldots, \mathbf{k}_N) = 2$. In this case, we have chosen that $\xi_1 = \xi$, $\xi_2 = \eta$, $\xi_3 = L_\eta - \eta$, and $\xi_4 = L_\xi - \xi$, such that $0 \leq \xi \leq L_\xi$ and $0 \leq \eta \leq L_\eta$. As a result, the Goursat problem (4.81)–(4.82) is rewritten as the PDE problem (4.11)–(4.14) with the boundary values (4.16). Theorem 4.7 does not guarantee well-posedness of (4.11)–(4.14), while explicit Fourier series solutions (4.51)–(4.54) do (see Proposition 4.6).

5. Summary and open problems. We have shown that the coupled-mode equations can be used for analysis and modeling of resonant interaction of Bloch waves in low-contrast cubic-lattice three-dimensional photonic crystals. The analytical solutions for the linear stationary transmission problem are found by using separation of variables and generalized Fourier series. We have proved that the linear stationary boundary-value problem is well-posed in the context of four counter-propagating and two oblique waves on the plane. We have also given general results on well-posedness of the general linear stationary transmission problem.

It remains an open problem to prove well-posedness of the nonlinear stationary boundary-value problem for small-norm and finite-norm solutions. Nonstationary transmission problems are also of interest, and very few analytical results are available on local and global well-posedness of the nonstationary nonlinear coupled-mode equations. Finally, numerical approximations of the stationary and nonstationary, fully nonlinear coupled-mode equations can be constructed in bounded domains with the method of orthogonal polynomials [17]. All these problems are beyond the scope of the present work.

Appendix A. Nonlinear coupled-mode equations with cubic nonlinearities. Modeling of nonlinear photonic band-gap crystals with cubic (Kerr) nonlinearities is based on the Maxwell equations, where the polarization vector depends nonlinearly on the electric field vector $\mathbf{E}$ (see [13]). When the nonlinearity terms are small, nonlocal (dispersive) terms in the polarization vector can be neglected, and the low-contrast weakly nonlinear photonic crystals can be modeled with the Maxwell equations (1.1), where the refractive index $n = n(\mathbf{x}, |\mathbf{E}|^2)$ is decomposed into the linear and nonlinear parts [23]:

$$n(\mathbf{x}, |\mathbf{E}|^2) = n_0 + \epsilon n_1(\mathbf{x}) + \epsilon n_2(\mathbf{x})|\mathbf{E}|^2, \tag{A.1}$$

where n_0 is constant and ϵ is of small parameter. When the photonic crystal has cubic-lattice structure, the periodic functions $n_1(\mathbf{x})$ and $n_2(\mathbf{x})$ are expanded into the

triple Fourier series (3.4) and

$$n_2(\mathbf{x}) = n_0 \sum_{(n,m,l)\in\mathbb{Z}^3} \beta_{n,m,l} e^{ik_0(nx+my+lz)}, \quad \beta_{n,m,l} = \bar{\beta}_{-n,-m,-l}, \tag{A.2}$$

where the factor n_0 is included for convenience. Derivation of the nonlinear coupled-mode equations is based on rigorous methods of Lyapunov–Schmidt reductions [16]. Equivalently, the formal derivation can be recovered with perturbation series expansions [19], which follows the formalism (3.2) and (3.3) outlined in section 3. The first-order correction term $\mathbf{E}_1(\mathbf{x},t)$ solves the nonhomogeneous problem (3.8) with additional nonlinear terms:

$$\begin{aligned} \nabla^2\mathbf{E}_1 - \frac{n_0^2}{c^2}\frac{\partial^2\mathbf{E}_1}{\partial t^2} &= 2\frac{n_0^2\omega}{c^2}\frac{\partial^2\mathbf{E}_0}{\partial T\partial t} - 2k\left(\nabla\cdot\nabla_X\right)\mathbf{E}_0 \\ &\quad + \frac{2n_0n_1(\mathbf{x})}{c^2}\frac{\partial^2\mathbf{E}_0}{\partial t^2} + \frac{2}{n_0}\nabla\left(\nabla n_1\cdot\mathbf{E}_0\right) \\ &\quad + \frac{2n_0n_2(\mathbf{x})}{c^2}|\mathbf{E}_0|^2\frac{\partial^2\mathbf{E}_0}{\partial t^2} + \frac{2}{n_0}\nabla\left(\nabla n_2|\mathbf{E}_0|^2\cdot\mathbf{E}_0\right). \end{aligned} \tag{A.3}$$

The cubic nonlinear terms generate N^3 terms from the leading-order solution (3.3), which all give resonant terms by means of the triple series (A.2). By removing the resonant terms, the nonlinear coupled-mode equations for $A_j(\mathbf{X},T)$, $j=1,\ldots,N$, are derived in the general form:

$$i\left(\frac{\partial A_j}{\partial T} + \left(\frac{\mathbf{k}_j}{k}\cdot\nabla_X\right)A_j\right) + \sum_{k\neq j}\hat{\alpha}_{j,k}A_k + \sum_{1\leq k_1,k_2,k_3\leq N}\hat{\beta}_{j,k_1,k_2,k_3}A_{k_1}A_{k_2}\bar{A}_{k_3} = 0, \tag{A.4}$$

where the elements $\{\hat{\beta}_{j,k_1,k_2,k_3}\}_{1\leq j,k_1,k_2,k_3\leq N}$ are related to the Fourier coefficients of the resonant waves $\{\beta_{n,m,l}\}_{(n,m,l)\in\mathcal{S}}$. The explicit forms of the nonlinear coupled-mode equations are given below for two and four counter-propagating and two oblique resonant Bloch waves.

The nonlinear coupled-mode equations for two counter-propagating waves (3.10) generalize the linear equations (3.12)–(3.13) as follows:

$$\begin{aligned} &i\left(\frac{\partial A_+}{\partial T} + \frac{\partial A_+}{\partial Z}\right) + \alpha A_- + \beta_{0,0,0}(|A_+|^2 + 2|A_-|^2)A_+ \\ &\quad + \beta_{0,0,1}(2|A_+|^2 + |A_-|^2)A_- + \beta_{0,0,-1}A_+^2\bar{A}_- + \beta_{0,0,2}\bar{A}_+A_-^2 = 0, \end{aligned} \tag{A.5}$$

$$\begin{aligned} &i\left(\frac{\partial A_-}{\partial T} - \frac{\partial A_-}{\partial Z}\right) + \alpha A_+ + \beta_{0,0,0}(2|A_+|^2 + |A_-|^2)A_- \\ &\quad + \beta_{0,0,-1}(|A_+|^2 + 2|A_-|^2)A_+ + \beta_{0,0,1}\bar{A}_+A_-^2 + \beta_{0,0,-2}A_+^2\bar{A}_- = 0. \end{aligned} \tag{A.6}$$

The system (A.5)–(A.6) is reviewed in [8, 23] for $\beta_{0,0,1} = \beta_{0,0,2} = 0$ and analyzed in [14, 15] for $\beta_{0,0,1} \neq 0$ and $\beta_{0,0,2} = 0$. When $\beta_{0,0,1}, \beta_{0,0,2} \neq 0$, the system (3.12)–(3.13) is the most general coupled-mode system for Bragg resonance of two counter-propagating waves [6, 22].

The nonlinear coupled-mode equations for four counter-propagating waves (3.14) generalize the linear equations (3.16)–(3.19) as follows:

$$i\left(\frac{\partial A_+}{\partial T} + \frac{\partial A_+}{\partial X} + \frac{\partial A_+}{\partial Y}\right) + \alpha A_- + \beta\left(B_+ + B_-\right) + F_+(A_+, A_-, B_+, B_-) = 0, \tag{A.7}$$

$$i\left(\frac{\partial A_-}{\partial T}-\frac{\partial A_-}{\partial X}-\frac{\partial A_-}{\partial Y}\right)+\alpha A_+ +\beta\,(B_+ + B_-)+F_-(A_+,A_-,B_+,B_-)=0, \tag{A.8}$$

$$i\left(\frac{\partial B_+}{\partial T}+\frac{\partial B_+}{\partial X}-\frac{\partial B_+}{\partial Y}\right)+\beta\,(A_+ + A_-)+\alpha B_- +G_+(A_+,A_-,B_+,B_-)=0, \tag{A.9}$$

$$i\left(\frac{\partial B_-}{\partial T}-\frac{\partial B_-}{\partial X}+\frac{\partial B_-}{\partial Y}\right)+\beta\,(A_+ + A_-)+\alpha B_+ +G_-(A_+,A_-,B_+,B_-)=0, \tag{A.10}$$

where the cubic nonlinear functions are given by

$$\begin{aligned}
F_+ =\ & \beta_{0,0,0}((|A_+|^2+2|A_-|^2+2|B_+|^2+2|B_-|^2)A_+ + 2\bar{A}_-B_+B_-)\\
&+\beta_{0,-1,0}(A_+^2\bar{B}_+ + 2A_+\bar{A}_-B_-)+\beta_{-1,0,0}(A_+^2\bar{B}_- + 2A_+\bar{A}_-B_+)\\
&+\beta_{1,1,0}((2|A_+|^2+|A_-|^2+2|B_+|^2+2|B_-|^2)A_- + 2\bar{A}_+B_+B_-)\\
&+\beta_{0,1,0}((2|A_+|^2+2|A_-|^2+|B_+|^2+2|B_-|^2)B_+ + 2A_+A_-\bar{B}_-)\\
&+\beta_{-1,1,0}(2A_+B_+\bar{B}_- + \bar{A}_-B_+^2)+\beta_{1,-1,0}(2A_+\bar{B}_+B_- + \bar{A}_-B_-^2)\\
&+\beta_{1,0,0}((2|A_+|^2+2|A_-|^2+2|B_+|^2+|B_-|^2)B_- + 2A_+A_-\bar{B}_+)\\
&+\beta_{2,0,0}(\bar{A}_+B_-^2 + 2A_-\bar{B}_+B_-)+\beta_{2,1,0}(2\bar{A}_+A_-B_- + A_-^2\bar{B}_+)\\
&+\beta_{1,2,0}(A_-^2\bar{B}_- + 2\bar{A}_+A_-B_+)+\beta_{2,-1,0}\bar{B}_+B_-^2+\beta_{-1,2,0}B_+^2\bar{B}_-\\
&+\beta_{0,2,0}(\bar{A}_+B_+^2 + 2A_-B_+\bar{B}_-)+\beta_{-1,-1,0}A_+^2\bar{A}_- + \beta_{2,2,0}\bar{A}_+A_-^2,\\
F_- =\ & \beta_{-1,-1,0}((|A_+|^2+2|A_-|^2+2|B_+|^2+2|B_-|^2)A_+ + 2\bar{A}_-B_+B_-)\\
&+\beta_{-1,-2,0}(A_+^2\bar{B}_+ + 2A_+\bar{A}_-B_-)+\beta_{-2,-1,0}(A_+^2\bar{B}_- + 2A_+\bar{A}_-B_+)\\
&+\beta_{0,0,0}((2|A_+|^2+|A_-|^2+2|B_+|^2+2|B_-|^2)A_- + 2\bar{A}_+B_+B_-)\\
&+\beta_{-1,0,0}((2|A_+|^2+2|A_-|^2+|B_+|^2+2|B_-|^2)B_+ + 2A_+A_-\bar{B}_-)\\
&+\beta_{-2,0,0}(2A_+B_+\bar{B}_- + \bar{A}_-B_+^2)+\beta_{0,-2,0}(2A_+\bar{B}_+B_- + \bar{A}_-B_-^2)\\
&+\beta_{0,-1,0}((2|A_+|^2+2|A_-|^2+2|B_+|^2+|B_-|^2)B_- + 2A_+A_-\bar{B}_+)\\
&+\beta_{1,-1,0}(\bar{A}_+B_-^2 + 2A_-\bar{B}_+B_-)+\beta_{1,0,0}(2\bar{A}_+A_-B_- + A_-^2\bar{B}_+)\\
&+\beta_{0,1,0}(A_-^2\bar{B}_- + 2\bar{A}_+A_-B_+)+\beta_{1,-2,0}\bar{B}_+B_-^2+\beta_{-2,1,0}B_+^2\bar{B}_-\\
&+\beta_{-1,1,0}(\bar{A}_+B_+^2 + 2A_-B_+\bar{B}_-)+\beta_{-2,-2,0}A_+^2\bar{A}_- + \beta_{1,1,0}\bar{A}_+A_-^2,\\
G_+ =\ & \beta_{0,-1,0}((|A_+|^2+2|A_-|^2+2|B_+|^2+2|B_-|^2)A_+ + 2\bar{A}_-B_+B_-)\\
&+\beta_{0,-2,0}(A_+^2\bar{B}_+ + 2A_+\bar{A}_-B_-)+\beta_{-1,-1,0}(A_+^2\bar{B}_- + 2A_+\bar{A}_-B_+)\\
&+\beta_{1,0,0}((2|A_+|^2+|A_-|^2+2|B_+|^2+2|B_-|^2)A_- + 2\bar{A}_+B_+B_-)\\
&+\beta_{0,0,0}((2|A_+|^2+2|A_-|^2+|B_+|^2+2|B_-|^2)B_+ + 2A_+A_-\bar{B}_-)\\
&+\beta_{-1,0,0}(2A_+B_+\bar{B}_- + \bar{A}_-B_+^2)+\beta_{1,-2,0}(2A_+\bar{B}_+B_- + \bar{A}_-B_-^2)\\
&+\beta_{1,-1,0}((2|A_+|^2+2|A_-|^2+2|B_+|^2+|B_-|^2)B_- + 2A_+A_-\bar{B}_+)\\
&+\beta_{2,-1,0}(\bar{A}_+B_-^2 + 2A_-\bar{B}_+B_-)+\beta_{2,0,0}(2\bar{A}_+A_-B_- + A_-^2\bar{B}_+)\\
&+\beta_{1,1,0}(A_-^2\bar{B}_- + 2\bar{A}_+A_-B_+)+\beta_{2,-2,0}\bar{B}_+B_-^2+\beta_{-1,1,0}B_+^2\bar{B}_-\\
&+\beta_{0,1,0}(\bar{A}_+B_+^2 + 2A_-B_+\bar{B}_-)+\beta_{-1,-2,0}A_+^2\bar{A}_- + \beta_{2,1,0}\bar{A}_+A_-^2,\\
G_- =\ & \beta_{-1,0,0}((|A_+|^2+2|A_-|^2+2|B_+|^2+2|B_-|^2)A_+ + 2\bar{A}_-B_+B_-)\\
&+\beta_{-1,-1,0}(A_+^2\bar{B}_+ + 2A_+\bar{A}_-B_-)+\beta_{-2,0,0}(A_+^2\bar{B}_- + 2A_+\bar{A}_-B_+)\\
&+\beta_{0,1,0}((2|A_+|^2+|A_-|^2+2|B_+|^2+2|B_-|^2)A_- + 2\bar{A}_+B_+B_-)\\
&+\beta_{-1,1,0}((2|A_+|^2+2|A_-|^2+|B_+|^2+2|B_-|^2)B_+ + 2A_+A_-\bar{B}_-)\\
&+\beta_{-2,1,0}(2A_+B_+\bar{B}_- + \bar{A}_-B_+^2)+\beta_{0,-1,0}(2A_+\bar{B}_+B_- + \bar{A}_-B_-^2)\\
&+\beta_{0,0,0}((2|A_+|^2+2|A_-|^2+2|B_+|^2+|B_-|^2)B_- + 2A_+A_-\bar{B}_+)
\end{aligned}$$

$$+\beta_{1,0,0}(\bar{A}_+B_-^2+2A_-\bar{B}_+B_-)+\beta_{1,1,0}(2\bar{A}_+A_-B_-+A_-^2\bar{B}_+)$$
$$+\beta_{0,2,0}(A_-^2\bar{B}_-+2\bar{A}_+A_-B_+)+\beta_{1,-1,0}\bar{B}_+B_-^2+\beta_{-2,2,0}B_+^2\bar{B}_-$$
$$+\beta_{-1,2,0}(\bar{A}_+B_+^2+2A_-B_+\bar{B}_-)+\beta_{-2,-1,0}A_+^2\bar{A}_-+\beta_{1,2,0}\bar{A}_+A_-^2.$$

The nonlinear coupled-mode equations for two oblique waves (2.16) generalize the linear equations (3.20)–(3.21) as follows:

$$i\left(\frac{\partial A_1}{\partial T}+\frac{p}{\sqrt{p^2+q^2}}\frac{\partial A_1}{\partial X}+\frac{q}{\sqrt{p^2+q^2}}\frac{\partial A_1}{\partial Y}\right)+\alpha A_2+\beta_{0,0,0}(|A_1|^2+2|A_2|^2)A_1$$
$$+\beta_{-n,-m,0}(2|A_1|^2+|A_2|^2)A_2+\beta_{n,m,0}A_1^2\bar{A}_2+\beta_{-2n,-2m,0}\bar{A}_1A_2^2=0, \tag{A.11}$$

$$i\left(\frac{\partial A_2}{\partial T}+\frac{p+2n}{\sqrt{p^2+q^2}}\frac{\partial A_2}{\partial X}+\frac{q+2m}{\sqrt{p^2+q^2}}\frac{\partial A_2}{\partial Y}\right)+\alpha A_1+\beta_{0,0,0}(2|A_1|^2+|A_2|^2)A_1$$
$$+\beta_{n,m,0}(|A_1|^2+2|A_2|^2)A_1+\beta_{-n,-m,0}\bar{A}_1A_2^2+\beta_{2n,2m,0}A_1^2\bar{A}_2=0. \tag{A.12}$$

The system (A.11)–(A.12) and its generalization to three oblique resonant waves are reviewed in [18, 19].

Acknowledgments. The authors thank Walter Craig, Ted Sargent, and Jamin Sheriff for collaboration and useful discussions. This paper was completed during the visit of D. P. to Universidad Autonoma del Estado de Mexico, organized by Dr. M. Aguero Granados and supported by research grant SEPPROMEP Mexico /103.5/03/309.

REFERENCES

[1] N. Akozbek and S. John, *Optical solitary waves in two- and three-dimensional nonlinear photonic band-gap structures*, Phys. Rev. E, 57 (1998), pp. 2287–2319.
[2] N. Akozbek and S. John, *Self-induced transparency solitary waves in a doped nonlinear photonic band gap material*, Phys. Rev. E, 58 (1998), pp. 3876–3895.
[3] A. Arraf and C.M. de Sterke, *Coupled-mode equations for quadratically nonlinear deep gratings*, Phys. Rev. E, 58 (1998), pp. 7951–7958.
[4] A. Babin and A. Figotin, *Nonlinear photonic crystals:* I. *Quadratic nonlinearity*, Waves Random Media, 11 (2001), pp. R31–R102.
[5] A. Babin and A. Figotin, *Nonlinear photonic crystals:* II. *Interaction classification for quadratic nonlinearities*, Waves Random Media, 12 (2002), pp. R25–R52.
[6] N. Bhat and J.E. Sipe, *Optical pulse propagation in nonlinear photonic crystals*, Phys. Rev. E, 64 (2001), paper 056604.
[7] E.A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw–Hill, New York, 1955.
[8] R.H. Goodman, M.I. Weinstein, and P.J. Holmes, *Nonlinear propagation of light in one-dimensional periodic structures*, J. Nonlinear Sci., 11 (2001), pp. 123–168.
[9] I.S. Gradshtein and I.M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, New York, 2000.
[10] C. Kittel, *Introduction to Solid-State Physics*, John Wiley & Sons, New York, 1996.
[11] A.N. Kolmogorov and S.V. Fomin, *Elements of the Theory of Functions and Functional Analysis*, Nauka, Moscow, 1989.
[12] P. Kuchment, *Floquet Theory for Partial Differential Operators*, Birkhäuser, Basel, 1993.
[13] P. Kuchment, *The mathematics of photonic crystals*, in Mathematical Modeling in Optical Sciences, G. Bao, L. Cowsar, and W. Masters, eds., Frontiers in Appl. Math 22, SIAM, Philadelphia, 1999, pp. 207–272.
[14] D. Pelinovsky, J. Sears, L. Brzozowski, and E.H. Sargent, *Stable all-optical limiting in nonlinear periodic structures.* I*: Analysis*, J. Opt. Soc. Amer. B Opt. Phys., 19 (2002), pp. 43–53.
[15] D.E. Pelinovsky and A. Scheel, *Spectral analysis of stationary light transmission in nonlinear photonic structures*, J. Nonlinear Sci., 13 (2003), pp. 347–396.

[16] G. Schneider and H. Uecker, *Existence and stability of modulating pulse solutions in Maxwell's equations describing nonlinear optics*, Z. Angew. Math. Phys., 54 (2003), pp. 677–712.
[17] J. Shen, *Efficient spectral-Galerkin method* I*: Direct solvers of second- and fourth-order equations using Legendre polynomials*, SIAM J. Sci. Comput., 15 (1994), pp. 1489–1505.
[18] J.L. Sheriff, I.A. Goldthorpe, and E.H. Sargent, *Optical limiting and intensity-dependent diffraction from low-contrast nonlinear periodic media: Coupled-mode analysis*, Phys. Rev. E, 70 (2004), paper 036616.
[19] J.L. Sheriff, *Coupled-mode theory in low-contrast nonlinear photonic crystals*, B.Sc. thesis, ECE Department, University of Toronto, Toronto, ON, 2003.
[20] C.M. de Sterke and J.E. Sipe, *Envelope-function approach for the electrodynamics of nonlinear periodic structures*, Phys. Rev. A, 38 (1988), pp. 5149–5165.
[21] C.M. de Sterke and J.E. Sipe, *Extensions and generalizations of an envelope-function approach for the electrodynamics of nonlinear periodic structures*, Phys. Rev. A, 39 (1989), pp. 5163–5178.
[22] C.M. de Sterke, D.G. Salinas, and J.E. Sipe, *Coupled-mode theory for light propagation through deep nonlinear gratings*, Phys. Rev. E, 54 (1996), pp. 1969–1989.
[23] C.M. de Sterke and J.E. Sipe, *Gap solitons*, Progress in Optics, 33 (1994), pp. 203–260.
[24] W. Strauss, *Partial Differential Equations: An Introduction*, John Wiley & Sons, New York, 1992.

SIAM J. APPL. MATH.
Vol. 65, No. 4, pp. 1130–1152

A UNIVERSAL PROCEDURE FOR NORMALIZING n-DEGREE-OF-FREEDOM POLYNOMIAL HAMILTONIAN SYSTEMS*

SUSANA GUTIÉRREZ-ROMERO†, JESÚS F. PALACIÁN†, AND PATRICIA YANGUAS†

Abstract. We depart from an n-degree-of-freedom Hamiltonian formed by the sum of homogeneous polynomials in n coordinates and n momenta with arbitrary coefficients. By extending formally an integral of the principal part of the system to the full Hamiltonian and truncating higher-order terms, we obtain a simplified Hamiltonian. This "normalization" procedure can be used to extract qualitative features of the departure system. In this paper we present the symbolic routines needed to achieve the normalization. The power and generality of the algorithm are exhibited through two examples.

Key words. polynomial Hamiltonians, generalized normal forms, formal integrals, reduction and simplification, monodromy, invariants

AMS subject classifications. 34C14, 34C20, 34C27, 58K10, 70H09, 70H12, 70H15, 70H33

DOI. 10.1137/S0036139903434390

1. Introduction.

This work is devoted to the algorithmic implementation of the theory developed in [25, 26, 27] dealing with "normalization" of polynomial Hamiltonian systems in n degrees of freedom (n DOFs). We are interested in the actual construction of "normal form" Hamiltonians, as well as their associated changes of coordinates and formal integrals. The motivation for this series of works is the wide rank of applications of normal form theory in the last 30 years in various fields, such as classical mechanics, astrodynamics, qualitative theory of ordinary differential equations, or molecular physics.

We consider Hamilton functions of the form

$$\mathcal{H}(\boldsymbol{x};\varepsilon) = \sum_{i\geq 0} \frac{\varepsilon^i}{i!}\,\mathcal{H}_i(\boldsymbol{x}), \tag{1.1}$$

where $\boldsymbol{x} = (x_1,\ldots,x_n,X_1,\ldots,X_n)$ is a $(2n)$-dimensional vector in the coordinates $x_1,\ldots,x_n$ and corresponding momenta $X_1,\ldots,X_n$. Moreover, each $\mathcal{H}_i$ is an arbitrary homogeneous polynomial in $\boldsymbol{x}$ of degree $i+p+2$ for some fixed integer $p \geq -1$. Note that with this choice of p we can deal with polynomial Hamiltonians starting at degree one and therefore we do not restrict ourselves to local analysis around equilibria points. Indeed if our starting Hamiltonian function is a polynomial our approach is of global nature; however, if we are interested in the neighborhood of a critical point, then p should be 0. Thus, we consider the most generic class of Hamiltonians in the polynomial context.

*Received by the editors December 5, 2003; accepted for publication (in revised form) July 16, 2004; published electronically April 14, 2005. This work was partially supported by project BFM2002-03157 of Ministerio de Ciencia y Tecnología (Spain), by Project Resolución 92/2002 of Departamento de Educación y Cultura, Gobierno de Navarra (Spain), by project ACPI2002/04 of Departamento de Educación y Cultura, Gobierno de La Rioja (Spain), and by project API02/20 of Universidad de La Rioja (Spain). The work of the first author was also supported by the predoctoral grant AP2002-1548.

http://www.siam.org/journals/siap/65-4/43439.html

†Departamento de Matemática e Informática, Universidad Pública de Navarra, 31006 Pamplona, Navarra, Spain (susana.gutierrez@unavarra.es, palacian@unavarra.es, yanguas@unavarra.es).

Our purpose is to simplify system (1.1) by introducing a constant of motion independent of $\mathcal{H}$, up to a certain degree. The tool for carrying out our normalizing procedures are the well-known Lie transformations for canonical systems [6]. From the practical point of view the normalization up to any order becomes a cumbersome task due to two drawbacks principally. (1) The form of the principal term $\mathcal{H}_0$ dictates the form acquired by the homological equation, which is the one needed to be solved at each order to calculate the "normal form." Thus, for certain types of $\mathcal{H}_0$ the procedure should be stopped at a certain order and should not be extended up to any order. (2) The computational trouble caused by the huge amount of terms in those cases where the calculations are carried out to very high degrees or when the number of DOFs is big, let us say, greater than four.

From our side we have adopted the compromise of building an algorithm being as generic as possible, sacrificing on some occasions the possibility of writing an optimal set of routines for some types of $\mathcal{H}_0$. The reason for this choice lies on the fact that we intend to present a universal procedure. Thus, we have designed our algorithm by following three criteria: (i) $\mathcal{H}_0$ can be any homogenous polynomial of degree $p+2$ in $\boldsymbol{x}$ with real or complex coefficients, $p \geq -1$ being an integer. Indeed, it is not necessary to bring $\mathcal{H}_0$ to a prescribed form to resolve the corresponding homological equations. (ii) The classical techniques of normal forms for Hamiltonians [20] (see also the general case in [18, 3, 19, 22]) are enlarged, as we make use of the concept of generalized normal forms [25, 26]. This allows us either to execute the usual approach of normalization or to extend other integrals of $\mathcal{H}_0$ up to a certain order. (iii) Our algorithm is based on the routines we have prepared with MATHEMATICA, but it can be programmed in other similar algebraic manipulators, such as MAPLE or MACSYMA.

Poincaré [30] is considered a pioneer in the simplification of systems of differential equations, as he developed a method applicable to systems of not necessarily Hamiltonian nature. Birkhoff [1] considered the Hamiltonian version thereafter. The normalization of semisimple systems in equilibrium at the origin was carried out for the planar case by Whittaker [34]. Moser [24] extended the work of Birkhoff to resonant Hamiltonians in $2n$ dimensions. Thereafter there appeared a generalization by Meyer [20], who presented the general solution for a Hamiltonian system whose corresponding matrix related to the linear part was semisimple. Normal forms for nonsemisimple matrices have been studied by Meyer and Schmidt [23], Sokol'skij [32], and van der Meer [18], for instance. Furthermore, Meyer [21] established the normal form theorem for any type of equilibrium point in systems of ordinary differential equations whose main part is linear. The Hamiltonian case appears in Meyer and Hall [22].

In the classical theory of normal forms the dominant term $\mathcal{H}_0$ is a homogeneous polynomial of degree two (so $p=0$). Associated with it is a linear differential system of equations. When the matrix defining the linear system has a nonnull semisimple part, the normal form theorem for the general equilibrium [22] is the best choice for reducing the number of degrees of freedom of the original system, provided that the reduced system would not have a trivial flow. However, if the matrix is nilpotent or if the reduced flow is trivial, we should resort to the extended normal form approach proposed in [25, 26].

This article is divided into five sections. Section 2 revises the basic theory on the subject of transformations bringing a system to a generalized normal form. In section 3 we offer a detailed description of the routines included in our algorithms. In section 4 we give two examples where the advantages of the exposed algorithm with respect to other current theories are shown. The first example consists in a 1-DOF

Hamiltonian vector field whose dominant part is a homogenous polynomial of degree 3. We show that under certain conditions on the coefficients of the Hamiltonian it is possible to introduce a polynomial formal integral. This allows the reduction of the system to another of 0 DOF. The motivation for the second example is the occurrence of Hamiltonian–Hopf bifurcations and monodromy in a swing spring. Finally, in section 5 we outline the conclusions of the paper.

2. Generalized normal forms and formal integrals. In this section we broach the normalization of Hamiltonian systems through the construction of formal integrals. First, we recall the normal form theorem [21, 22]. (See also the previous contributions by van der Meer [18, 19], Cushman, Deprit, and Mosak [3], and Churchill, Kummer, and Rod [2].) Thereafter we continue with the concept of generalized normal forms.

2.1. Normal form theorem. Let us recall this standard result, which combines results by van der Meer and Meyer.

THEOREM 2.1. *Let*

$$\mathcal{H}(\boldsymbol{x};\varepsilon) = \sum_{i\geq 0} \frac{\varepsilon^i}{i!}\,\mathcal{H}_i(\boldsymbol{x}) \tag{2.1}$$

be a Hamilton function such that each $\mathcal{H}_i(\boldsymbol{x})$ is a homogeneous polynomial of degree $i+2$ in $\boldsymbol{x}\in\boldsymbol{R}^{2n}$. It defines a differential system of n DOFs whose quadratic part is $\mathcal{H}_0(\boldsymbol{x}) = \frac{1}{2}\,\boldsymbol{x}^t\,B\,\boldsymbol{x}$, where B is a symmetric $(2n\times 2n)$-matrix. Let $\mathcal{J}$ be the standard skew-symmetric matrix of order $2n$

$$\mathcal{J} = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix}$$

(I_n stands for the identity matrix of order n), and let $A = \mathcal{J}\,B$ be the matrix associated with the linear system of differential equations defined by $\mathcal{H}_0$. Let $\boldsymbol{x} = \boldsymbol{X}(\boldsymbol{y};\varepsilon) = \boldsymbol{y} + \cdots$ be the symplectic change of coordinates that transforms $\mathcal{H}$ into its normal form (up to an order $L\geq 1$), the convergent Hamiltonian $\mathcal{K}(\boldsymbol{y}) = \mathcal{K}_0(\boldsymbol{y}) + \mathcal{K}_1(\boldsymbol{y}) + \cdots + \mathcal{K}_L(\boldsymbol{y}) + \mathcal{O}(\varepsilon^{L+1})$. Let $A = S + N$ represent the Jordan decomposition of A into its semisimple (S) plus nilpotent (N) parts. Then, the quadratic polynomial $\mathcal{I}_S(\boldsymbol{y}) = \mathcal{I}_S(\boldsymbol{x}) = -\frac{1}{2}\,\boldsymbol{x}^t\,\mathcal{J}\,S\,\boldsymbol{x}$ is an integral of $\mathcal{H}_0$, provided that $S\neq 0$. Moreover, by means of the normal form transformation, $\mathcal{I}_S(\boldsymbol{y})$ becomes an integral (formal integral) of $\mathcal{K}$ independent of it and the number of DOFs of the system related to $\mathcal{K}$, after truncation, is at most $n-1$.

Proof. See the papers by van der Meer [19] and Meyer [21] and the book by Meyer and Hall [22]. □

First, if we define $\mathcal{K}_0^t = \frac{1}{2}\boldsymbol{y}^t R\boldsymbol{y}$ with $R = \mathcal{J}B\mathcal{J} = -\mathcal{J}A^t$, when applying the normal form theorem, Theorem 2.1, the terms $\mathcal{K}_i$ for $i\geq 1$ are built so that they satisfy $\{\mathcal{K}_i, \mathcal{K}_0^t\} = 0$. That is, the terms of the normal form are invariant under the flow defined by $\exp(A^t s)$. In this respect one can understand that the normal form Hamiltonian is simpler than the original Hamiltonian.

As is well known, normal forms are in general divergent; see, for instance, [29]. Indeed the transformation proposed in Theorem 2.1 does not converge in general. However, we can say that $\mathcal{K}$ is convergent as we are including the tail $\mathcal{O}(\varepsilon^{L+1})$ in its definition.

When $S\neq 0$ one can build a formal integral of $\mathcal{H}$ up to a certain order of approximation, as we shall detail later on. Therefore, the initial system is reduced by

the calculation of a formal integral. Nevertheless, when $S = 0$ we cannot construct a new formal integral by the application of the normal form theorem. In such a situation we resort to another strategy to reduce the number of DOFs of $\mathcal{H}$.

The number of DOFs defined by $\mathcal{K}$ is $n - 1$ at most. It is smaller when the semisimple component of $\mathcal{H}_0$ is composed by the sum of semisimple terms such that there is no resonance among all of them. In these situations the reduction procedure could be too drastic in the sense that one would not obtain enough information from the reduced system [28]. A way to overcome this problem would be to take another integral of $\mathcal{H}_0$ to be extended to the whole system. This is the subject of the following subsections.

2.2. Extension of an integral of the principal part to the full system. We provide a result which can be used to enlarge the applicability of the normal form theorem. We start with some useful definitions.

DEFINITION 2.2. *Given two scalar fields $\mathcal{P}$ and $\mathcal{Q}$ defined over an open domain of $\boldsymbol{R}^{2n}$, the Poisson bracket of $\mathcal{P}$ and $\mathcal{Q}$ is given by the relation*

$$\{\mathcal{P}, \mathcal{Q}\} = \sum_{i=1}^{n} \left(\frac{\partial \mathcal{P}}{\partial x_i} \frac{\partial \mathcal{Q}}{\partial X_i} - \frac{\partial \mathcal{P}}{\partial X_i} \frac{\partial \mathcal{Q}}{\partial x_i} \right).$$

DEFINITION 2.3. *Given two scalar fields $\mathcal{P}$ and $\mathcal{Q}$, the Lie operator associated with $\mathcal{Q}$ is given by means of the Poisson bracket $\mathcal{L}_{\mathcal{Q}}(\mathcal{P}) = \{\mathcal{P}, \mathcal{Q}\}$.*

THEOREM 2.4. *Let the integers $L \geq 1$ and $p \geq -1$ be given. Let Hamiltonian* (1.1) *be given. Let $\{\mathcal{P}_i\}_{i=0}^{L}$ be the sequence of the linear spaces of all homogeneous polynomials of degree $i + p + 2$ in $\boldsymbol{x} \in \boldsymbol{R}^{2n}$. Let $\{\mathcal{Q}_i\}_{i=1}^{L}$ be the sequence of some subsets of the sets $\mathcal{P}_i$ and let $\{\mathcal{R}_i\}_{i=1}^{L}$ be the sequence of some linear spaces of $\mathcal{C}^k$-functions ($k \geq 0$) for $i = 1, \ldots, L$. Let $\mathcal{G}$ be a polynomial in $\mathcal{P}_j$ for some $j \in \{0, \ldots, L\}$. Let $\Omega \subseteq \boldsymbol{R}^{2n}$ be the common domain where the sequences $\{\mathcal{P}_i\}_{i=0}^{L}$, $\{\mathcal{Q}_i\}_{i=1}^{L}$, and $\{\mathcal{R}_i\}_{i=1}^{L}$ are defined. Moreover, suppose that the following properties are satisfied:*

(i) $\mathcal{H}_i \in \mathcal{P}_i$, $i = 0, 1, \ldots, L$;
(ii) $\{\mathcal{P}_i, \mathcal{R}_j\} \subseteq \mathcal{P}_{i+j}$, $i + j = 1, \ldots, L$;
(iii) *for any $D \in \mathcal{P}_i$, $i = 1, \ldots, L$, there exist $E \in \mathcal{Q}_i$ and $F \in \mathcal{R}_i$ such that*

$$E = D + \{\mathcal{H}_0, F\} \quad \text{and} \quad \{\mathcal{G}, E\} = 0.$$

Then, there exists a $\mathcal{C}^k$-function

$$\mathcal{W}(\boldsymbol{x}; \varepsilon) = \sum_{i=0}^{L-1} \frac{\varepsilon^i}{i!} \mathcal{W}_{i+1}(\boldsymbol{x})$$

with $\mathcal{W}_i \in \mathcal{R}_i$, $i = 1, \ldots, L$, such that the change of coordinates $\boldsymbol{x} = \boldsymbol{X}(\boldsymbol{y}; \varepsilon)$ is the general solution of the initial value problem

$$\begin{aligned} \frac{d\,\boldsymbol{x}}{d\,\varepsilon} &= \mathcal{J} \frac{\partial \mathcal{W}}{\partial \boldsymbol{x}}(\boldsymbol{x}; \varepsilon), \\ \boldsymbol{x}(0) &= \boldsymbol{y}, \end{aligned}$$

and transforms the convergent Hamiltonian (1.1) *into the convergent Hamiltonian*

$$\mathcal{K}(\boldsymbol{y}; \varepsilon) = \sum_{i=0}^{L} \frac{\varepsilon^i}{i!} \mathcal{K}_i(\boldsymbol{y}) + \mathcal{O}(\varepsilon^{L+1})$$

with $\mathcal{K}_i \in \mathcal{Q}_i$, $i = 1, \ldots, L$, *such that each* $\mathcal{K}_i$ *is a polynomial in* $\boldsymbol{y}$ *of degree* $i + p + 2$ *with* $\{\mathcal{K}_i, \mathcal{G}\} = 0$ *for* $i = 1, \ldots, L$.

Additionally, if $\{\mathcal{H}_0, \mathcal{G}\} = 0$, *then* $\mathcal{G}$ *is an integral of* $\mathcal{K}$.

Proof. See [26]. □

The smoothness of the spaces $\mathcal{R}_i$ and of $\mathcal{W}$ depends on the possible appearance of nonpolynomial terms at some order i. For instance, in the example of section 4.1, $\mathcal{W}$ is $\mathcal{C}^2$, while in section 4.2 it is analytic.

As in the standard approach, the application of Theorem 2.4 does not produce generally a convergent normal form Hamiltonian. Nevertheless, the reader should notice that in the definition of $\mathcal{K}$ we are putting higher-order terms inside the error $\mathcal{O}(\varepsilon^{L+1})$.

The reader should notice that whenever $\mathcal{G}$ is an integral of $\mathcal{H}_0$, the effect of constructing $\mathcal{K}$, where $\mathcal{K}_i \in \ker(\mathcal{L}_\mathcal{G})$ (i.e., the Poisson bracket $\{\mathcal{K}_i, \mathcal{G}\}$ vanishes) for $i = 1, \ldots, L$, is to extend formally the integral of the unperturbed system, $\mathcal{G}$, to the whole Hamiltonian $\mathcal{K}$ after truncating it at order L.

We stress that the integral $\mathcal{G}$ of $\mathcal{H}_0$ must be chosen a priori and it can be other than the part of $\mathcal{H}_0$ whose associated matrix is semisimple. This is the main difference with respect to the usual treatment of normal forms, although Theorem 2.4 can be interpreted as a generalization of the standard approach. In fact, under further hypotheses, if $\mathcal{G}$ corresponds to the semisimple part of $\mathcal{H}_0$, both normal form transformations yield the same results; see the details in [26, 28]. Our way of computing formal integrals extends the classical approaches of Whittaker [33, 34], Gustavson [13], and Giorgilli [12] as we can obtain approximate integrals of polynomial Hamiltonians in n DOFs, with $\mathcal{H}_0$ being any homogeneous (not necessarily quadratic) polynomial.

Theorem 2.4 can be applied to calculate formal integrals of polynomial Hamiltonians whose dominant parts are related to nilpotent matrices A. Furthermore, this approach can be used to calculate different normalized Hamiltonians whose flows lie on different reduced phase spaces. Thus, performing several reductions allows us to analyze the original Hamiltonian from different points of view. The theorem also can be applied to lower the number of DOFs by two or more units. Indeed, the number of DOFs of the resulting Hamiltonian system depends on the type of integral introduced through the reduction. Nevertheless, this needs a careful analysis based on Lie groups and invariant theory; see [28] for details.

The drawback of this generalized method is that at each step $i \geq 1$, $\mathcal{W}_i$ is not necessarily a polynomial function, as is always the case with the application of the normal form theorem. Indeed, the occurrence of polynomial generating functions in a specific normal form computation can be known in advance by analyzing the dimensions of the kernels of various linear spaces of homogeneous polynomials. Thus, depending on the choice of $\mathcal{G}$, as well as on $\mathcal{H}_0$ and the type of terms in the perturbation, $\mathcal{W}$ could be a polynomial of degree $i + 2$. (In this optimal case the linear spaces $\mathcal{R}_i$ correspond to the spaces of homogeneous polynomials of degree $i + 2$ in $\boldsymbol{x} \in \boldsymbol{R}^{2n}$.) For more information, see [26, 28]. Nevertheless, in the design of the algorithm we do not perform this algebraic analysis prior to the computation of the normal form. Instead, we prefer to try first to solve the homological equation with a homogeneous polynomial. When it is not possible we resort to finding the solution of the usual partial differential equation (PDE). This will be a fundamental step in the design of the algorithms presented in the next section.

2.3. Types of normalization according to $\mathcal{H}_0$. The cases where $\mathcal{H}_0$ is not a quadratic polynomial cannot be treated by the normal form theorem and we have to

resort to Theorem 2.4 for its reduction. The choice of the integral $\mathcal{G}$ to be extended depends on each problem and can be, for instance, $\mathcal{H}_0$, as we will see in section 4.1.

If $\mathcal{H}_0$ is a quadratic homogeneous polynomial we perform the following classification according to the type of the associated matrix A:

(a) semisimple case: $A = S$;
(b) semisimple plus nilpotent case: $A = S + N$, with $S, N \neq 0$;
(c) nilpotent case: $A = N$.

In cases (a) and (b), as the semisimple part is not zero, both Theorem 2.1 and Theorem 2.4 apply. Let us see when one should apply the first and when the second.

On the one hand, by applying the normal form theorem, $\mathcal{I}_S$ (the semisimple part of $\mathcal{H}_0$) becomes a formal integral of the reduced system. This would be equivalent to choosing $\mathcal{G} = \mathcal{I}_S$ and applying Theorem 2.4. Nevertheless, the reduced Hamiltonians in both cases may not be the same [26]. In this situation the application of the normal form theorem would be preferable because the generating function is polynomial and may not be polynomial with Theorem 2.4.

On the other hand, when the application of the normal form theorem leads to $\mathcal{K}$ defining a system of 0 DOFs ($n > 1$), the reduction does not allow one to extract information from the flow of the reduced system. Then, the application of the generalized normal form theorem by choosing $\mathcal{G}$ an integral of $\mathcal{H}_0$ different from $\mathcal{I}_S$ is recommended.

Moreover, in both situations, when one intends to obtain information about the original system from different reduced Hamiltonians, one should resort to Theorem 2.4, as the normal form theorem allows only one reduction.

In case (c), as the semisimple part is zero the application of the normal form theorem does not lead to a reduction in the number of DOFs. Thus, for this purpose we use the generalized method by taking $\mathcal{G} = \mathcal{H}_0$, for instance. Obviously, other integrals could be chosen depending on the system. Besides, if $\mathcal{H}_0 = 0$, we could even select $\mathcal{G}$ among the integrals of $\mathcal{H}_k$, provided that $\mathcal{H}_i = 0$ for $i = 0, \ldots, k-1$, with $k \geq 1$.

Once we have made the choice of $\mathcal{G}$ the next step consists in performing the normal form transformation, i.e., in calculating $\mathcal{K}$ and $\mathcal{W}$ up to order L of the process. This is done through Lie transformations and will be treated algorithmically in section 3.

Next we calculate the direct and inverse changes of coordinates related to the generalized normal form transformation. This is useful when going back to the original Hamiltonian. For example, one can think of a qualitative study of a certain Hamilton vector field based on the calculation of normal forms. In this situation it is usually convenient to get the expressions in the original coordinates of the invariant objects of the original system obtained from the ones in the normal form Hamiltonians.

Finally, it is possible to determine a formal integral of the Hamiltonian $\mathcal{H}$ using $\mathcal{G}$, that is, the function extended to become an integral of $\mathcal{K}$, and the generating function related to the normal form construction. Specifically, if $\mathcal{G}$ commutes with $\mathcal{K}$, the result of applying the inverse change of coordinates to $\mathcal{G}$, which we denote by $\mathcal{I}_{\mathcal{G}}(\boldsymbol{x}; \varepsilon)$, is a formal symmetry of $\mathcal{H}$, i.e., $\{\mathcal{H}, \mathcal{I}_{\mathcal{G}}\} = \mathcal{O}(\varepsilon^{L+1})$. Note that if $\mathcal{W}$ is of polynomial nature, $\mathcal{I}_{\mathcal{G}}$ is also a polynomial. However, when $\mathcal{W}$ contains nonpolynomial terms, $\mathcal{I}_{\mathcal{G}}$ will be in general nonpolynomial.

The direct and inverse changes of coordinates will be explained in section 3.4 from the point of view of the computations which are needed.

For a complete description of the methodology and some examples, see [25, 26, 27] and references therein.

3. The algorithms. This section shows the main features of the routines we have developed so as to apply the reduction process to an arbitrary n-DOF polynomial Hamiltonian of the type (1.1) up to a certain order of approximation. We used MATHEMATICA to write the routines, but other symbolic processors, such as MAPLE or MACSYMA, would be also adequate. We present the entire collection of routines to give a full description of the normal form procedure for polynomial Hamiltonians, emphasizing the main features of the implementation. In addition, our exposition provides the computational cost of all steps and, to our knowledge, it is the first time all these issues appear together in one paper.

3.1. Choice of the integral $\mathcal{G}$. According to $\mathcal{H}_0$ and our purposes we start by choosing the type of normal form we are going to calculate. If $\mathcal{H}_0$ is quadratic we implement an auxiliary procedure to split this Hamilton function into its semisimple and nilpotent parts. The so-called Jordan decomposition of the associated Hamiltonian matrix A is the algebraic tool which provides us with this result.

ALGORITHM 1 (choice of the integral to be extended).

INPUT:

↪ $\mathcal{H}_0$ *and* p*: the main (or unperturbed) part of* $\mathcal{H}$ *and the degree of* $\mathcal{H}_0$ *minus two, respectively;*

↪ n*: number of DOFs of the system;*

↪ `option` $\in \{\{\texttt{no}\}, \{\texttt{yes}, \mathcal{G}\}\}$*: chance of either allowing the program to choose the integral to be extended or the user to choose a formal integral* $\mathcal{G}$ *of* $\mathcal{H}_0$*, resp. In case* $\mathcal{H}_0$ *is not quadratic the user should choose* $\mathcal{G}$*;*

↪ `ForceNFT`*: Boolean variable which indicates whether the user wants to apply the normal form theorem.*

OUTPUT:

↩ $\mathcal{I}_S, \mathcal{I}_N$*: decomposition of* $\mathcal{H}_0$ *into its semisimple and nilpotent parts, resp. (when* $\mathcal{H}_0$ *is quadratic, otherwise both are taken to be zero);*

↩ $\mathcal{G}$*: integral of* $\mathcal{H}_0$ *to be extended to higher orders;*

↩ `NFT`*: it takes the value* `True` *if the normal form theorem is going to be used or* `False` *otherwise.*

CODE:

`NFT` = `False`;
If ($p = 0$) *then*
  Compute the Jordan decomposition of the Hamiltonian matrix A *associated with* $\mathcal{H}_0$*, that is,* $A = S + N$*;*
  Construct $\mathcal{I}_S, \mathcal{I}_N$ *from the preceding step;*
  If (`ForceNFT` = `True`) *then*
    If ($\mathcal{I}_S = 0$) *show the message "After the normalization,* $\mathcal{K}$ *will define a system with the same number of DOFs as* $\mathcal{H}$*";*
    Calculate $\mathcal{G}$ *as in the normal form theorem (*$\mathcal{G} = \mathcal{K}_0^t$*);*
    `NFT` = `True`;
  Else
    If (`option` = `no`) *then*
      Compute $\mathcal{G}$ *following the guidelines of section* 2.3.
    Else
      Check if $\mathcal{G}$ *is an integral of* $\mathcal{H}_0$*. When* $\mathcal{G}$ *does not commute with* $\mathcal{H}_0$ *(i.e.,* $\{\mathcal{G}, \mathcal{H}_0\} \neq 0$*), a warning message is displayed saying that no integral will be introduced through the transformation;*
Else

Set $\mathcal{I}_S = \mathcal{I}_N = 0$*;*
Define $\mathcal{G}$ *as the second component of* $\{\texttt{yes}, \mathcal{G}\}$.

Remark 3.1. The algorithm allows the user to normalize Hamiltonians satisfying $\mathcal{H}_i = 0$, for $i = 1, \ldots, k-1$ and $\mathcal{H}_k \neq 0$; the input variable `option` brings such possibility by means of giving it the value $\{\texttt{yes}, \mathcal{G}\}$ and introducing any homogeneous polynomial which commutes with $\mathcal{H}_k$ as the second component, $\mathcal{G}$.

3.2. Resolution of the homological equation. The construction of the transformed Hamiltonian $\mathcal{K}$ is done order by order, i.e., one has to proceed in an ascendant way from $i = 1$ to $i = L$ so as to determine each $\mathcal{K}_i$. For that, the homological equation

$$\mathcal{L}_{\mathcal{H}_0}(\mathcal{W}_i) + \mathcal{K}_i = \bar{\mathcal{H}}_i \tag{3.1}$$

has to be solved, with the extra condition $\{\mathcal{K}_i, \mathcal{G}\} = 0$ for $i = 1, \ldots, L$. Note that the terms $\bar{\mathcal{H}}_i$ are the ones known from the previous orders through the Poisson bracket calculations. The solution of (3.1) is the pair $(\mathcal{W}_i, \mathcal{K}_i)$, where $\mathcal{W}_i$ denotes the generating function at order i.

To actually solve (3.1) we split $\bar{\mathcal{H}}_i$ as $\bar{\mathcal{H}}_i = \bar{\mathcal{H}}_i^* + \bar{\mathcal{H}}_i^{\#}$, where $\bar{\mathcal{H}}_i^* \in \ker(\mathcal{L}_{\mathcal{G}})$ and $\bar{\mathcal{H}}_i^{\#} = \bar{\mathcal{H}}_i - \bar{\mathcal{H}}_i^*$, for each $i = 1, \ldots, L$. In this way, we choose $\mathcal{K}_i = \bar{\mathcal{H}}_i^*$ and $\mathcal{W}_i$ as a solution of

$$\mathcal{L}_{\mathcal{H}_0}(\mathcal{W}_i) = \bar{\mathcal{H}}_i^{\#}. \tag{3.2}$$

When Theorem 2.4 is applied, for the general case nonpolynomial terms appear in $\bar{\mathcal{H}}_i$ at a certain order i. To solve (3.1) these terms are always considered to belong to $\bar{\mathcal{H}}_i^{\#}$. Thus, in the program we introduce a simple algorithm to separate the polynomial terms of $\bar{\mathcal{H}}_i$ from the nonpolynomial ones.

ALGORITHM 2 (splitting in polynomial and nonpolynomial terms).
INPUT:
↪ `Expr`*: a function in the coordinates and respective momenta;*
↪ *n: number of DOFs of the system.*
OUTPUT:
↩ `Pol` *and* `NoPol`*:* `Pol` *collects the polynomial terms of* `Expr` *while* `NoPol` *collects the nonpolynomial ones.*

Remark 3.2. (i) The only nonpolynomial terms are rational, logarithmic, arctangent functions and combinations of them. They appear in the generating function at a certain order as the solution of (3.2) and are propagated through the intermediate Hamiltonians needed to proceed to higher orders (see also section 3.3).

(ii) Every symbolic processor has its own functions to distinguish between polynomial and nonpolynomial terms. For example, MATHEMATICA makes use of the Boolean built-in function `PolynomialQ` which recognizes whether a single term is polynomial in a given set of variables. Thus, the issue of splitting the polynomial terms from the others becomes a simple matter.

Now, we describe the main algorithm of this section—the one for the resolution of (3.1). We suppose that Algorithm 1 has been already applied.

ALGORITHM 3 (resolution of the homological equation).
INPUT:
↪ *i: step in the normalizing procedure;*
↪ $\bar{\mathcal{H}}_i$*: terms computed in order* $i-1$ *through the Lie transformation;*
↪ $\mathcal{G}$, `NFT`, *and s: the latter denoting the degree of* $\mathcal{G}$*;*

OUTPUT:
↩ *The pair* $(\mathcal{K}_i, \mathcal{W}_i)$*: solution of* (3.1)*;*
↩ `resolved`*: it takes the value* `True` *if* (3.1) *has been solved successfully.*
CODE:
Split $\bar{\mathcal{H}}_i$ *into its polynomial* (`Pol`) *and nonpolynomial* (`NoPol`) *components by making use of Algorithm 2;*
Redefine $\bar{\mathcal{H}}_i = \mathtt{Pol}$*;*
Define $\mathtt{arb}_1$*, an arbitrary polynomial of degree* $i+p+2$ *in* $\boldsymbol{x}$*, and solve the equation* $\mathcal{L}_{\mathcal{G}}(\mathtt{arb}_1) = 0$ *in the coefficients of* $\mathtt{arb}_1$*; store the resulting conditions for the coefficients of* $\mathtt{arb}_1$ *in a new variable called* $\mathtt{sol}_\mathtt{p}$*;*
Define $\mathtt{arb}_2$*, a homogeneous polynomial of degree* $i+2$ *in* $\boldsymbol{x}$*;*
If ($\mathtt{NoPol} = 0$) *then*
    Try to solve $\{\mathcal{H}_0, \mathtt{arb}_2\} + \mathtt{arb}_1 = \bar{\mathcal{H}}_i$ *in the coefficients of* $\mathtt{arb}_1$ *and* $\mathtt{arb}_2$*, including the restrictions given by* $\mathtt{sol}_\mathtt{p}$*. The solution, if any, is stored in* $\mathtt{sol} = (\mathtt{arb}_1, \mathtt{arb}_2)$*;*
Else
    $\mathtt{sol} = \emptyset$*;*
If ($\mathtt{sol} \neq \emptyset$) *then*
    Set $(\mathcal{K}_i, \mathcal{W}_i) = \mathtt{sol}$ *and* $\mathtt{resolved} = \mathtt{True}$*;*
Else
    Define $\mathcal{K}_i$ *as the terms of* $\bar{\mathcal{H}}_i$ *that are in* $\ker(\mathcal{L}_{\mathcal{G}})$*;*
    Try to solve the PDE $\{\mathcal{H}_0, \mathcal{W}_i\} + \mathcal{K}_i = \bar{\mathcal{H}}_i + \mathtt{NoPol}$ *in the unique unknown* $\mathcal{W}_i$*. We call the solution, if any,* `Dsol`*;*
    If (`Dsol` *is computed properly*) *then*
        $\mathtt{resolved} = \mathtt{True}$*;*
    Else
        An error message telling that $\mathcal{W}_i$ *cannot be obtained is displayed; in addtion* $\mathtt{resolved} = \mathtt{False}$ *and the algorithm is aborted.*

Remark 3.3. (i) If the number of DOFs of our problem is n and we are at order i, then the number of different coefficients of $\mathtt{arb}_1$ is $\binom{2n+i+p+1}{i+p+2}$. Besides, since the degrees of $\mathcal{G}$ and $\mathtt{arb}_1$ are, respectively, s and $i+p+2$, then $\mathcal{L}_{\mathcal{G}}(\mathtt{arb}_1)$ is a homogeneous polynomial of degree $i+p+s$ in $\boldsymbol{x}$. Hence, the solution of $\mathcal{L}_{\mathcal{G}}(\mathtt{arb}_1) = 0$ is obtained by matching the coefficients of $\mathtt{arb}_1$ so that the latter expression becomes identically zero. Therefore, one needs to solve a system of $\binom{2n+i+p+1}{i+p+2}$ linear homogeneous equations with $\binom{2n+s-1}{s}$ unknowns. This system always has an infinite number of nontrivial solutions provided that the number of unknowns is smaller than or equal to the number of linear equations; this is satisfied whenever $s \leq i+p+2$. The solution is the one denoted by $\mathtt{sol}_\mathtt{p}$.

(ii) When $\mathtt{NoPol} = 0$ we try to resolve the PDE: $\mathcal{L}_{\mathcal{H}_0}(\mathtt{arb}_2) = \bar{\mathcal{H}}_i - \mathtt{arb}_1$ for $\mathtt{arb}_2$ (once the conditions for $\mathtt{arb}_1$ are imposed) by matching the coefficients of $\mathtt{arb}_2$. Since we try to get $\mathtt{arb}_2$ as a polynomial in $\boldsymbol{x}$, $\bar{\mathcal{H}}_i$ and $\mathtt{arb}_1$ are homogeneous polynomials of degree $i+p+2$ and the degree of $\mathcal{H}_0$ is $p+2$, the degree of $\mathtt{arb}_2$ should be $i+2$, and hence, the number of different coefficients for $\mathtt{arb}_2$ is $\binom{2n+i+1}{i+2}$. Now, the solution of the PDE is achieved by solving a system of $\binom{2n+i+p+1}{i+p+2}$ linear equations with $\binom{2n+i+1}{i+2}$ unknowns. We store the solution on the variable `sol`. Unlike what occurs in Remark 3.3(i), this system does not always have a solution for $\mathtt{arb}_2$. In this case $\mathtt{sol} = \emptyset$ (we are taking $\emptyset$ as the empty subset) and the algorithm tries to solve $\mathcal{L}_{\mathcal{H}_0}(\mathcal{W}_i) = \bar{\mathcal{H}}_i{}^{\#}$ as a PDE to determine $\mathcal{W}_i$. For that we make use of the MATHEMATICA function `DSolve`. This second situation is worse than the first in terms of computational cost.

Indeed, there are a few examples where even the command `DSolve` does not return the desired result. In these cases, a suitable change of coordinates could sort out the trouble.

(iii) If $\mathcal{W}_i$ admits a polynomial expression, that is, when $\texttt{sol} \neq \emptyset$, the undetermined coefficients of $\texttt{arb}_1$ and $\texttt{arb}_2$ are settled to zero, so as to avoid an increase of unnecessary monomials in $\mathcal{K}_i$ and $\mathcal{W}_i$.

(iv) During the process we express $\mathcal{K}_i$ in Cartesian coordinates. However, for some systems, from a computational point of view, it is usually convenient to use an adequate set of symplectic variables such that the homological equation presents a "nicer" aspect, for example, spherical or complex variables. Sometimes it is better to use a combination of them. See examples of this in [27].

(v) Every time Algorithm 3 is applied, a message showing us if the homological equation is solved correctly at order i is displayed. Additionally, we are informed whether the generating function is a polynomial.

(vi) Taking $i = 0$ at all steps and considering also that $p = 0$, the algorithm computes versal deformations of linear Hamiltonian vector fields; see, for instance, [11]. This feature is included in the next subsection.

(vii) A little algorithm to compute Poisson brackets should be added to Algorithm 3. In MATHEMATICA it is a simple matter to compute the Poisson bracket of two scalar functions $\mathcal{P}, \mathcal{Q} : \Omega \subseteq \boldsymbol{R}^{2n} \to \boldsymbol{R}$ by using the built-in function for derivatives `D`.

(viii) Tests to check if the algorithm works properly are introduced at several steps in the procedure. For instance, when redefining $\bar{\mathcal{H}}_i$ as a polynomial we introduce a line in the code to check if $\bar{\mathcal{H}}_i$ is in fact a homogeneous polynomial in $\boldsymbol{x} = (x_1, \ldots, x_n, X_1, \ldots, X_n)$. If it is not, the algorithm is aborted and an error message is shown.

3.3. Lie–Deprit method. Once the order we want to reach in the transformation, say L, is fixed, we are ready to proceed to the construction of both $\mathcal{K}$ and $\mathcal{W}$. The way to obtain them is by going ascendantly order by order from $i = 1$ to $i = L$. Notice that we need to calculate $\mathcal{K}_i$ for $i = 1, \ldots, L$, such that

$$\mathcal{K}(\boldsymbol{y}; \varepsilon) \,=\, \sum_{i=0}^{L} \frac{\varepsilon^i}{i!}\, \mathcal{K}_i(\boldsymbol{y}),$$

where each $\mathcal{K}_i$ is a polynomial in $\boldsymbol{y}$ of degree $i+p+2$ and $\{\mathcal{K}_i, \mathcal{G}\} = 0$ for $i = 1, \ldots, L$. Besides, we need to obtain

$$\mathcal{W}(\boldsymbol{x}; \varepsilon) \,=\, \sum_{i=0}^{L-1} \frac{\varepsilon^i}{i!}\, \mathcal{W}_{i+1}(\boldsymbol{x}),$$

and at each order i we will make use of Algorithm 3 to get $\mathcal{K}_i$ and $\mathcal{W}_i$. Indeed, the advance from one order to the other is done through the recursion formula

$$\mathcal{H}_{j,k} \,=\, \mathcal{H}_{j+1,k-1} + \sum_{\ell=0}^{j} \binom{j}{\ell} \{\mathcal{H}_{j-\ell,k-1}, \mathcal{W}_{\ell+1}\} \tag{3.3}$$

for $j \geq 0$ and $k \geq 1$. Here it is assumed that $\mathcal{H}_{i,0} = \mathcal{H}_i$ and $\mathcal{H}_{0,i} = \mathcal{K}_i$ for all $i \geq 0$. At the end of the process one has computed the Hamiltonians $\mathcal{H}_{j,k}$ with $0 \leq j \leq L$, $1 \leq k \leq L$, and $j + k \leq L$. Hence, the intermediate Hamiltonians form the so-called

Lie–Deprit triangle. See Deprit [6] for the original version of the method and for different applications [2, 22, 5, 26, 27, 10].

ALGORITHM 4 (Lie–Deprit method).

INPUT:

↪ *L: number of steps to be performed in the Lie–Deprit process;*
↪ *n: number of DOFs of the vector field related to $\mathcal{H}$;*
↪ *$\mathcal{H}_0, \ldots, \mathcal{H}_L$: the first, second, $\ldots, L+1$th terms of the Hamiltonian $\mathcal{H}$;*
↪ `VDef`*: Boolean variable which offers the user the possibility of carrying out a transformation within linear versal deformations.*

OUTPUT:

↩ *$(\mathcal{K}_i, \mathcal{W}_i)$: for $i = 1, \ldots, L$.*

CODE:

If (`VDef`) *then*
  Set the auxiliary variable $d_{\mathtt{aux}} = 0$*;*
Else
  Set $d_{\mathtt{aux}} = 1$*;*
For d from 1 to L do
  Define $\mathcal{H}_{d,0} = \mathtt{aux}_1 = \mathcal{H}_d$*;*
  Set $g_d = 0$*;*
  For j from $d-1$ to 0 do

$$\textit{Compute } \mathcal{H}_{j,d-j} = \mathtt{aux}_1 = \mathtt{aux}_1 + \sum_{\ell=0}^{j} \binom{j}{\ell} \{\mathcal{H}_{j-\ell,d-j-1}, g_{\ell+1}\};$$

  $(\mathcal{H}_{0,d}, \mathtt{aux}_2)$ *= solution of applying Algorithm 3 to* $\mathtt{aux}_1$ *(with index* $i = d \cdot d_{\mathtt{aux}}$*);*
  Set $g_d = -\mathtt{aux}_2$*;*
  Or
    $d = 1$*;*
    $\mathtt{aux}_2 = \mathcal{H}_{0,d} - \mathtt{aux}_1$*;*
    For j from $d-1$ to 1 do
      $\mathcal{H}_{d-j,j} = \mathcal{H}_{d-j,j} + \mathtt{aux}_2$*;*
Set $\mathcal{K}_i = \mathcal{H}_{0,i}$*, for* $i = 0, \ldots, L$ *and* $\mathcal{W}_i = g_i$ *for* $i = 1, \ldots, L$.

Remark 3.4. (i) The number of Poisson brackets needed to achieve the Lie–Deprit method is $\binom{L+2}{3}$.

(ii) Since at each order i, according to Remark 3.2(i) and (ii), one must solve a system of linear equations with $\binom{2n+s-1}{s}$ variables plus another system with $\binom{2n+i+1}{i+2}$ variables, after completing the entire process of the Lie transformation, one has solved $2\,L$ systems of linear equations. The total number of unknowns which have been determined is $\binom{2n+L+1}{L+2} + L\binom{2n+s-1}{s} - (2n+1)(n+1)$.

3.4. The algorithm of the inverse for Lie transformations. Once the normal form transformation has been obtained, a further step consists in calculating the explicit expressions of the changes of variables relating the original coordinates $\boldsymbol{x}$ with the transformed $\boldsymbol{y}$. In other words, if the direct change, called $\boldsymbol{x} = \boldsymbol{X}(\boldsymbol{y}; \varepsilon)$, is responsible for writing the "old" coordinates $\boldsymbol{x}$ in terms of the "new" $\boldsymbol{y}$, we need to find an expression for $\boldsymbol{X}$. Equivalently, if the inverse change puts the new coordinates $\boldsymbol{y}$ in terms of the old $\boldsymbol{x}$ and this change is denoted by $\boldsymbol{y} = \boldsymbol{Y}(\boldsymbol{x}; \varepsilon)$, we have to look for an expression of $\boldsymbol{Y}$.

For this purpose we have designed Algorithms 5 and 6. In particular, the corresponding codes have been developed according to the algorithm of the inverse for Lie

transformations due to Henrard [16]. The method proposed in that paper reduces the amount of computations compared with the usual method for the direct and inverse changes proposed by Deprit [6]. For example, in the Lie transformations utilized in the analytical lunar theory [7, 8], the number of Poisson brackets evaluated was reduced by a factor of three.

We give a brief description of Henrard's method [16]. Notice that its purpose is a bit wider than the matter of constructing the direct and inverse changes of coordinates. For instance, we can use the method proposed by Henrard to write any function expressed in terms of the variables $\boldsymbol{x}$ as a function of the variables $\boldsymbol{y}$ and vice versa, provided only that we know the generating function $\mathcal{W}$.

The coefficients of the transformed function

$$g(\boldsymbol{y};\varepsilon) \,=\, \sum_{i\geq 0} \frac{\varepsilon^i}{i!}\, g_i(\boldsymbol{y}),$$

corresponding to a certain smooth function

$$f(\boldsymbol{x};\varepsilon) \,=\, \sum_{i\geq 0} \frac{\varepsilon^i}{i!}\, f_i(\boldsymbol{x}),$$

under the inverse of the transformation generated by the smooth function

$$\mathcal{V}(\boldsymbol{x};\varepsilon) \,=\, \sum_{i\geq 0} \frac{\varepsilon^i}{i!}\, \mathcal{V}_{i+1}(\boldsymbol{x}),$$

are computed recursively by the formula

$$g_i \,=\, \sum_{j=0}^{i} \binom{i}{j} f_{j,i-j}. \tag{3.4}$$

The intermediate terms $f_{j,k}$ are calculated taking into account that

$$f_{0,k} \,=\, f_k, \qquad f_{j,k} \,=\, -\sum_{\ell=1}^{j} \binom{j-1}{\ell-1} \{f_{j-\ell,k}, \mathcal{V}_\ell\} \quad \text{for} \quad j>0.$$

Now we can use the above to calculate the direct and inverse changes as follows.

ALGORITHM 5 (direct change).

INPUT:

↪ *L: number of steps of the transformation process;*

↪ *n: number of DOFs of the system;*

↪ *$f_0, \ldots, f_L$: the first, second, $\ldots$, $L+1$th terms of a function f of $\boldsymbol{x}$;*

↪ *$\mathcal{W}_1, \ldots, \mathcal{W}_L$: the first, second, $\ldots$, Lth terms of the generating function $\mathcal{W}$.*

OUTPUT:

↩ *g: the expression of f as a function of $\boldsymbol{y}$ up to an approximation of order $\mathcal{O}(\varepsilon^{L+1})$.*

CODE:

Define $g_0 = g_{0,0} = f_0$*;*

For d *from* 0 *to* $L-1$ *do*

For j *from* 1 *to* $L-d$ *do*

$$g_{j,d} \,=\, -\sum_{k=1}^{j} \binom{j-1}{k-1} \{g_{j-k,d}, \mathcal{W}_k\};$$

$$g_{d+1} = f_{d+1} - \sum_{\ell=1}^{d+1} \binom{d+1}{\ell} g_{\ell,d+1-\ell};$$

Compute $g = \sum_{i=0}^{L} \frac{\varepsilon^i}{i!} g_i$.

Remark 3.5. (i) For the direct change of coordinates we need to apply Algorithm 5, replacing f_0 by each component of $\boldsymbol{x}$ and putting $f_i = 0$ for $1 \leq i \leq L$.

(ii) To put a function g in terms of another function f one needs to calculate $\binom{L+2}{3}$ Poisson brackets.

The advantage of the algorithm of the inverse is that it can be used backward as well as forward. More precisely, using (3.4) one can express the terms f_i as functions of the g_j as follows:

$$f_i = g_i - \sum_{j=1}^{i} \binom{i}{j} f_{j,i-j}. \tag{3.5}$$

Thus, we modify Algorithm 5 accordingly.

ALGORITHM 6 (inverse change).

INPUT:

↪ *L: the number of steps of the transformation process;*

↪ *n: number of DOFs of the system;*

↪ *$g_0, \ldots, g_L$: the first, second, $\ldots, L+1$th terms of a function g of $\boldsymbol{y}$;*

↪ *$\mathcal{W}_1, \ldots, \mathcal{W}_L$: the first, second, $\ldots, L$th terms of the generating function $\mathcal{W}$.*

OUTPUT:

↩ *f: the expression of g as a function of $\boldsymbol{x}$ up to an approximation of order $\mathcal{O}(\varepsilon^{L+1})$.*

CODE:

For d from 0 to L do

Define $g_{0,d} = g_d$;

If $g_d \neq 0$ then

For ℓ from 1 to $L - d$ do

$$g_{\ell,d} = -\sum_{k=1}^{\ell} \binom{\ell-1}{k-1} \{g_{\ell-k,d}, \mathcal{W}_k\};$$

Else

$g_{\ell,d} = 0$ *for* $\ell = 1, \ldots, L-d$;

$$f_d = \sum_{\ell=0}^{d} \binom{d}{\ell} g_{\ell,d-\ell};$$

Compute $f = \sum_{i=0}^{L} \frac{\varepsilon^i}{i!} f_i$.

Remark 3.6. (i) The inverse change of coordinates is obtained after application of Algorithm 6 substituting g_0 by each component of $\boldsymbol{y}$ and setting $g_i = 0$ for $1 \leq i \leq L$.

(ii) The first "if" command of Algorithm 6 is due to the advantageous fact that each column of $g_{i,d}$ for $i = 0, \ldots, L-d$ is independent from the others and it is computed recursively from the first to the last element. Hence, if the first element of any column vanishes, the entire column vanishes as well. Thereby, in some particular but typical situations, only some of these columns have to be computed. Consequently, the number of operations is reduced drastically.

(iii) Given a Hamilton function $\mathcal{H}$ together with its normal form Hamiltonian $\mathcal{K}$ and the generating function $\mathcal{W}$ constructed such that $\mathcal{K}_i$ commutes with $\mathcal{G}$ for

$i \in \{0, \ldots, L\}$, a formal integral of $\mathcal{H}$ can be built up with the use of Algorithm 6, after replacing $\mathcal{G}$ by g_0 and setting $g_i = 0$ for $1 \le i \le L$. Thus, the formal integral $\mathcal{I}_{\mathcal{G}}$ will be the output function f.

(iv) The number of Poisson brackets used in the computation of a function g as a function of f is $\binom{L+2}{3}$. However, if g is such that $g_i = 0$ for $1 \le i \le L$, as it occurs for situations (i) and (iii), the number of Poisson brackets to be evaluated gets reduced to $\binom{L+1}{2}$, due to the consideration explained in (ii).

4. Applications. This section displays the features of the algorithm exposed in the preceding paragraphs through two examples.

4.1. Normalization of a Hamiltonian with null quadratic terms. First we have chosen a polynomial Hamiltonian where the first nonnull term is a homogeneous polynomial of degree 3.

In this application we assume that the vector $\boldsymbol{x}$ is two-dimensional, i.e., $\boldsymbol{x} = (x, X)$; the lowercase x stands for the position, whereas the uppercase X refers to its associated momentum.

We start by defining the Hamilton function through

$$\mathcal{H}(\boldsymbol{x}) = \sum_{i=0}^{\infty} \frac{1}{i!}\, \mathcal{H}_i(\boldsymbol{x}), \tag{4.1}$$

where

$$\begin{aligned} \mathcal{H}_0(\boldsymbol{x}) &= x^2 X, \\ \mathcal{H}_i(\boldsymbol{x}) &= x^2 \sum_{j=1}^{i+2} h_{i,j}\, x^{i+2-j} X^{j-1}, \qquad i = 1, \ldots, \infty, \end{aligned} \tag{4.2}$$

and $h_{i,j} \in \boldsymbol{R}$. Note that the vector field associated with $\mathcal{H}$ is a 1-DOF system. With the notation introduced in section 2 we have that $A = (0)$, that is, the square matrix of order two whose entries are all zero. In this situation, it is clear that the normal form theorem, Theorem 2.1, does not produce any change in $\mathcal{H}$ because the quadratic part is zero. As a consequence, if we want to simplify (4.1) we need to change to the setting of generalized normal forms.

RESULT 4.1. *We select $\mathcal{G}(\boldsymbol{x}) = \mathcal{H}_0(\boldsymbol{x})$ and apply Theorem 2.4 to Hamiltonian (4.1). Note that s, the degree of $\mathcal{G}$, is 3, whereas $p = 1$. Thus, each $\mathcal{H}_i$ is a homogenous polynomial of degree $i+3$. For all $i \ge 1$ the ith term of the generating function $\mathcal{W}(\boldsymbol{x})$ of the transformation has the form $\mathcal{W}_i(\boldsymbol{x}) = x\, p_{i+1} + x^2 \log|x|\, q_i$, with p_{i+1} and q_i homogeneous polynomials in $\boldsymbol{x}$ of degrees $i+1$ and i, respectively. Furthermore, if the new variables are called $\boldsymbol{y} = (y, Y)$, the new Hamiltonian will be a polynomial $\mathcal{K}$ composed of homogeneous polynomials $\mathcal{K}_i$. More precisely, for all $i \ge 0$ we have that $\mathcal{K}_{3\,i} = \bar{h}_i\, (y^2\, Y)^{i+1}$ (with $\bar{h}_i$ a constant polynomial in the coefficients $h_{i\,j}$), whereas $\mathcal{K}_{3\,i+1} = \mathcal{K}_{3\,i+2} = 0$.*

Proof. We make a single-induction process over i, the order of the Lie transformation. We have the following.

Step 1. *First-order normalization.* Applying the algorithms of section 3 at order $i = 1$ we get

$$\begin{aligned} \mathcal{K}_1 &= 0, \\ \mathcal{W}_1 &= \frac{x}{3}\left(h_{1,1}\, x^2 - h_{1,3}\, X^2\right) + x^2 \log|x|\, h_{1,2}\, X. \end{aligned}$$

This is straightforward, noting that the induction hypothesis is true for $i = 1$.

Step 2. *Higher-order terms.* Suppose now that the assertion is true up to order $i-1$. Thus we have that $\mathcal{K}_{i-1}$ will be zero when $i-1$ does not divide 3 or $\bar{h}_{(i-1)/3}\,(y^2\,Y)^{(i+2)/3}$ otherwise. Besides, $\mathcal{W}_{i-1} = x\,p_i + x^2\,\log|x|\,q_{i-1}$ with p_i and q_{i-1} homogeneous polynomials in $\boldsymbol{x}$ of degrees i and $i-1$, respectively. In addition, since $\mathcal{W}_{i-1}$ satisfies the identity $\mathcal{L}_{\mathcal{H}_0}(\mathcal{W}_{i-1}) + \mathcal{K}_{i-1} = \bar{\mathcal{H}}_{i-1}$, we can put $\bar{\mathcal{H}}_{i-1} = x^2\,\bar{p}_i + x^3\,\log|x|\,\bar{q}_{i-1}$ with $\bar{p}_i$ and $\bar{q}_{i-1}$ homogeneous polynomials in $\boldsymbol{x}$ of degrees i and $i-1$, respectively. Therefore, using the properties of the Poisson brackets we know that all intermediate Hamiltonians $\mathcal{H}_{j,k}$, with $j+k=i-1$, are of the form $x^2\,\bar{p}_i^{(j,k)} + x^3\,\log|x|\,\bar{q}_{i-1}^{(j,k)}$ for some polynomials $\bar{p}_i^{(j,k)}$ and $\bar{q}_{i-1}^{(j,k)}$ of respective degrees i and $i-1$.

We need to conclude that it remains true at order i. Notice that now we have to calculate the diagonal $\mathcal{H}_{j,k}$ for $j+k=i$ starting at $\mathcal{H}_{i-1,1}$ and ending at $\mathcal{H}_{0,i}$ using Algorithm 4. Specifically, we set $\mathcal{W}_i = 0$ and compute

$$\mathcal{H}_{i-1,1} \;=\; \mathcal{H}_i + \sum_{\ell=0}^{i-1} \binom{i-1}{\ell} \{\mathcal{H}_{i-1-\ell,0}, \mathcal{W}_{\ell+1}\}.$$

Taking into account the forms of $\mathcal{W}_j$ for each $j < i$ and of $\mathcal{H}_k$ for $k \le i$, it is not hard to observe that all Poisson brackets $\{\mathcal{H}_{i-1-\ell,0}, \mathcal{W}_{\ell+1}\}$ produce terms like $x^2\,\bar{p}_{i+1}^{(\ell)} + x^3\,\log|x|\,\bar{q}_i^{(\ell)}$ with $\bar{p}_{i+1}^{(\ell)}$ and $\bar{q}_i^{(\ell)}$ homogeneous polynomials of degrees $i+1$ and i. Using the recurrence (3.3) we obtain that all Hamiltonians $\mathcal{H}_{j,k}$ with $j+k=i$ are written as $x^2\,\bar{p}_{i+1}^{(j,k)} + x^3\,\log|x|\,\bar{q}_i^{(j,k)}$, where $\bar{p}_{i+1}^{(j,k)}$ and $\bar{q}_i^{(j,k)}$ are polynomials of degrees $i+1$ and i; the same holds for $\bar{\mathcal{H}}_i$. Thus we can write $\bar{\mathcal{H}}_i = x^2\,\bar{p}_{i+1} + x^3\,\log|x|\,\bar{q}_i$. Next we need to extract $\mathcal{K}_i$ from $x^2\,\bar{p}_{i+1}$ (expressing it first in terms of $\boldsymbol{y}$). Note that $\mathcal{K}_i$ must belong to $\ker(\mathcal{L}_{\mathcal{G}})$ and this kernel is spanned by integer powers of $y^2\,Y$; thus $\mathcal{K}_i$ is zero when i is not a multiple of 3 or $\mathcal{K}_i = \bar{h}_{i/3}\,(y^2\,Y)^{(i+3)/3}$ otherwise. Finally we solve the PDE $\mathcal{L}_{\mathcal{H}_0}(\mathcal{W}_i) = \bar{\mathcal{H}}_i - \mathcal{K}_i$, but it is satisfied whenever $\mathcal{W}_i$ is of the form demanded in the induction hypothesis. Therefore, all assertions turn out to be true and Result 4.1 is obtained. □

The generating function $\mathcal{W}(\boldsymbol{x})$ contains polynomial and logarithmic terms. However, it satisfies $\mathcal{W}(\boldsymbol{x}) \to 0$ as $\boldsymbol{x} \to \boldsymbol{0}$. Besides, the normalization process will map the coordinates $\boldsymbol{x} = (x, X)$ into the new coordinates $\boldsymbol{y} = (y, Y)$ and vice versa. By construction (using Algorithms 5 and 6) both maps are $\mathcal{C}^2$ in $\boldsymbol{R}^2$. Note also that $\mathcal{W}$ is $\mathcal{C}^\infty$ in $\boldsymbol{R}^2 \setminus \{x=0\}$ and therefore this treatment could be used to analyze a piece of the phase space, excluding the ray $x=0$, whenever the Hamiltonian $\mathcal{H}$ does not come from an expansion around the origin of $\boldsymbol{R}^2$.

One should observe that when trying to solve the homological equation (3.1) at any order i, the Hamiltonians $\mathcal{H}_i$ need to be of the type (4.2), as the corresponding generating function can be obtained as $\mathcal{W}_i = p_i + \log|x|\,q_i$. This is requested to overcome the problem of dropping significant terms when truncating the system at order $L \ge 1$.

Remark 4.1. It is clear that after the normalizing process, we arrive at a system of 0 DOFs. We could extend this result for n-DOF Hamiltonians with $n > 1$ and $\mathcal{H}_0 = x^2\,X$. Similar conclusions would be drawn and the method would lead to an $(n-1)$-DOF Hamiltonian vector field through $\mathcal{C}^2$-maps. Note in addition that a formal integral of $\mathcal{H}$ can be calculated through Algorithm 6 using $\mathcal{W}$ and $\mathcal{G}$.

4.2. The swing spring. Our aim is to show the occurrence of Hamiltonian–Hopf bifurcations in the swing spring (also called elastic or spring pendulum) through

the use of generalized normal forms and the algorithms developed in section 3. Furthermore, we also will show that the dynamical system of the swing spring has monodromy.

4.2.1. Description of the system. The spring pendulum is a mechanical system that exemplifies the motion of a point particle attached to a spring under a constant vertical gravitation field. The position of the particle is denoted by $\boldsymbol{r} = (x_1, x_2, x_3)$, whereas the velocity $\dot{\boldsymbol{r}}$ is the vector (X_1, X_2, X_3). This dynamical system is represented through the Hamilton function

$$\mathcal{H}(\boldsymbol{x}) = \frac{1}{2}(X_1^2 + X_2^2 + X_3^2) + U(\boldsymbol{r}), \tag{4.3}$$

where the potential U is written as

$$U(\boldsymbol{r}) \;=\; x_3 + \frac{1}{2}\nu^2 \left(1 - \frac{1}{\nu^2} - \sqrt{x_1^2 + x_2^2 + x_3^2}\right)^2. \tag{4.4}$$

The parameter ν is related to the equilibrium and unstretched lengths of the spring, respectively, l and l_0, by $\nu = \sqrt{l/(l - l_0)}$. Thus, $\nu > 1$ since the frequency of the spring oscillation is bigger than the frequency of the small amplitude pendulum oscillations, that is, $l \geq l_0$.

The potential has an equilibrium point when the forces of gravity and the spring balance. It corresponds to $\boldsymbol{r}_0 = (0, 0, -1)$ and is linearly stable. We will carry out four steps of the Lie–Deprit process so as to make a proper analysis that we will explain later on. As a consequence, we expand the potential $U(\boldsymbol{r})$ around $(0, 0, -1)$ up to terms of degree 6. Hence, we calculate the Hamiltonian approximated up to degree 6 with the equilibrium shifted to the origin

$$\mathcal{H}(\boldsymbol{x}) \;=\; \mathcal{H}_0 + \sum_{i=1}^{4} \mathcal{H}_i(\boldsymbol{x}), \tag{4.5}$$

where

$$\mathcal{H}_0 \;=\; \frac{1}{2}(X_1^2 + X_2^2 + X_3^2) + \frac{1}{2}(x_1^2 + x_2^2 + \nu^2 x_3^2)$$

and each $\mathcal{H}_i$ is a homogeneous polynomial in $\boldsymbol{x}$ of degree $i+2$ for $i \in \{1, \ldots, 4\}$. Note that x_3 accounts for the displacement from -1. Note also that the coefficients of $\mathcal{H}_i$ depend on ν. We could introduce a small parameter by stretching the coordinates $\boldsymbol{x} = \varepsilon\, \boldsymbol{x}'$ and scaling time. However, we prefer to set $\varepsilon = 1$ and use the theory of section 2 and algorithms of section 3, proceeding straightforwardly.

We easily see that the system is a $1:1:\nu$ resonant Hamiltonian. The $1:1:2$ case has already been analyzed by Dullin, Giacobbe, and Cushman [9]. In our treatment, we broach the nonresonant case; equivalently, we take $\nu \in \boldsymbol{R} \setminus \mathbf{Q}$ together with $\nu > 1$.

4.2.2. Integrable approximation. First, it is straightforward to note that the system is symmetric with respect to the z-axis. Equivalently, Hamiltonian (4.5) is invariant under the flow defined by the field associated with the third component of the angular momentum $L_3 = x_1 X_2 - x_2 X_1$. Therefore, the departure (i.e., the original) Hamiltonian defines, in fact, a 2-DOF system.

The next step is the introduction of a new (formal) symmetry in the system to achieve an integrable approximation. Now it should be clear that the application

of the normal form theorem would lead to a 0-DOF Hamiltonian dynamical system. The reason for this is the following. After the normalization procedure (i.e., after applying Theorem 2.1 and truncating higher-order terms), one introduces in principle one (formal) integral, which in this case, due to the semisimple character of $\mathcal{H}_0$, is precisely $\mathcal{H}_0$. However the number of independent integrals introduced in the process is two—$I_1 = (X_1^2 + x_1^2)/2 + (X_2^2 + x_2^2)/2$ and $I_2 = (X_3^2 + \nu^2 x_3^2)/2$—due to the absence of resonant terms involving x_3 and X_3 with x_1, x_2, X_1, or X_2. Now, one still has to take into account the occurrence of the (exact) integral L_3 yielding three independent integrals. As a result, the reduction process using invariant theory leads to the so-called fully reduced normalized system, which is indeed a trivial integrable Hamiltonian system. In this sense this normalization would be too drastic. Thence, we have to make use of the generalized method for constructing normal forms as stated in Theorem 2.4.

First reduction. We choose $\mathcal{G}(\boldsymbol{x}) = \nu^2 x_3^2 + X_3^2$. At this point, we make use of Algorithms 3 and 4 presented in section 3 to compute the generalized normal form with the aim of extending the integral $\mathcal{G}$ up to some degree. The normalized and truncated Hamiltonian up to order 4 (that is, $L = 4$ and the polynomials are of degree 6) is denoted by $\mathcal{K}$, thus $\{\mathcal{K}, \mathcal{G}\} = 0$. The reason for going to order 6 is motivated by the fact that odd orders give 0 and at order 2 the system has nonisolated equilibria and thus is not structurally stable.

The reduction is regular. Moreover, the corresponding homological equations give polynomial results in all orders $i \in \{1, \ldots, 4\}$. Henceforth the generating function of the transformation $\mathcal{W}$ is a polynomial of degree 6 and always would be a polynomial if we would push the calculations up to any order. This is due to the nonresonant situation existing between $\mathcal{G}$ and $\frac{1}{2}(x_1^2 + X_1^2) + \frac{1}{2}(x_2^2 + X_2^2)$. An asymptotic integral of $\mathcal{H}$ could be found by means of Algorithm 6 applied to $\mathcal{G}$ with the aid of $\mathcal{W}$. The result would be $\mathcal{I}_\mathcal{G}$, a polynomial in $\boldsymbol{x}$ of degree 6 such that $\{\mathcal{H}, \mathcal{I}_\mathcal{G}\} = p_7(\boldsymbol{x})$, where p_7 is a polynomial whose lowest degree is 7.

The invariants associated with this reduction are

$$i_1 = x_1, \quad i_2 = X_1, \quad i_3 = x_2, \quad i_4 = X_2, \quad i_5 = \nu^2 x_3^2 + X_3^2.$$

Then, we fix a value for the integral, i.e., we set $\mathcal{G} = i_5 = j_1 \geq 0$. Next, the reduced phase space is the hyperplane $i_5 = j_1$ defined in the four-dimensional space determined by $i_1, \ldots, i_4$ that we call P_{j_1}. In this way the transformed Hamiltonian $\mathcal{K}$ in that hyperplane can be expressed as $\mathcal{K}(i_1, i_2, i_3, i_4; j_1)$.

Second reduction. As we have stressed previously, the system is axially symmetric with the third component of the angular momentum L_3 an integral of motion. Thus, a second (and exact) reduction can be performed straightforwardly.

The invariants associated with this second reduction are

$$\varphi_1 = i_1^2 + i_3^2 = x_1^2 + x_2^2, \qquad \varphi_2 = i_2^2 + i_4^2 = X_1^2 + X_2^2,$$
$$\varphi_3 = i_1 i_2 + i_3 i_4 = x_1 X_1 + x_2 X_2, \qquad \varphi_4 = i_1 i_4 - i_2 i_3 = x_1 X_2 - x_2 X_1.$$

Again, we fix a value for the integral L_3; i.e., we let $\varphi_4 = j_2$ for some constant $j_2 \in \boldsymbol{R}$. Now, the twice-reduced phase space is the two-dimensional hyperbolic paraboloid defined by

$$V_{j_2} = \{\varphi \in \boldsymbol{R}^3 \mid R_{j_2}(\varphi) = 0\},$$

where $R_{j_2}(\varphi) = \varphi_1 \varphi_2 - \varphi_3^2 + j_2^2$. Depending on the value of j_2 there are two cases; see Figure 4.1. If $j_2 \neq 0$, the reduction is regular, whereas if $j_2 = 0$, the reduction is singular with the origin of the frame $0, \varphi_1, \varphi_2, \varphi_3$ being a singular point.

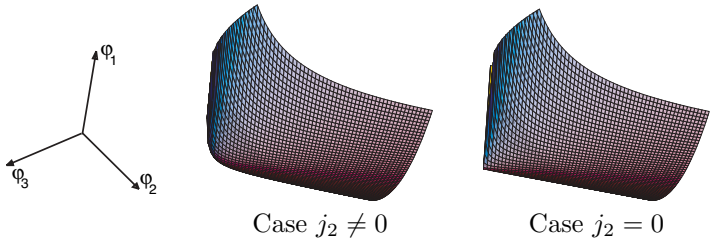


FIG. 4.1. *Twice-reduced phase space.*

Then, putting the twice-reduced Hamiltonian as a function of these invariants, we arrive at

$$\begin{aligned}\mathcal{S}^{\nu}(\varphi_1,\varphi_2,\varphi_3;j_1,j_2) \;=\; & \frac{1}{768\,\nu^6\,(\nu^2-4)^4\,(64-148\,\nu^2+9\,\nu^4)}\\ & \times [c_1\,\varphi_1^3 + c_2\,\varphi_2^3 + c_3\,\varphi_1\,\varphi_2^2 + c_4\,\varphi_1^2\\ & \quad + c_5\,\varphi_2^2 + c_6\,\varphi_1\,\varphi_2 + c_7\,\varphi_1 + c_8\varphi_2 + c_9],\end{aligned} \tag{4.6}$$

where all the c_i are polynomials in ν, j_1, and j_2 having integer coefficients. In the remainder, we will denote this twice-reduced Hamilton function as $\mathcal{S}^{\nu}$.

4.2.3. Hamiltonian–Hopf bifurcations at the origin. We have to look at the points on the twice-reduced phase space V_{j_2} where the gradient $\nabla\mathcal{S}^{\nu}$ is parallel to the gradient ∇R_{j_2} to get the equilibria of the system defined by $\mathcal{S}^{\nu}$. Equivalently, this can be interpreted geometrically as seeking whether the level set $\{\mathcal{S}^{\nu}(\varphi) = h\}$ is tangent to V_{j_2}, which always occurs at the singular point of V_{j_2}.

Now, we restrict ourselves to the singular case $j_2 = 0$. In this situation the reduced phase space is singular at the origin, which is always an equilibrium point of this space. This singularity reflects the fact that the S^1-action defined by the axial symmetry is not free.

Next we study the stability of the origin, which we denote by φ^0 (so, this equilibrium point satisfies $\varphi_i = 0$ for $i \in \{1,2,3,4\}$). We need to calculate the Poisson brackets of the φ_i. We do that with the aim of using the symplectic structure on the reduced phase space. The matrix containing all brackets appears in Table 4.1.

TABLE 4.1
The Poisson bracket relations $\{\varphi_i,\varphi_j\}$ *for* $i,j = 1,\ldots,4$.

$\{\ ,\ \}$	φ_1	φ_2	φ_3	φ_4
φ_1	0	$4\varphi_3$	$2\varphi_1$	0
φ_2	$-4\varphi_3$	0	$-2\varphi_2$	0
φ_3	$-2\varphi_1$	$2\varphi_2$	0	0
φ_4	0	0	0	0

Now we can obtain the Hamiltonian vector field $X_{\mathcal{S}^{\nu}}$ associated with the Hamiltonian $\mathcal{S}^{\nu}(\varphi_1,\varphi_2,\varphi_3;j_1,j_2)$ and, after that, we set $j_2 = 0$ to analyze whether that system undergoes a Hamiltonian–Hopf bifurcation.

The Hessian matrix obtained from its linearization evaluated at the origin turns out to be

$$DX_{\mathcal{S}_0^\nu}(\varphi^0) = \begin{pmatrix} 0 & 0 & \frac{\lambda_{\nu j_1}}{128\,\nu^4\,(\nu^2-4)^3} \\ 0 & 0 & \frac{-2\,j_1+(32+5\,j_1)\,\nu^2-8\,\nu^4}{8\,\nu^2\,(\nu^2-4)} \\ \frac{-2\,j_1+(32+5\,j_1)\,\nu^2-8\,\nu^4}{4\,\nu^2\,(\nu^2-4)} & \frac{\lambda_{\nu j_1}}{64\,\nu^4\,(\nu^2-4)^3} & 0 \end{pmatrix},$$

where we have defined

$$\begin{aligned}\lambda_{\nu j_1} &= 128\,\nu^4\,(\nu^2-4)^3 + 16\,j_1\,\nu^2\,(\nu^2-4)^2(-6+7\,\nu^2+2\,\nu^4) \\ &\quad + j_1^2\,(-144+436\,\nu^2-300\,\nu^4-105\,\nu^6+59\,\nu^8).\end{aligned}$$

The eigenvalues of $DX_{\mathcal{S}_0^\nu}(\varphi^0)$ vanish if and only if j_1 takes the values

$$\begin{aligned} j_{11} &= \tfrac{8\,\nu^2\,(\nu^2-4)}{5\,\nu^2-2}, \\ j_{12} &= -\frac{16\,\nu^2\,(\nu^2-4)^{3/2}}{\sqrt{\nu^2-4}\,(2\,\nu^4+7\,\nu^2-6)-\sqrt{(\nu^2+2)\,(72-286\,\nu^2+351\,\nu^4-114\,\nu^6+4\,\nu^8)}}, \\ j_{13} &= -\frac{16\,\nu^2\,(\nu^2-4)^{3/2}}{\sqrt{\nu^2-4}\,(2\,\nu^4+7\,\nu^2-6)+\sqrt{(\nu^2+2)\,(72-286\,\nu^2+351\,\nu^4-114\,\nu^6+4\,\nu^8)}}. \end{aligned}$$

For irrational values of $\nu > 1$, the denominators of j_{11} and j_{13} do not vanish, whereas the denominator of j_{12} is zero provided that $\nu = 1.6635156185484876$. (As the infinite-precision expressions are very involved, we use floating-point arithmetic.) Thus, for this particular value of ν, the values of j_1 for which $DX_{\mathcal{S}^{1.66\ldots}}(\varphi^0)$ has three null eigenvalues get reduced to $j_{11} = 0.9513184442193461$.

Now, if we vary j_1 and make it pass through j_{11}, j_{12}, or j_{13}, we find that the origin changes its stability from elliptic to hyperbolic point or vice versa. In fact, one of the eigenvalues of $DX_{\mathcal{S}_0^\nu}(\varphi^0)$ is always zero as the system is 1-DOF, and we are using a three-dimensional frame with the constraint given through R_{j_2}. Besides, the other two eigenvalues are either pure imaginary (elliptic point, also called center) or real ones with different sign (hyperbolic point, also called saddle). More concretely, depending on the value of the external parameter ν, we have three cases. We have

(1) for $\nu \in (1, 1.5117504938658013)$: *elliptic*;
(2) for $\nu \in (1.5117504938658013, 1.6635156185484876)$: *hyperbolic* if ($j_1 < j_{13}$ or $j_1 > j_{12}$) and *elliptic* for $j_1 \in (j_{13}, j_{12})$;
(3) for $\nu \in (1.6635156185484876, 2)$: *elliptic* for $j_1 < j_{13}$ and *hyperbolic* for $j_1 > j_{13}$;
(4) for $\nu > 2$: *elliptic* for $j_1 < j_{11}$ and *hyperbolic* for $j_1 > j_{11}$.

Recall that the origin φ^0 is an isolated equilibrium point in all the cases since we have pushed the calculations to degree 6. (At degree 4 the origin was not an isolated critical point.) Note as well that the twice-reduced Hamiltonian is a family of Hamiltonians depending on two parameters (ν and j_1). However, ν is an external parameter while j_2 is an internal or distinguished one.

Next we perform a suitable change of coordinates so that $\mathcal{S}^\nu$ acts as a Morse function on V_{j_2}. This is achieved through two successive changes of coordinates. We define

$$\sigma_1 : \boldsymbol{R}^3 \to \boldsymbol{R}^3 : (\varphi_1, \varphi_2, \varphi_3) \mapsto (\varsigma_1 + \varsigma_2, \varsigma_1 - \varsigma_2, \varsigma_3)$$

and

$$\sigma_2 : \boldsymbol{R}^3 \to \boldsymbol{R}^3 : (\varsigma_1, \varsigma_2, \varsigma_3) \mapsto \left(\tfrac{1}{4}(\tau_1^2 - \tau_2^2), \tfrac{1}{2}\,\tau_1\tau_2, \tfrac{1}{4}\,\tau_3\right).$$

By doing so we transform $R_{j_2}(\varphi)$ into $R_{j_2}(\tau) = \frac{1}{16}(\tau_3^2 - \tau_1^4 - 2\tau_1^2\tau_2^2 - \tau_2^4) - j_2^2$. Consequently, $R_0(\tau) = \frac{1}{16}(\tau_3^2-\tau_1^4-2\tau_1^2\tau_2^2-\tau_2^4)-j_2^2$, and $R_0(\tau) = 0$ defines a parabolic surface. Thus, through the changes σ_1 and σ_2 we have removed the singularity from the twice-reduced phase space V_{j_2}.

As the next step we determine the 2-jet related to $\mathcal{S}^{\nu}_{j_2=0}$. It leads to the expression

(4.7)
$$\begin{aligned}\mathcal{S}^{2\text{-jet}}_{j_2=0}&(\tau_1,\tau_2)\\ &= \tfrac{2j_1-(32+5j_1)\nu^2+8\nu^4}{32\nu^2(\nu^2-4)}\,\tau_2^2\\ &\quad+\tfrac{128\nu^4(\nu^2-4)^3+16j_1\nu^2(\nu^2-4)^2(2\nu^4+7\nu^2-6)+j_1^2(59\nu^8-105\nu^6-300\nu^4+436\nu^2-144)}{512\nu^4(\nu^2-4)^3}\,\tau_1^2.\end{aligned}$$

Note that the composition $\sigma_2 \circ \sigma_1$ maps the origin φ^0 to the origin τ^0 in the new variables.

The 2-jet plays the role of a Morse function. As a consequence, given a fixed value of $\nu > 1$, the analysis is carried out by replacing j_1 by its value before the bifurcations, after the bifurcations, and in the bifurcations value itself. From the expression (4.7) it is easy to deduce that when ν and j_1 are taken so that the stability at the origin is of elliptic type, the 2-jet takes the form

$$\mathcal{S}^{2\text{-jet}}_{j_2=0}(\tau_1,\tau_2) \;=\; a_1\,\tau_1^2 + a_2\,\tau_2^2,$$

while when we select ν and j_1 such that the origin τ^0 is of hyperbolic character, we have

$$\mathcal{S}^{2\text{-jet}}_{j_2=0}(\tau_1,\tau_2) \;=\; a_1\,\tau_1^2 - a_2\,\tau_2^2,$$

where a_1, a_2 are nonnull constants such that they satisfy $a_1\,a_2 > 0$. Therefore, the origin will be stable in the first case and unstable in the second. Besides, in the bifurcation values, the 2-jet shows an unstable behavior of τ^0.

The above paragraphs enable us to conclude that indeed a Hamiltonian–Hopf bifurcation takes place whenever one of these situations occurs (see also the different regions and bifurcation lines in Figure 4.2):

(1) for $\nu \in (1.5117504938658013, 1.6635156185484876)$ and $j_1 = j_{12}$,
(2) for $\nu \in (1.5117504938658013, 2)$ and $j_1 = j_{13}$,
(3) for $\nu > 2$ and $j_1 = j_{11}$.

A very detailed description of Hamiltonian–Hopf bifurcations is available in [14].

4.2.4. Monodromy. We end our example with an analysis of the occurrence of monodromy in the Liouville-integrable system $(\mathcal{K}, \mathcal{G}, L_3, \boldsymbol{R}^6, \omega)$, with $\omega = dx_1 \wedge dX_1 + dx_2 \wedge dX_2 + dx_3 \wedge dX_3$. Nontrivial monodromy describes the global twisting of a family of invariant two-dimensional tori parameterized by a circle of regular values of the energy-momentum map of a certain integrable system. Its presence is determined by the existence of a singular fiber of the energy-momentum map, topologically a torus with one or two pinched points. If an integrable system has monodromy, then one cannot label the tori in a unique way by values of the actions [5]. To unveil the monodromy feature of our departure Hamiltonian we start by studying the 2-DOF integrable system $(\mathcal{K}(i_1, i_2, i_3, i_4; j_1), L_3, P_{j_1}, \{\ ,\ \})$, where $\mathcal{K}$ represents the normal form Hamiltonian expressed in the invariants associated with the first reduction and in P_{j_1} we take $j_1 > 0$. We note that $\mathcal{K}$ and L_3 commute, which is an essential hypothesis for proving monodromy. In the case where we have a hyperbolic behavior of the origin φ^0, that is, whenever

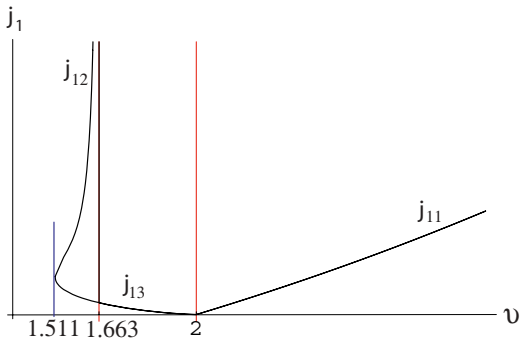


FIG. 4.2. *The bifurcation lines in the plane defined by the external parameter ν and the internal one j_1 are shown. In the ν-axis we are taking into account only the irrational values. The origin of the coordinates in the picture is the point* ($\nu = 1.5117504938658013$, $j_1 = 0$).

(1) $\nu \in (1.5117504938658013, 1.6635156185484876)$ and $j_1 < j_{13}$ or $j_1 > j_{12}$,
(2) $\nu \in (1.6635156185484876, 2)$ and $j_1 > j_{13}$,
(3) $\nu > 2$ and $j_1 > j_{11}$,

we see that the system satisfies the hypotheses of the monodromy theorem for Hamiltonian systems as stated in Matveev [17] and Zung [35]; see also a similar case in [5] and the non-Hamiltonian situation in [4].

Looking at the energy momentum mapping

$$\mathcal{EM}_{j_1} : P_{j_1} \subseteq \boldsymbol{R}^4 \to \boldsymbol{R}^2 : (i_1, i_2, i_3, i_4) \mapsto (\mathcal{K}_{j_1}(i_1, i_2, i_3, i_4), L_3(i_1, i_2, i_3, i_4)), \tag{4.8}$$

fixing values for the energy, h, and for the third component of the angular momentum, j_2, it is straightforward to observe that the fiber $\mathcal{EM}_{j_1}^{-1}(h, j_2)$ is compact and connected by simply inverting $\mathcal{EM}_{j_1}$ and putting the invariants i_j in terms of j_2 and h. So at the critical value $(j_1/2, 0)$ of $\mathcal{EM}_{j_1}$ we have a once-pinched 2-torus. Therefore, applying the monodromy theorem, if Γ represents a closed curve around the critical value $(j_1/2, 0)$ we have that the 2-torus bundle $\mathcal{EM}_{j_1}(\Gamma)$ has a monodromy linear mapping whose matrix is

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

This implies, following reasoning due to Rink [31], that the system defined through the system (4.3) also has monodromy. The analysis of the Hamiltonian–Hopf bifurcations in the original system could be done by reconstructing the flow near φ^0, using the direct change of coordinates, that is, by making use of Algorithm 5. The analysis of monodromy for the spring pendulum when $\nu = 2$ can be found in the work by Dullin, Giacobbe, and Cushman [9].

5. Conclusions. We present a set of algorithms to deal with generalized normal forms for polynomial Hamiltonians. Our initial Hamiltonian $\mathcal{H}$ is a Hamiltonian vector field with n DOFs that can be written as a sum of homogeneous polynomial Hamiltonians, $\mathcal{H}_i$, starting at degree $p + 2$ (with $p \geq -1$ a fixed integer) up to a certain degree. The coefficients of the polynomials $\mathcal{H}_i$ are arbitrary and can be real or complex. Given a polynomial $\mathcal{G}$ of degree $s \geq 1$ our aim is to construct a polynomial Hamiltonian $\mathcal{K}$ up to a certain degree $L + p + 2$ and a generating function $\mathcal{W}$ (sometimes also a polynomial of degree $L+p+2$ and generically a smooth function in some domain) such that the Poisson brackets $\{\mathcal{K}_i, \mathcal{G}\}$ vanish for all $i \in \{1, \ldots, L\}$.

So, if $\mathcal{H}_0$ commutes with $\mathcal{G}$, then $\mathcal{G}$ becomes an integral of $\mathcal{K}$, after truncation of higher-order terms.

One of our algorithms looks for a polynomial generating function when possible. If at an order i there is not a solution of a certain linear system, it tries to solve a PDE corresponding to the homological equation with the aim of determining a smooth generating function $\mathcal{W}_i$. We have included within our approach the case of the normal form theorem, selecting $\mathcal{G}$ adequately to yield the same result.

The resolution of the homological equation is not optimal from the computational point of view as the algorithm we use requires handling, at least, two systems of linear equations. However, it can be applied to a large class of Hamilton functions in n DOFs. Besides, one does not need to write the unperturbed part of the Hamiltonian, that is, $\mathcal{H}_0$, in normal form. Moreover, the computation of the direct and inverse changes of variables is optimized as much as possible since we use the algorithm of the inverse introduced by Henrard [16].

We apply the theory of generalized normal forms together with the algorithms developed in section 3 to deal with two applications that could not be studied with the current techniques of the standard approaches based on normal forms.

Acknowledgments. We thank the anonymous referees for remarks that helped to improve a previous version of the paper.

REFERENCES

[1] G. D. Birkhoff, *Dynamical Systems*, Amer. Math. Soc. Colloq. Publ. 9, AMS, Providence, RI, 1927.

[2] R. C. Churchill, M. Kummer, and D. L. Rod, *On averaging, reduction, and symmetry in Hamiltonian systems*, J. Differential Equations, 49 (1983), pp. 359–414.

[3] R. Cushman, A. Deprit, and R. Mosak, *Normal form and representation theory*, J. Math. Phys., 24 (1983), pp. 2102–2117.

[4] R. Cushman and J. J. Duistermaat, *Non-Hamiltonian monodromy*, J. Differential Equations, 172 (2001), pp. 42–58.

[5] R. Cushman and D. A. Sadovskií, *Monodromy in the hydrogen atom in crossed fields*, Phys. D, 142 (2000), pp. 166–196.

[6] A. Deprit, *Canonical transformations depending on a small parameter*, Celestial Mech., 1 (1969), pp. 12–30.

[7] A. Deprit, J. Henrard, and A. Rom, *Analytical lunar ephemeris. Delaunay's theory*, Astronom. J., 76 (1971), pp. 269–272.

[8] A. Deprit, J. Henrard, and A. Rom, *Analytical lunar ephemeris. The variational orbit*, Astronom. J., 76 (1971), pp. 723–726.

[9] H. Dullin, A. Giacobbe, and R. Cushman, *Monodromy in the resonant swing spring*, Phys. D, 190 (2004), pp. 15–37.

[10] S. Ferrer, H. Hanßmann, J. Palacián, and P. Yanguas, *On perturbed oscillators in* 1-1-1 *resonance: The case of axially symmetric cubic potentials*, J. Geom. Phys., 40 (2002), pp. 320–369.

[11] D. M. Galin, *Versal deformations of linear Hamiltonian systems*, in Amer. Math. Soc. Transl. (Ser. 2) 118, AMS, Providence, RI, 1982, pp. 1–12.

[12] A. Giorgilli, *A computer program for integrals of motion*, Comput. Phys. Comm., 16 (1979), pp. 331–343.

[13] F. G. Gustavson, *On constructing formal integrals of a Hamiltonian system near an equilibrium point*, Astronom. J., 71 (1966), pp. 670–686.

[14] H. Hanßmann and J.-C. van der Meer, *On the Hamiltonian Hopf bifurcations in the* 3*D Hénon–Heiles family*, J. Dynam. Differential Equations, 14 (2002), pp. 675–695.

[15] M. Hénon and C. Heiles, *The applicability of the third integral of motion: Some numerical experiments*, Astronom. J., 69 (1964), pp. 73–79.

[16] J. Henrard, *The algorithm of the inverse Lie transform*, in Recent Advances in Dynamical Astronomy, V. Szebehely and B. D. Tapley, eds., D. Reidel, Dordrecht, The Netherlands, 1973, pp. 250–259.

[17] V. S. Matveev, *Integrable Hamiltonian systems with two degrees of freedom. Topological structure of saturated neighborhoods of points of focus-focus and saddle-saddle types*, Mat. Sb., 187 (1996), pp. 29–58; translation in Sb. Math., 187 (1996), pp. 495–524.
[18] J.-C. van der Meer, *Nonsemisimple* 1:1 *resonance at an equilibrium*, Celestial Mech., 27 (1982), pp. 131–149.
[19] J.-C. van der Meer, *The Hamiltonian Hopf Bifurcation*, Lecture Notes in Math. 1160, Springer-Verlag, Berlin, Heidelberg, 1985.
[20] K. R. Meyer, *Normal forms for Hamiltonian systems*, Celestial Mech., 9 (1974), pp. 517–522.
[21] K. R. Meyer, *Normal forms for the general equilibrium*, Funkcial. Ekvac., 27 (1984), pp. 261–271.
[22] K. R. Meyer and G. R. Hall, *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*, Appl. Math. Sci. 90, Springer-Verlag, New York, 1992.
[23] K. R. Meyer and D. S. Schmidt, *Periodic orbits near $\mathcal{L}_4$ for mass ratios near the critical mass ratio of Routh*, Celestial Mech., 4 (1971), pp. 99–109.
[24] J. Moser, *New aspects in the theory of stability of Hamiltonian systems*, Comm. Pure Appl. Math., 11 (1958), pp. 81–114.
[25] J. Palacián and P. Yanguas, *Simplification of perturbed Hamiltonians through Lie transformations*, in Hamiltonian Systems and Celestial Mechanics (HAMSYS-98), J. Delgado, E. A. Lacomba, E. Pérez-Chavela, and J. Llibre, eds., World Sci. Monogr. Ser. Math. 6, World Scientific, Singapore, 2000, pp. 284–302.
[26] J. Palacián and P. Yanguas, *Reduction of polynomial Hamiltonians by the construction of formal integrals*, Nonlinearity, 13 (2000), pp. 1021–1055.
[27] J. Palacián and P.Yanguas, *Reduction of polynomial planar Hamiltonians with quadratic unperturbed part*, SIAM Rev., 42 (2000), pp. 671–691.
[28] J. Palacián and P. Yanguas, *Equivariant n-DOF Hamiltonians via generalized normal forms*, Commun. Contemp. Math., 5 (2003), pp. 449–480.
[29] R. Pérez-Marco, *Convergence and generic divergence of the Birkhoff normal form*, Ann. of Math. (2), 157 (2003), pp. 557–574.
[30] H. Poincaré, *Sur les courbes définies par les équations différentielles*, J. Math. Pures Appl. (4), 1 (1885), pp. 167–244.
[31] B. Rink, *A Cantor set of tori with monodromy near a focus-focus singularity*, Nonlinearity, 17 (2004), pp. 1–10.
[32] A. G. Sokol'skij, *On the stability of an autonomous Hamiltonian system with two degrees of freedom in the case of equal frequencies*, Prikl. Mat. Meh., 38 (1974), pp. 791–799; translation in J. Appl. Math. Mech., 38 (1974), pp. 741–749.
[33] E. T. Whittaker, *On the adelphic integral of the differential equations of dynamics*, Proc. Roy. Soc. Edinburgh Sect. A, 37 (1918), pp. 95–116.
[34] E. T. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies*, Cambridge University Press, Cambridge, UK, 1927.
[35] N. T. Zung, *A note on focus-focus singularities*, Differ. Geom. Appl., 7 (1997), pp. 123–130.

SIAM J. APPL. MATH.
Vol. 65, No. 4, pp. 1153–1174

SPATIALLY DISCRETE FITZHUGH–NAGUMO EQUATIONS*

CHRISTOPHER E. ELMER† AND ERIK S. VAN VLECK‡

Abstract. We consider pulse and front solutions to a spatially discrete FitzHugh–Nagumo equation that contains terms to represent both depolarization and hyperpolarization of the nerve axon. We demonstrate a technique for deriving candidate solutions for the McKean nonlinearity and present and apply solvability conditions necessary for existence. Our equation contains both spatially continuous and discrete diffusion terms.

Key words. traveling fronts and pulses, discrete diffusion

AMS subject classifications. 35K57, 74N99

DOI. 10.1137/S003613990343687X

1. Introduction. By considering an electrical circuit model with a complicated nonlinear resistor, Hodgkin and Huxley (along with Katz) modeled the ionic conductances that generate the action potential of nerve fibers. To develop their model they performed voltage and space clamping experiments on the giant axon of squids, axons which were *relatively* easy to work with because of their size. The Hodgkin–Huxley equations are a four-variable model which may be reduced to a two-variable model (the FitzHugh–Nagumo (FH-N) ODE model) which preserves much of the dynamics of the Hodgkin–Huxley system by considering fast and slow variables and slaving the other variables. When considering a chain of electrical circuits, diffusion is added as a means of propagation in the spatial variable, thus obtaining the FH-N PDE model. Since the seminal work of Hodgkin, Huxley, and Katz, similar experiments have been performed on nerve axons of vertebrates and its been discovered that, electrically, nerve fibers behave as spatially discrete periodic structures in vertebrates. This is due to the periodically spaced active channels (nodes of Ranvier) in the myelin insulation (in the coating by Schwann cells or oligodendrocytes). Thus it is not only appropriate but correct to model motor nerves in vertebrates with equations which also have a spatially discrete periodic structure, to model with nonlinear differential-difference equations (DDEs), in particular an FH-N DDE model.

Our contribution in this paper is to consider front and pulse solutions for a FitzHugh–Nagumo system with both continuous and discrete diffusion, thus allowing one to compare and contrast the dynamics generated by spatially continuous and spatially discrete models of action potential propagation. By employing a piecewise linear bistable nonlinearity we reduce the problem to a linear inhomogeneous equation, for which candidate solutions can be derived using transform methods. The candidate solutions are then shown to be consistent with our ansatz of a front or pulse solution, a necessary condition for existence. We focus on one-front and one-pulse solutions and prove their existence using two approaches: (1) we show that consistency of the

*Received by the editors October 29, 2003; accepted for publication (in revised form) October 27, 2004; published electronically April 14, 2005. This work was supported in part by NSF grants DMS-0204573, DMS-9973393, and DMS-0139824.
http://www.siam.org/journals/siap/65-4/43687.html
†Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ 07102 (elmer@oak.njit.edu).
‡Department of Mathematics, University of Kansas, Lawrence, KS 66045 (evanvleck@math.ukans.edu).

infinite interval solution may be reduced to showing consistency on a finite interval, and (2) we use the implicit function theorem, thus showing that consistency can be obtained with a perturbation argument. Once we have derived the exact solutions and verified their existence, we investigate the solution behavior as a function of the problem parameters.

Mathematical models of the electrical behavior of axons often come from postulating an equivalent electrical circuit model (leaky underwater cable theory) of the excitable axonal membrane. Consider a single nerve fiber (axon) coated with a lipid material called myelin with periodically spaced gaps (which are commonly call nodes). Assuming the axial currents are constant, the intracellular, I_i, and the extracellular, I_e, currents between two consecutive nodes are (by Ohm's law)

$$Lr_iI_{i,n} = -(v_{i,n+1} - v_{i,n}) \qquad \text{and} \qquad Lr_eI_{e,n} = -(v_{e,n+1} - v_{e,n}),$$

where L is the length of the myelin sheath between the nodes, r_i and r_e are the intracellular and extracellular resistances per unit length of material, and $v_{i,n}$ and $v_{e,n}$ are the intracellular and extracellular voltages in the nth node. Using Kirchoff's laws, one obtains

$$I_{i,n-1} - I_{i,n} = I_{e,n} - I_{e,n-1} = \mu p\left(C\frac{\partial v_n}{\partial t} + I_{ion,n}\right),$$

where the quantities in the parentheses are the capacitive current and the ionic current flowing through the nth node from inside to outside, $v_n = v_{i,n} - v_{e,n}$, μ is the length of each node (here they will all be assumed to be the same), p is the perimeter length of the axon (assumed to be constant), C is the capacitance, and $I_{ion,n}$ is the ionic current at each node. The total transmembrane current at a node n is thus given by

$$p\left(C\frac{\partial v_n}{\partial t} + I_{ion,n}\right) = \frac{1}{\mu L(r_e + r_i)}(v_{n+1} - v_n + v_{n-1} - v_n). \tag{1.1}$$

The change of variables $\tau = t/(CR)$ nondimensionalizes time (where R has units $\Omega\, cm^2$) and (1.1) becomes

$$\frac{dv_n}{d\tau} = \rho(v_{n+1} - 2v_n + v_{n-1}) - RI_{ion,n},$$

where $\rho = R/(\mu Lp(r_i + r_e))$. We want the transmembrane ionic current at each node to possess both a sodium and a potassium component (like an actual nerve we want both a "front" and "back" to our traveling waves), thus we use the analytically simple $RI_{ion,n} = f(v_n) + w_n$, where $f(v_n)$ represents the sodium ion current component and w_n represents the potassium ion current contribution, and we add the governing equation

$$\frac{\partial w_n}{\partial t} = b(v_n - rw_n)$$

for our potassium recovery variable w_n. Note that by setting $b = 0$, one can assume that the behavior is dominated by the leading edge behavior and that recovery is so slow that it can be treated as constant.

Related to our work on the discrete FitzHugh–Nagumo equation is the work of Anderson and Sleeman [1] on the existence and stability of equilibrium solutions, the work of Binczak, Eilbeck, and Scott [9] on ephaptic coupling in systems related to

systems of discrete FitzHugh–Nagumo equations, the work of Tonnelier [38, 39, 40], the work of Carpio and Bonilla [8], as well as the enlightening books of Keener and Sneyd [28] and Scott [36]. Work on the discrete Nagumo equation includes the work of Bell and Cosner [4], Keener [26, 27], and Zinner [44, 45] on existence, stability, and propagation failure, and the work of Mallet-Paret [31, 32] establishing a Fredholm theory for linear mixed type delay equations. Other works on discrete Nagumo type equations include [3, 5, 6, 7, 11, 12, 13, 14, 19, 23, 24]. Notable work on the existence and stability on monotone traveling fronts for the Nagumo PDE include that of Aronson and Weinberger [2] and Fife and McLeod [21] and the original work of Nagumo, Arimoto, and Yoshizawa [34]. Existence and stability of fronts and pulses for the FitzHugh–Nagumo PDE begins from the work of FitzHugh [22] (see also [34]) and includes the work on stability of Jones [25], Maginu [30], and Yanagida [43] (see also [29]), the work on existence of Deng [10] and existence and stability results of Evans [15, 16, 17, 18], Feroe [20], Wang [41, 42], Rinzel and Keller [35], and McKean [33] for the piecewise linear nonlinearity considered here.

This paper is organized as follows. In section 2 we present the model equations to be considered, including the nonlinearity, and derive traveling wave equations. In section 3, using transform techniques, we derive the general form for candidate front and pulse solutions. We consider one-front solutions in section 4 and show similarities with monotone one-front solutions of Nagumo type equations. In sections 5 and 6, using the form of the candidate solutions found in section 3, we derive conditions for the existence of one-pulse solutions. Two approaches are considered: one is perturbative in that it shows under certain conditions the existence of one-pulse solutions in a neighborhood of an existing one-pulse solution, while the other shows the existence of a one-pulse solution more directly, but with assumptions that are more difficult to verify. We present plots of the relationship between the driving force and the speed of wave propagation, and we present waveforms obtained numerically, in section 7.

2. Models. The continuous FitzHugh–Nagumo equations (the FH-N PDEs) can be derived as above by considering a smooth spatial domain (or a spatial scale where the local behavior appears homogenous) and thus allowing the spatial difference terms to go zero. This gives the model

$$\begin{cases} v_t = v_{xx} - f(v) - w, & x \in \mathbb{R}^N,\ t > 0, \\ w_t = b(v - rw), \end{cases} \tag{2.1}$$

where $b > 0$ relates the time scales of the pulse front and the recovery, the pulse's tail, and $r \geq 0$ indicates the strength of recovery.

In this paper we consider a differential-difference equation of FitzHugh–Nagumo type which contains both the diffusion term derived by considering periodically nodes, as in the introduction, and the diffusion term obtained by allowing the spatial domain to be uniform. While this may not be a valid first principle derivation, it does allow us to compare and contrast propagation of action potential in the two perspectives (allowing for different length scale assumptions). The equations of interest are

$$\begin{cases} \dot{v}(\eta,t) = \displaystyle\sum_{i=1}^{N} d_i L_i v(\eta,t) + \sum_{i=1}^{N} \gamma_i \frac{\partial^2 v}{\partial \eta_i^2}(\eta,t) - f(v(\eta,t)) - w(\eta,t), \\ \dot{w}(\eta,t) = b[v(\eta,t) - rw(\eta,t)], \end{cases} \tag{2.2}$$

for

- $\eta \in \mathbb{R}^N$, $t \in \mathbb{R}^+$, and η_i is the ith element of η,
- "$\dot{}$" denotes differentiation with respect to t,
- $d_i \geq 0$, $\gamma_i \geq 0$, $i = 1, \ldots, N$, $b > 0$, and $r \geq 0$ are parameters,
- $L_i v(\eta, t) = v(\eta + e_i, t) - 2v(\eta, t) + v(\eta - e_i, t)$, where e_i is the unit vector with 1 in the ith element, and
- in general, f is a function of "cubic" shape, but for our investigations, we employ the piecewise linear f (as was done in [33, 35, 20, 41, 42, 19, 7, 12, 13, 38, 39, 40]),

$$f(v) \equiv v - h(v-a), \quad \text{where} \quad h(v-a) \equiv \begin{cases} 0, & v < a, \\ [0,1], & v = a, \\ 1, & v > a, \end{cases} \tag{2.3}$$

  where $a \in (0,1)$ is a "detuning" parameter allowing for tuning the behavior based on the behavior of the sodium channels.

While we have discussed only the derivation of the one-dimensional model of propagation along a single nerve axon, the equations presented in (2.2) are three-dimensional. This is simply a generalization we choose to explore and it may (or may not) be used to gain insight into three-dimensional biological domains such as cardiac tissue. We intend to study traveling waves (plane waves), and thus we now specify a direction of propagation with the direction normal $\sigma = \{\sigma_1, \ldots, \sigma_N\}^T \in \mathbb{R}^N$, with $\sum_{i=1}^N \sigma_i^2 = 1$, and apply the classic traveling wave ansatz $\phi(\eta \cdot \sigma - ct) = v(\eta, t)$ and $\psi(\eta \cdot \sigma - ct) = w(\eta, t)$ to (2.2) to obtain the system of differential-difference equations:

$$\begin{cases} -c\phi'(\xi) = \displaystyle\sum_{i=1}^N d_i[\phi(\xi + \sigma_i) - 2\phi(\xi) + \phi(\xi - \sigma_i)] + \gamma\phi''(\xi) - f(\phi(\xi)) - \psi(\xi), \\ -c\psi'(\xi) = b[\phi(\xi) - r\psi(\xi)], \end{cases} \tag{2.4}$$

where $\gamma := \sum_{i=1}^N \gamma_i \sigma_i^2$ and c is the unknown wave speed.

3. Multiple pulse and front solutions. Although our interest in this paper is in one-pulse and one-front solutions, in this section we construct candidate solutions with any number of pulses or fronts, i.e., for $m \in \mathbb{Z}^+$ we construct

- m-pulse solutions where $\phi(-\infty) = \phi(+\infty) = 0$ and $\psi(-\infty) = \psi(+\infty) = 0$, homoclinic connections between constant stable equilibrium solution 0 of (2.4), and
- m-front solutions for $r > 0$ such that $\phi(-\infty) = 0$, $\phi(+\infty) = \frac{r}{1+r}$, and $\psi(-\infty) = 0$, $\psi(+\infty) = \frac{1}{1+r}$, heteroclinic connections between constant stable equilibrium solutions of (2.4) for $0 < a < \frac{r}{1+r}$.

Before we begin construction, using linear transforms, we now take a close look at the piecewise linear nonlinearity f and its effects and a close look at the characteristic equation of (2.4).

3.1. The nonlinearity. Because we intend to apply linear transforms to (2.4) we rewrite the piecewise linear nonlinearity as

$$f(\phi(\xi)) = \phi(\xi) - h(\phi(\xi) - a) = \phi(\xi) - \sum_{k=0}^{n} (-1)^k h(\xi - \xi_k), \tag{3.1}$$

where the ξ_k are the unknown values of ξ, where $\phi = a$, $\phi' \neq 0$, $n = 2m - 1$ for pulse solutions and $n = 2m - 2$ for front solutions. This implies that when finding an

m-pulse solution, one also needs to seek the values $\xi_0 < \xi_1 < \cdots < \xi_{2m-1}$ such that $\phi(\xi_k) = a$ for $k = 0, 1, \ldots, 2m-1$ with

- $\phi(\xi) < a$ for $\xi < \xi_0$, for $\xi_k < \xi < \xi_{k+1}$ with k odd, and for $\xi > \xi_{2m-1}$; and
- $\phi(\xi) > a$ for $\xi_k < \xi < \xi_{k+1}$ with k even.

Similarly, for an m-front solution one also needs to seek $0 = \xi_0 < \xi_1 < \cdots < \xi_{2m-2}$ where $\phi(\xi_k) = a$ for $k = 0, 1, \ldots, 2m-2$ with

- $\phi(\xi) < a$ for $\xi < \xi_0$ and for $\xi_k < \xi < \xi_{k+1}$ with k odd; and
- $\phi(\xi) > a$ for $\xi > \xi_{2m-2}$ and for $\xi_k < \xi < \xi_{k+1}$ with k even.

Because the solutions we seek are translationally invariant we pin down the solution by choosing $\xi_0 = 0$.

Remark 3.1. Due to the set-valued nature of the nonlinearity (2.3) and the corresponding Heaviside functions in (3.1) we have that from (2.4)

$$\lim_{\xi \to \xi_k -} c\phi'(\xi) + \gamma\phi''(\xi) \neq \lim_{\xi \to \xi_k +} c\phi'(\xi) + \gamma\phi''(\xi)$$

and

$$\begin{cases} -c\phi'(\xi_k) - \gamma\phi''(\xi_k) \in \displaystyle\sum_{i=1}^{N} d_i(\phi(\xi_k + \sigma_i) - 2\phi(\xi_k) + \phi(\xi_k - \sigma_i)) - f(\phi(\xi_k)) - \psi(\xi_k), \\ -c\psi'(\xi_k) = b(\phi(\xi_k) - r\psi(\xi_k)) \end{cases}$$

for $k = 0, 1, \ldots, n$.

3.2. The characteristic equation. We consider connecting orbits, homoclinic and heteroclinic connections, between homogeneous equilibria. A central aspect is the eigenstructure of the linearization about these equilibrium solutions. In contrast with the case of continuous diffusion in which the characteristic equation is written in terms of a polynomial, in the case of discrete diffusion the characteristic equation is a transcendental equation with an infinite number of solutions. Three aspects are especially important:

(i) that the equilibria are hyperbolic in the sense that there are not purely imaginary solutions to the characteristic equation;
(ii) that the dominant eigenvalues, those with smallest real part among those with positive real part and those with largest real part among those with negative real part, possess a gap (up to complex conjugates in the case of dominant complex eigenvalue) in their real parts with respect to other eigenvalues; and
(iii) whether the dominant eigenvalues are real or a complex conjugate pair.

To study the characteristic equation of (2.4) we begin by linearizing around a constant equilibrium solution such that $f(\phi) \neq a$ to obtain the following linear differential equation:

$$\begin{cases} -cx'(\xi) = \displaystyle\sum_{i=1}^{N} d_i(x(\xi + \sigma_i) - 2x(\xi) + x(\xi - \sigma_i)) + \gamma x''(\xi) - x(\xi) - y(\xi), \\ -cy'(\xi) = b(x(\xi) - ry(\xi)). \end{cases}$$

On substituting $x(\xi) = \kappa_1 \exp(\lambda\xi)$ and $y(\xi) = \kappa_2 \exp(\lambda\xi)$ we obtain

$$\begin{cases} -c\lambda x(\xi) = \displaystyle\sum_{i=1}^{N} d_i(\exp(\lambda\sigma_i) - 2 + \exp(-\lambda\sigma_i))x(\xi) + \gamma\lambda^2 x(\xi) - x(\xi) - y(\xi), \\ -c\lambda y(\xi) = b(x(\xi) - ry(\xi)). \end{cases} \tag{3.2}$$

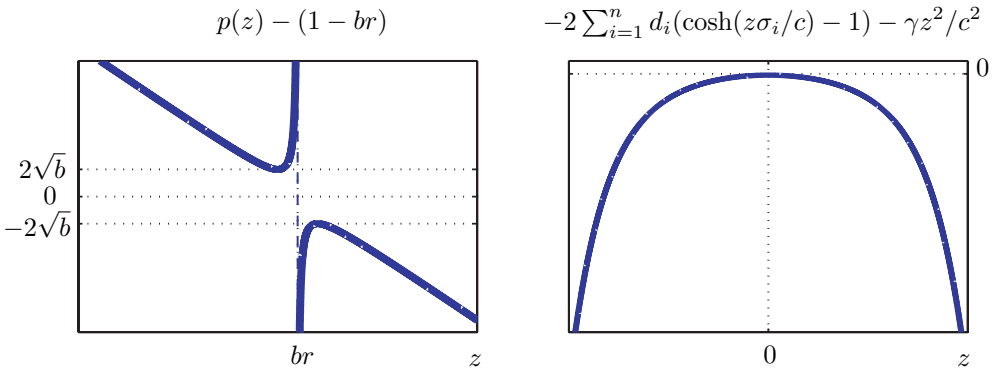


FIG. 3.1. *On the left is a plot of the function* $p(z)-(1-br)$. *To find the real roots of* $p(z)$ *one only needs to plot the constant valued function* $br-1$. *On the right is a plot of* $-2\sum_{i=1}^{n} d_i(\cosh(z\sigma_i/c)-1) - \gamma z^2/c^2$ *when* c *is finite.*

When $br \neq c\lambda$ the second equation of (3.2) is $y(\xi) = bx(\xi)/(br - c\lambda)$, which on substitution into the first equation of (3.2) yields the characteristic equation

$$\Delta(\lambda) := 1 - 2\sum_{i=1}^{n} d_i(\cosh(\lambda\sigma_i) - 1) - \gamma\lambda^2 + \frac{b}{(br - c\lambda)} - c\lambda = 0. \tag{3.3}$$

Consider the change of variables $z = \lambda c$ and let $|c| \to \infty$. Then we have

$$p(z) = 0, \quad \text{where} \quad p(z) \equiv \lim_{|c|\to\infty} \Delta(z/c) = 1 + \frac{b}{(br - z)} - z.$$

For the following it may be illustrative to refer to Figure 3.1. Thus $p'(z) = b/(br-z)^2 - 1$, which equals zero when $z_\pm = br \pm \sqrt{b}$, one value on each side of the vertical asymptote $z = br$. The function p has a maximum of $p(z_+) = 1 - 2\sqrt{b} - br$ at $z_+ = br+\sqrt{b}$ and a minimum of $p(z_-) = 1+2\sqrt{b}-br$ at $z_- = br-\sqrt{b}$. Therefore, in the limit as $|c| \to \infty$, there are two positive real roots (greater than br) to the characteristic equation $\Delta(z/c)$ if $p(z_+) > 0$, i.e., if $r < (1 - 2\sqrt{b})/b$, and two positive real roots (less than br) in the limit if $p(z_-) < 0$, i.e., if $r > (1 + 2\sqrt{b})/b$. If $(1 - 2\sqrt{b})/b < r < (1 + 2\sqrt{b})/b$, then the roots in the limit are complex.

For c finite, the characteristic equation $\Delta(z/c) = p(z) - 2\sum_{i=1}^{n} d_i(\cosh(z\sigma_i/c) - 1) - \gamma z^2/c^2$ always has one negative real root. If the roots in the limit are complex with positive real part, then for all finite c there will not be real positive roots to the characteristic equation. We are interested in cases where the characteristic equation does not admit purely imaginary or zero solutions. In this case the following lemma provides justification for our calculations.

LEMMA 3.1. *Let* (ϕ, ψ) *be a solution of* (2.3), (2.4) *for some* $c \neq 0$. *Then there exists* $\delta_0 > 0$ *such that for some* $K > 0$,

$$|\phi(\xi)| \leq Ke^{\delta_0\xi}, \ |\psi(\xi)| \leq Ke^{\delta_0\xi}, \text{ for } \xi \leq 0.$$

Proof. The proof follows the proof of Lemma 4.1 of [7]; see also the proof of Lemma 3.1 of [13] and the proof of Lemma 2.1 of [12]. ∎

3.3. Construction of candidate solutions. We are now ready to construct candidate solutions by employing the Fourier transform

$$\hat{\phi}_\delta(s) = \int_{-\infty}^{+\infty} e^{-is\xi}\phi_\delta(\xi)d\xi \quad \text{with} \quad \phi_\delta(\xi) = e^{-\delta\xi}\phi(\xi)$$

(and similarly for ψ) and $\delta > 0$ is sufficiently small. Convergence of the integral is guaranteed by Lemma 3.1, which implies that $\phi_\delta(\xi) \to 0$ and $\psi_\delta(\xi) \to 0$ exponentially fast, both as $\xi \to -\infty$ and $\xi \to +\infty$ for $0 < \delta < \delta_0$. Using (2.4) and (3.1) we have that $(\phi_\delta, \psi_\delta)$ satisfy

(3.4)
$$\begin{cases} (-c - 2\gamma\delta)\phi_\delta'(\xi) = \displaystyle\sum_{i=1}^{N} d_i[e^{\delta\sigma_i}\phi_\delta(\xi+\sigma_i) - 2\phi_\delta(\xi) + e^{-\delta\sigma_i}\phi_\delta(\xi-\sigma_i)] + \gamma\ddot{\phi}_\delta(\xi) \\ \qquad\qquad - (1 - c\delta - \gamma\delta^2)\phi_\delta(\xi) + e^{-\delta\xi}\displaystyle\sum_{k=0}^{n}(-1)^k h(\xi - \xi_k) - \psi_\delta(\xi), \\ -c\psi_\delta'(\xi) = b\phi_\delta(\xi) - [br - c\delta]\psi_\delta(\xi). \end{cases}$$

By applying the Fourier transform to (3.4) we obtain the following matrix equation:

$$M(s - i\delta)\begin{pmatrix} \hat{\phi}_\delta(\xi) \\ \hat{\psi}_\delta(\xi) \end{pmatrix} = \frac{1}{is+\delta}\begin{pmatrix} \sum_{k=0}^{n}(-1)^k e^{-is\xi_k} \\ 0 \end{pmatrix} \quad \text{with} \quad M(s) := \begin{pmatrix} R(s) & 1 \\ -b & B(s) \end{pmatrix},$$

where

(3.5)
$$R(s) = -cis + A(s), \quad A(s) = 1 + \gamma s^2 + 2\sum_{i=1}^{N} d_i(1 - \cos(\sigma_i s)), \quad \text{and } B(s) = -cis + br.$$

The matrix function $M(s)$ is invertible near the real axis. To see this note that we have $\det(M(s)) = R(s)B(s) + b$, and the imaginary part of the determinant is bounded away from zero for $s \neq 0$ near the real axis, while for $s = 0$ the real part of the determinant is bounded away from zero since $b > 0$ and $r \geq 0$.

Solving we obtain

$$\hat{\phi}_\delta(s) = \frac{B(s - i\delta)}{(is+\delta)[R(s-i\delta)B(s-i\delta) + b]}\sum_{k=0}^{n}(-1)^k e^{-is\xi_k}$$

and

$$\hat{\psi}_\delta(s) = \frac{b}{(is+\delta)[R(s-i\delta)B(s-i\delta) + b]}\sum_{k=0}^{n}(-1)^k e^{-is\xi_k},$$

and on applying the Fourier inversion theorem we obtain

$$\begin{aligned} \phi(\xi) = e^{\delta\xi}\phi_\delta(\xi) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty}\hat{\phi}_\delta(s)e^{(is+\delta)\xi}ds \\ &= \frac{1}{2\pi i}\int_{-i\delta-\infty}^{-i\delta+\infty}\frac{B(s)}{s(R(s)B(s)+b)}\sum_{k=0}^{n}(-1)^k e^{is(\xi-\xi_k)} \\ &= \frac{1}{2\pi i}\left(\int_{C_\delta} + \int_{S_\delta}\right)\frac{B(s)}{s(R(s)B(s)+b)}\sum_{k=0}^{n}(-1)^k e^{is(\xi-\xi_k)}ds \end{aligned} \tag{3.6}$$

and

$$\psi(\xi) = e^{\delta\xi}\psi_\delta(\xi) = \frac{1}{2\pi i}\left(\int_{C_\delta} + \int_{S_\delta}\right)\frac{b}{s(R(s)B(s)+b)}\sum_{k=0}^{n}(-1)^k e^{is(\xi-\xi_k)}ds, \tag{3.7}$$

where C_δ denotes the two half-lines $(-\infty, -\delta]$ and $[\delta, \infty)$, and S_δ is the half-circle $t \to \delta e^{it}$ for $-\pi \le t \le 0$. Upon simplification, and taking $\delta \to 0$ we next obtain

$$\begin{aligned}\phi(\xi) = \left(\frac{1+(-1)^n}{4}\right)\left(\frac{r}{1+r}\right) &+ \int_0^\infty W(s)\sum_{k=0}^{n}(-1)^k \sin(s(\xi-\xi_k))ds \\ &+ \int_0^\infty X(s)\sum_{k=0}^{n}(-1)^k \cos(s(\xi-\xi_k))ds \quad \text{and} \\ \psi(\xi) = \left(\frac{1+(-1)^n}{4}\right)\left(\frac{1}{1+r}\right) &+ \int_0^\infty Y(s)\sum_{k=0}^{n}(-1)^k \sin(s(\xi-\xi_k))ds \\ &+ \int_0^\infty Z(s)\sum_{k=0}^{n}(-1)^k \cos(s(\xi-\xi_k))ds,\end{aligned} \tag{3.8}$$

where

$$W(s) = \frac{1}{\pi}\left[\frac{b^2 r + C(s)A(s)}{sD(s)}\right], \qquad X(s) = \frac{c}{\pi}\left[\frac{-b+C(s)}{D(s)}\right], \tag{3.9}$$

$$Y(s) = \frac{b}{\pi}\left[\frac{brA(s)+b-c^2s^2}{sD(s)}\right], \quad \text{and} \quad Z(s) = \frac{bc}{\pi}\left[\frac{A(s)+br}{D(s)}\right], \tag{3.10}$$

with $C(s) := b^2r^2 + c^2s^2$ and

$$D(s) := (R(s)B(s)+b)(R(-s)B(-s)+b) = c^2s^2(A(s)+br)^2 + (brA(s) - c^2s^2 + b)^2.$$

Remark 3.2. The construction of candidate solutions is also applicable to more general diffusive operators. For example, consider for $N = 1$, the term (see also [3]) $\sum_{i=1}^{N} d_i(v(\eta+e_i,t) - 2v(\eta,t) + v(\eta-e_i,t)) + \sum_{i=1}^{N}\gamma_i \frac{\partial^2 v}{\partial \eta_i^2}(\eta,t)$ in (2.2) replaced by

$$d\left(-2v(\eta,t) + \sum_{j=1}^{\infty}\alpha_j\{v(\eta+j,t)+v(\eta-j,t)\}\right) + \gamma\frac{\partial^2 v}{\partial \eta_i^2}(\eta,t), \quad \sum_{j=1}^{\infty}\alpha_j = 1.$$

The main change to the derivation is that $A(s)$ in (3.5) becomes

$$A(s) = 1 + \gamma s^2 + 2d\sum_{j=1}^{\infty}\alpha_j(1-\cos(js)).$$

4. Further discussion of one-front candidate solutions. The potential solutions derived in (3.8)–(3.10) are one-front solutions when $n = 0$. Notice that to satisfy the boundary conditions, the detuning parameter a must be restricted so that $a \in [0, r/(1+r)]$. From (3.8)–(3.10) we have the following symmetry property:

$$\phi(\xi, c) = \frac{r}{1+r} - \phi(-\xi, -c) \quad \text{and} \quad \psi(\xi, c) = \frac{1}{1+r} - \psi(-\xi, -c). \tag{4.1}$$

By (3.6) we have that $|\phi(\xi)| = O(e^{\delta\xi})$ as $\xi \to -\infty$, so the boundary condition $\phi(-\infty) = 0$ holds. By (4.1), $\phi(+\infty) = \frac{r}{1+r}$ and the boundary conditions for ψ are satisfied similarly using (3.7) and (4.1).

In certain limits the existence of one-front solutions is known. For instance, if we let $b = \epsilon$ in (2.4), then in the limit as $\epsilon \to 0$, $\psi(\xi) = \psi_0$, a constant, and we can then consider the Nagumo equation with nonlinearity $\tilde{f}(\phi) = f(\phi) - \psi_0$ provided $1 - a < \psi_0 < a$, in which case we seek a one-front solution such that $\phi(-\infty) = \psi_0$ and $\phi(+\infty) = 1 + \psi_0$. In this case previous results (see, e.g., [7]) concerning one-front solutions scaled so that $\tilde{\phi}(-\infty) = 0$ and $\tilde{\phi}(+\infty) = 1$ are applicable by considering a fixed ψ_0 and considering the correspondence $\phi(\xi) = \tilde{\phi}(\xi) + \psi_0$.

Similarly, if we let $r = 1/\epsilon$ and let $\epsilon \to 0$, then $\psi_0 = 0$ and results in [7] on propagation failure, monotonicity of one-front solutions, and monotonicity of the (a, c) relationship are directly applicable. Furthermore, we expect these behaviors to persist in a neighborhood of the limiting parameter value.

4.1. Verification of candidate one-front solution. First, we have assumed $\phi(0) = a$, so by (3.8),

$$a = \frac{1}{2}\left(\frac{r}{1+r}\right) + \int_0^\infty X(s)ds. \tag{4.2}$$

The candidate solution found in (3.8)–(3.10) is consistent with our ansatz of (3.1) with $n = 0$ provided $\phi(\xi) > a$ for $\xi > 0$ and $\phi(\xi) < a$ for $\xi < 0$, where ϕ is defined by (3.8) and a is defined in (4.2). Since the boundary conditions are satisfied, if the roots of the corresponding characteristic equation (3.3) do not lie on the imaginary axis, then for $|\xi|$ large enough the solution ϕ in (3.8) is bounded away from a. Thus it is enough to check $\phi(\xi) > a$ for $\xi > 0$ and $\phi(\xi) < a$ for $\xi < 0$ over a finite interval of values ξ. Clearly, we expect to have one-front solutions for $b \approx 0$ and for r large enough.

5. Further discussion of one-pulse candidate solutions. The derivation in section 3 relied on the assumption that there exists an m-pulse solution (or a front solution). This allowed us to write the nonlinear term as a linear term and a sum of Heaviside functions. In this section we give conditions under which these assumptions may be verified for one-pulse solutions.

Existence of ξ_1. Our assumption for one-pulse solutions was that $\phi(0) = a$ and $\phi(\xi_1) = a$ for some $\xi_1 > 0$ and $\phi(\xi) < a$ for $\xi < 0$ and for $\xi > \xi_1$ with $\phi(\xi) > a$ for $0 < \xi < \xi_1$. Using the form of the candidate solutions (3.8), (3.9), and (3.10), to have $\phi(0) = \phi(\xi_1) = a$, there must be $\xi_1 > 0$ such that $g(\xi_1) = 0$, where

$$\int_0^\infty X(s)(2 - 2\cos(s\xi))ds = \frac{2c}{\pi}\int_0^\infty \frac{-b + C(s)}{D(s)}(1 - \cos(s\xi_1))ds \equiv \frac{2c}{\pi}g(\xi).$$

The existence of such a $\xi_1 > 0$ that satisfies $g(\xi_1) = 0$ for $c \neq 0$ is a *necessary* condition for the existence a one-pulse solution to (2.3) and (2.4). Let

$$Q(s) = -b + C(s),$$

so we can write

$$g(\xi_1) = \int_0^\infty \frac{Q(s)}{D(s)}(1 - \cos(s\xi_1))ds. \tag{5.1}$$

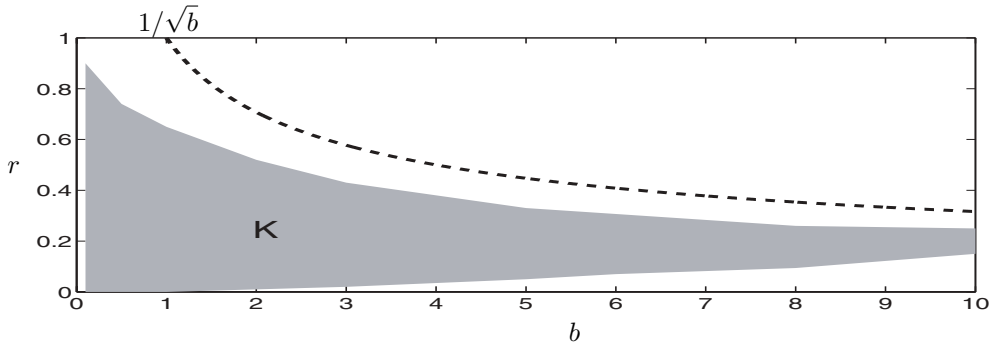


FIG. 5.1. *In this example, the shaded area K indicates values of* (b, r) *which satisfy the conditions of Theorem* 5.1. *The dashed line is* $br^2 = 1$. $N = \gamma = d_1 = c = 1$.

Remark 5.1. This same type of idea could be applied to verifying the existence of general m-pulse solutions. However, in that case, the existence of a zero for a system of $2m - 1$ equations in $2m - 1$ unknowns must be verified.

The idea behind the following theorem is to show that for $\xi_1 > 0$ g is positive for ξ_1 small and g is negative for ξ_1 large. It is motivated by the situation for the spatially continuous problem (2.1) with the piecewise linear nonlinearity (2.3). In the spatially continuous diffusion operator case, $D(s)$ is proportional to s^6 as $s \to \infty$ so that g, g', and g'' are defined by absolutely convergent integrals.

THEOREM 5.1. *There exists a positive zero of* g *defined in* (5.1) *for wave speed* $c \neq 0$ *provided* $br^2 < 1$,

$$\int_0^\infty \frac{Q(s)}{D(s)} ds < 0, \tag{5.2}$$

and for all $\nu > 0$ *sufficiently small*

$$-\int_0^{s^*} s^2 \frac{Q(s)}{D(s)} ds < \int_{s^*}^\infty s^2 \frac{Q(s)}{D(s) + \nu s^6} ds, \tag{5.3}$$

where $s^* = \sqrt{\frac{b(1-br^2)}{c^2}}$.

The shaded region of Figure 5.1 illustrates values of (b, r) which satisfy the conditions of this theorem, for $N = \gamma = d_1 = c = 1$. They were verified by comparison, bounding $D(s)$ with functions of the form $c_1 s^6 + c_2 s^3 + c_3$.

Proof. We have $D(s) > 0$ for $s \geq 0$ and since $br^2 < 1$, $Q(s^*) = 0$, $Q(s) < 0$ for $s < s^*$, and $Q(s) > 0$ for $s > s^*$. We want to show that $g(\xi_1) > 0$ for $\xi_1 > 0$ sufficiently small and $g(\xi_1) < 0$ for $\xi_1 > 0$ sufficiently large.

To this end note that for any bounded continuous function κ defined for $s \geq 0$, for $x > 0$

$$\int_0^\infty \kappa(s) \cos(sx) ds = \int_0^\infty \frac{1}{x} \kappa(u/x) \cos(u) du \to 0 \quad \text{as } x \to \infty.$$

Thus, $g(\xi_1) < 0$ for $\xi_1 > 0$ sufficiently large follows from (5.2).

Next observe that $g(0) = 0$ and define

$$h(\xi_1) = \int_0^{s^*} \frac{Q(s)}{D(s)} (1 - \cos(s\xi_1)) ds + \int_{s^*}^\infty \frac{Q(s)}{D(s) + \nu s^6} (1 - \cos(s\xi_1)) ds.$$

Then

$$h'(\xi_1) = \int_0^{s^*} s\frac{Q(s)}{D(s)}\sin(s\xi_1)ds + \int_{s^*}^{\infty} s\frac{Q(s)}{D(s)+\nu s^6}\sin(s\xi_1)ds,$$

and so $h'(0) = 0$. Similarly,

$$h''(\xi_1) = \int_0^{s^*} s^2\frac{Q(s)}{D(s)}\cos(s\xi_1)ds + \int_{s^*}^{\infty} s^2\frac{Q(s)}{D(s)+\nu s^6}\cos(s\xi_1)ds,$$

and since for $s \geq s^*$, $Q(s) \geq 0$ and $(D(s)+\nu s^6)^{-1} < (D(s))^{-1}$, (5.3) implies that g is increasing for $\xi_1 = 0$, so that $g(\xi_1) > 0$ for $\xi_1 > 0$ sufficiently small. Thus, since g is continuous and changes sign at least once, there exists a $\xi_1 > 0$ such that $g(\xi_1) = 0$. □

COROLLARY 5.1. *There exists a positive zero of g defined in* (5.1) *for wave speed $c \neq 0$ provided $br^2 < 1$ and either*

(i) *for $d_1 = \cdots = d_N = 0$ and $\gamma > 0$,* (5.2) *holds and* (5.3) *holds with $\nu = 0$, or*

(ii) *for $d_j \geq 0$, with at least one $d_j > 0$, $j = 1, \ldots, N$, and $\gamma = 0$,* (5.2) *holds.*

Proof. When $d_1 = \cdots = d_N = 0$ and $\gamma > 0$, then the integral on the right-hand side of (5.3) is absolutely convergent with $\nu = 0$. However, for $d_j \geq 0$, with at least one $d_j > 0$, and $\gamma = 0$, the right-hand side of (5.3) approaches $+\infty$ as $\nu \to 0$, and so in this case (5.3) is always satisfied. □

6. Existence of solutions. An important aspect of employing the McKean nonlinearity (2.3) is that (3.8)–(3.10) provide an explicit form (up to quadrature) for candidate solutions. This explicit form is useful in determining $\{\xi_k\}_k^n$ since for $\xi_0 = 0$ and $n \geq 1$, $\{\xi_k\}_k^n$ satisfies the system of nonlinear equations

$$\phi(0) - \phi(\xi_k) = 0, \quad k = 1, \ldots, n, \tag{6.1}$$

and in subsequently verifying

$$(-1)^k(\phi(0) - \phi(\xi)) > 0, \quad \xi \in (\xi_{k-1}, \xi_k), \quad k = 0, \ldots, n, \tag{6.2}$$

where we have set $\xi_{-1} = -\infty$ and $\xi_{n+1} = +\infty$. A necessary condition for (6.2) is that

$$(-1)^k \lim_{\xi \to \xi_k \pm} \phi'(\xi) > 0, \tag{6.3}$$

which for $c > 0$ involves only a one-sided limit since

$$(-1)^k \lim_{\xi \to \xi_k -} \phi'(\xi) > (-1)^k \lim_{\xi \to \xi_k +} \phi'(\xi). \tag{6.4}$$

One needs only to determine if the inequalities (6.2) are satisfied outside a neighborhood of $\{\xi_k\}_{k=1}^n$ if (6.3) holds.

The existence of $\{\xi_k\}_{k=0}^n$ is trivial for one-front solutions, $n = 0$, since we may choose $\xi_0 = 0$ by translation invariance, while for one-pulse solutions, $n = 1$, Theorem 5.1 and Corollary 5.1 provide criteria for the existence of $\xi_1 > 0$. In general for $n \geq 2$ establishing the existence of $\{\xi_k\}_k^n$ such that (6.1) holds is more difficult since this results in a nonlinear system of $n \geq 2$ equations in n unknowns. In the case of $n \geq 2$ perturbation/continuation techniques are more promising. One-front ($n = 0$) and one-pulse ($n = 1$) solutions provide building blocks for more general $n \geq 2$ solutions.

We are interested in connecting orbits between hyperbolic equilibria in which there is gap between

(1) the dominant solution(s) to the characteristic equation
(2) and the rest of the solutions to the characteristic equation.

The following lemmas show that for the characteristic equation (3.3) of (2.4) with (2.3) with $N = 1$, $d_1 = d > 0$, and $\gamma = 0$

- there are no purely imaginary solutions
- and if $\sinh((1 + 2d - \sqrt{4d^2 + c^2})/c) \neq -c/(2d)$, then all the solutions are simple.

LEMMA 6.1. *Let λ_1, λ_2 denote two solutions to the characteristic equation* (3.3). *Write $\lambda_j = a_j + ib_j$ for $j = 1, 2$ and let $c > 0$ be given. If $(br - ca_1)^2 + (cb_j)^2 > 0$ for $j = 1, 2$, then $a_1 = a_2$ implies $b_1 = \pm b_2$.*

Proof. We have that λ_j is a solution of the characteristic equation provided

$$\begin{aligned} &(br - ca_j)R_j + cb_jI_j = 0, && R_j = 1 - 2d(\mathrm{Re}(\cosh(\lambda_j)) - 1) - ca_j, \\ &\qquad\qquad \text{where} \\ &-cb_jR_j + (br - ca_j)I_j = 0, && I_j = -2d(\mathrm{Im}(\cosh(\lambda_j))) - cb_j, \end{aligned}$$

for $j = 1, 2$. Thus, if $a_1 = a_2$ and $(br - ca_1)^2 + (cb_j)^2 > 0$, then $R_j = 0$ and $I_j = 0$ for $j = 1, 2$, and so

$$1 - 2d(\cos(b_j)\cosh(a_j) - 1) - ca_j = 0, \quad -2d\sin(b_j)\sinh(a_j) - cb_j = 0 \tag{6.5}$$

for $j = 1, 2$. Using the first equation in (6.5), if $a_1 = a_2$, then $\cos(b_1) = \cos(b_2)$, and so $\sin(b_1) = \pm\sin(b_2)$. If $\sin(b_1) = \sin(b_2)$, then the second equation in (6.5) implies that $c(b_1 - b_2) = 0$, and if $\sin(b_1) = -\sin(b_2)$, then the second equation in (6.5) gives $c(b_1 + b_2) = 0$. ∎

LEMMA 6.2. *If $c > 0$, $b > 0$, $r \geq 0$, and $d > 0$, then any root $\lambda = x + iy$ of* (3.3) *has $x \neq 0$.*

Proof. If we write $\Delta(\lambda) = 0$ as $N(\lambda) + b/(br - c\lambda) = 0$, then $r = 0$ implies $\lambda \neq 0$. If $r > 0$ and $\lambda = 0$ is a solution, then $N(0) + 1/r = 0$, so $1 - 2d(\cos(0) - 1) + 1/r = 0$, which cannot occur for $r > 0$. If $x = 0$, but $y \neq 0$, then by the argument in the proof of Lemma 6.1, $\mathrm{Re}(N(\lambda)) = 0$ and $\mathrm{Im}(N(\lambda)) = 0$, so we have $1 + 2d - \cos(y) = 0$ and $-cy = 0$, which cannot both be simultaneously satisfied. ∎

LEMMA 6.3. *If for $c > 0$ and $d > 0$ one has $\sinh((1 + 2d - \sqrt{4d^2 + c^2})/c) \neq -c/(2d)$, then there does not exist a double root to* (3.3) *for λ not purely imaginary.*

Proof. If there is a double root, λ, then $\Delta(\lambda) = 0$ and $\Delta'(\lambda) = 0$. Write $\Delta(\lambda)$ as $N(\lambda) + b/(br - c\lambda)$, so $\Delta'(\lambda) = N'(\lambda) + cb/(br - c\lambda)^2$. Thus, $\Delta'(\lambda) = N'(\lambda) + \frac{c}{b}N^2(\lambda)$, so by the argument in the proof of Lemma 6.1, $\mathrm{Re}(N(\lambda)) = 0$ and $\mathrm{Im}(N(\lambda)) = 0$, and if there is a double root, then $\Delta'(\lambda) = N'(\lambda) = 0$. Then for $\lambda = x + iy$, $-2d\cos(y)\sinh(x) - c = 0$ and $-2d\sin(y)\cosh(x) = 0$, we have $\sin(y) = 0$, and since $\mathrm{Im}(N(\lambda)) = 0$, we have $-2d\sin(y)\sinh(x) - cy = 0$, so $y = 0$. Thus, $-2d\sinh(x) - c = 0$, which implies $\sinh(x) = -c/(2d)$ and $\cosh(x) = \sqrt{4d^2 + c^2}/(2d)$. Hence, the only way for $\mathrm{Re}(N(\lambda)) = 0$ is if $x = (1 + 2d - \sqrt{4d^2 + c^2})/c$. ∎

6.1. Existence of one-pulse solutions. Given existence of $\xi_1 > 0$ such that $g(\xi_1) = 0$ for g in (5.1), we now turn our attention to showing existence of one-pulse solutions. We proceed in two ways: in the first we assume the existence at a particular value of the wave speed c and then show existence (under certain conditions) in a neighborhood (Theorem 6.4); in the second we verify that $\phi(\xi) < a$ only when $\xi < 0$ and $\xi > \xi_1$ ($\phi(\xi) > a$ for $\xi \in (0, \xi_1)$) and then show that this condition can in certain cases be reduced to checking on finite intervals (Theorems 6.5 and 6.6).

THEOREM 6.4. *Suppose for fixed values of the parameters and* $(a, c) = (a^*, c^*)$ *with* $c^* > 0$ *we have a one-pulse solution defined for* $\xi_1 > 0$ *such that*

$$\text{for } \xi \in (0, \xi_1),\ \phi(\xi) - \phi(0) > 0, \quad \text{for } \xi < 0 \text{ and } \xi > \xi_1,\ \phi(\xi) - \phi(0) < 0, \tag{6.6}$$

$$\lim_{\xi \to 0^{\pm}} \phi'(\xi) > 0 \text{ and } \lim_{\xi \to \xi_1^{\pm}} \phi'(\xi) < 0, \tag{6.7}$$

and

$$0 < \left| \int_0^{\infty} \frac{sQ(s)}{D(s)} \sin(s\xi_1) ds \right| < \infty. \tag{6.8}$$

Then for all c *in a neighborhood of* c^* *there exist one-pulse solutions to* (2.3), (2.4).

Proof. We argue by the implicit function theorem. We need to show that ξ_1 depends smoothly on c. This follows from (6.8) since by direct calculation

$$\xi_1'(c) = -\frac{\int_0^{\infty} \frac{Q_1(s)}{D^2(s)}(1 - \cos(s\xi_1)) ds}{\int_0^{\infty} \frac{sQ(s)}{D(s)} \sin(s\xi_1) ds}, \tag{6.9}$$

where $Q_1(s) = Q_c(s)D(s) - Q(s)D_c(s)$ and Q_c and D_c are the derivatives of Q and D with respect to c, respectively. Now by the implicit function theorem, (6.6) and (6.7) hold in a neighborhood of c^* since the derived solution (3.8)–(3.10) and the one-sided derivatives at $\xi = 0$ and $\xi = \xi_1$ depend smoothly on ξ_1 and c. ∎

Inequalities (6.2), when satisfied over the entire real line, imply existence; see Theorem 6.5. Lemmas 6.1–6.3 that show the hyperbolicity of the equilibria and the gap condition are used to show that is sufficient to check these inequalities over a certain finite interval; see Theorem 6.6.

THEOREM 6.5. *If* $g(\xi_1) = 0$ *for* $\xi_1 > 0$ *and*

$$\int_0^{\infty} W(s)[\sin(s(\xi_1 + \delta)) - \sin(s\delta) - \sin(s\xi_1)] ds < \left| \int_0^{\infty} X(s)[\cos(s(\xi_1 + \delta)) - \cos(s\delta)] ds \right|, \tag{6.10}$$

if $\delta > 0$,

$$-\int_0^{\infty} W(s)[\sin(s(\xi_1 + \delta)) - \sin(s\delta) - \sin(s\xi_1)] ds < \int_0^{\infty} X(s)[\cos(s(\xi_1 + \delta)) - \cos(s\delta)] ds, \tag{6.11}$$

if $-\xi_1 < \delta < 0$, *then there exists a one-pulse solution to* (2.3), (2.4).

Proof. We assume there exists a positive zero, ξ_1, of g. Then for this $\xi_1 > 0$ we have $\phi(0) = a$ and $\phi(\xi_1) = a$. We show that (6.10) implies that $\phi(\xi) < a$ for $\xi < 0$ and $\xi > \xi_1$ and (6.11) implies $\phi(\xi) > a$ for $0 < \xi < \xi_1$. Recalling that $g(\xi_1) = 0$ implies $\int_0^{\infty} X(s)(1 - \cos(s\xi_1))\, ds = 0$ along with the solution equalities (3.8)–(3.10), (6.10) can be rewritten as

$$-\phi(0) < -\phi(-\delta) \qquad \text{and} \qquad -\phi(0) < -\phi(\delta + \xi_1), \qquad \delta > 0,$$

which implies $\phi(\xi) < a \equiv \phi(0)$ for $\xi < 0$ and for $\xi > \xi_1$. Similarly (6.11) can be rewritten as

$$\phi(0) < \phi(\delta + \xi_1), \qquad -\xi_1 < \delta < 0,$$

which implies $\phi(\xi) > a \equiv \phi(0)$ for $0 < \xi < \xi_1$. ∎

THEOREM 6.6. *Suppose for some value $c > 0$ and all other parameter values fixed there exists $\xi_1 > 0$ such that $g(\xi_1) = 0$. Suppose that the characteristic equation (3.3) has no solutions on the imaginary axis and order the solutions with positive real parts by the real parts: $\mathrm{Re}(\lambda_1^+) \le \mathrm{Re}(\lambda_2^+) \le \mathrm{Re}(\lambda_3^+) \le \cdots$ and similarly for the solutions with negative real parts as: $\mathrm{Re}(\lambda_1^-) \ge \mathrm{Re}(\lambda_2^+) \ge \mathrm{Re}(\lambda_3^+) \ge \cdots$. If $\lambda_1^+ = \bar{\lambda}_2^+$ assume $\epsilon_0^+ := \mathrm{Re}(\lambda_3^+) - \mathrm{Re}(\lambda_2^+) > 0$. If $\lambda_1^+ \neq \bar{\lambda}_2^+$ assume $\epsilon_0^+ := \mathrm{Re}(\lambda_2^+) - \mathrm{Re}(\lambda_1^+) > 0$. Similarly, assume $\epsilon_0^- > 0$ for the solutions with negative real parts. Then there exist $T^+ > 0$ and $T^- < 0$ such that (6.10) and (6.11) need only be checked on $(0, T^+)$ and $(0, -T^-)$, respectively.*

Proof. Consider (3.6) and observe that if $\lambda \in \mathbb{C}$ is a solution of the characteristic equation (3.3), then $-i\lambda$ is a zero of $R(s) + b/B(s)$ in (3.6). First assume that $0 < \epsilon^+ < \epsilon_0^+$ and shift the contour in (3.6) from $\mathrm{Im}\, s = -\delta$ to $\mathrm{Im}\, s = -(\mathrm{Re}(\lambda_1^+) + \epsilon^+)$. Then there are two cases: $\lambda_1^+ = \bar{\lambda}_2^+$ and $\lambda_1^+ \neq \bar{\lambda}_2^+$. If $\lambda_1^+ = \bar{\lambda}_2^+$, then we obtain

$$\begin{aligned}\phi(\xi) &= -\frac{B(s)\{e^{is\xi} - e^{is(\xi-\xi_1)}\}}{s(R'(s)B(s) + R(s)B'(s))}\bigg|_{s=-i\lambda_1^+} - \frac{B(s)\{e^{is\xi} - e^{is(\xi-\xi_1)}\}}{s(R'(s)B(s) + R(s)B'(s))}\bigg|_{s=-i\lambda_2^+} \\ &\quad + \frac{1}{2\pi i}\int_{-i(\mathrm{Re}(\lambda_1^+)+\epsilon^+)-\infty}^{-i(\mathrm{Re}(\lambda_1^+)+\epsilon^+)+\infty} \frac{B(s)\{e^{is\xi} - e^{is(\xi-\xi_1)}\}}{s(R(s)B(s) + b)} ds \\ &= C^+\{e^{\lambda_1^+\xi} - e^{\lambda_1^+(\xi-\xi_1)}\} + O(e^{(\lambda_1^+ + \epsilon^+)\xi})\end{aligned}$$

as $\xi \to -\infty$. Observe that

$$C^+ = -\frac{B(s)}{s(R'(s)B(s) + R(s)B'(s))}\bigg|_{s=-i\lambda_1^+} - \frac{B(s)}{s(R'(s)B(s) + R(s)B'(s))}\bigg|_{s=-i\lambda_2^+}.$$

If $\lambda_1^+ \neq \bar{\lambda}_2^+$ (and λ_1^+, λ_2^+ real), then we obtain

$$\begin{aligned}\phi(\xi) &= -\frac{B(s)\{e^{is\xi} - e^{is(\xi-\xi_1)}\}}{s(R'(s)B(s) + R(s)B'(s))}\bigg|_{s=-i\lambda_1^+} \\ &\quad + \frac{1}{2\pi i}\int_{-i(\lambda_1^+ + \epsilon^+)-\infty}^{-i(\lambda_1^+ + \epsilon^+)+\infty} \frac{B(s)\{e^{is\xi} - e^{is(\xi-\xi_1)}\}}{s(R(s)B(s) + b)} ds \\ &= C^+\{e^{\lambda_1^+\xi} - e^{\lambda_1^+(\xi-\xi_1)}\} + O(e^{(\lambda_1^+ + \epsilon^+)\xi})\end{aligned}$$

as $\xi \to -\infty$, where $C^+ = -\frac{B(s)}{s(R'(s)B(s)+R(s)B'(s))}|_{s=-i\lambda_1^+}$. Thus, there exists $T^- < 0$ such that $\phi(\xi) < \phi(0) =: a$ for $\xi \le T^-$.

The argument for solutions to the characteristic equation (3.3) with negative real parts is treated similarly by moving the contour up. ∎

7. Numerical results. In this section we present numerical results obtained by numerical integration of (3.8)–(3.10) and for one-pulse solutions by determining the positive zeros of (5.1). We approximate the integrals using the adaptive Gaussian quadrature code `adapt` of [37] after truncation to the interval $[0, 10^6]$. To find zeros of g we use the combined secant/bisection code `zero` of [37]. We focus on one-front, one-pulse, and two-pulse solutions, exhibit (a, c) curves, and waveforms of (3.8)–(3.10) with both continuous and discrete diffusion.

7.1. (a, c) curves. In Figure 7.1 we plot (a, c) curves for one-front solutions by approximating (4.2). We set $N = d_1 = d = 1$ and $\gamma = 0$ and vary b and r.

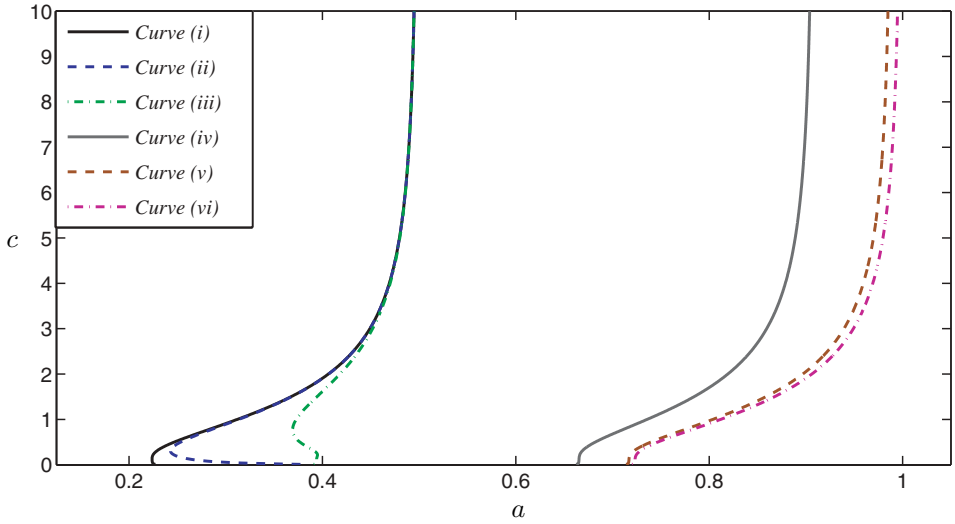


FIG. 7.1. *Plot of* (a, c) *curves (moving from left to right) for* (i) $(d, \gamma, b, r) = (1, 0, 10^{-4}, 1)$, (ii) $(d, \gamma, b, r) = (1, 0, 10^{-2}, 1)$, (iii) $(d, \gamma, b, r) = (1, 0, 1, 1)$, (iv) $(d, \gamma, b, r) = (1, 0, 1, 10^{1})$, (v) $(d, \gamma, b, r) = (1, 0, 1, 10^{2})$, *and* (vi) $(d, \gamma, b, r) = (1, 0, 0, 0)$, *the discrete Nagumo equation.*

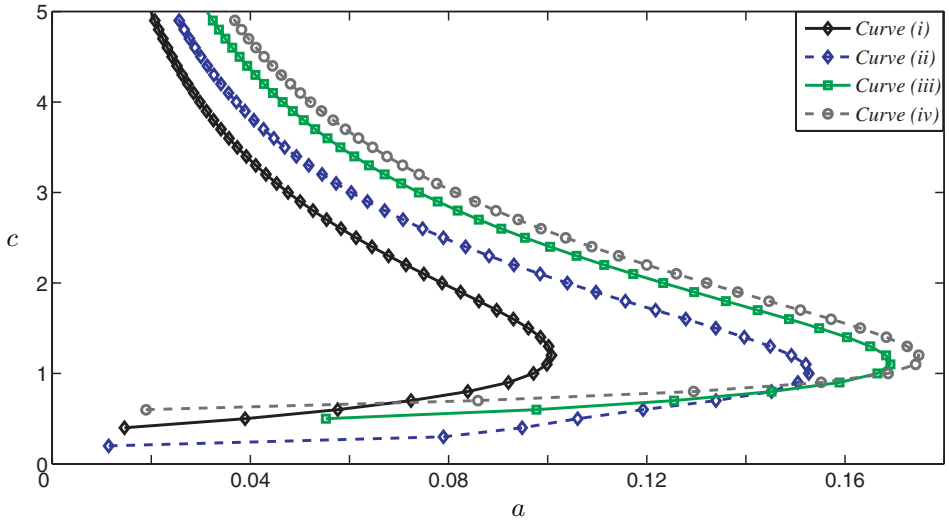


FIG. 7.2. *Plot of* (a, c) *curves (moving from left to right) for* (i) $(d, \gamma, b, r) = (1, 0, 1, 0)$, (ii) $(d, \gamma, b, r) = (0.8, 0.2, 1, 0)$, (iii) $(d, \gamma, b, r) = (0.4, 0.6, 1, 0)$, *and* (iv) $(d, \gamma, b, r) = (0, 1, 1, 0)$.

Notice the nonuniqueness suggested for $(b, r) = (10^{-2}, 1)$ and $(b, r) = (1, 1)$. The curves limit to $r/(r + 1)$ as $c \to \infty$. In Figure 7.2 we plot (a, c) curves for one-pulse solutions (these are actually (a, c) curves obtained when there exists $\xi_1 > 0$ such that $g(\xi_1) = 0$) for various values of the parameters d, γ, b, r. The plot illustrates the difference between the behavior with continuous and discrete diffusion. In the case of continuous diffusion it is known that the fast waves, i.e., those above the tip on the (a, c) curve, are stable (see [43, 30, 25, 29]), while the slow waves (those below the tip)

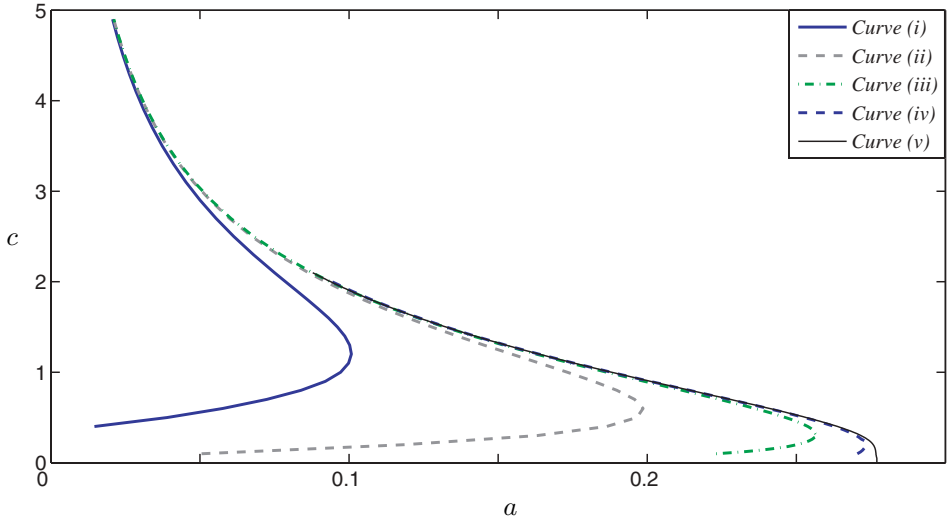


FIG. 7.3. *Plot of* (a, c) *curves (moving from left to right) for* (i) $(d, \gamma, b, r) = (1, 0, 1, 0)$, (ii) $(d, \gamma, b, r) = (1, 0, 10^{-1}, 0)$, (iii) $(d, \gamma, b, r) = (1, 0, 10^{-2}, 0)$, (iv) $(d, \gamma, b, r) = (1, 0, 10^{-3}, 0)$, *and the limiting* (a, c) *curve obtained from the discrete Nagumo equation.*

are unstable. In Figure 7.3 we highlight the dependence of the (a, c) relationship on the parameter $b > 0$ and compare it with the limiting (a, c) curve obtained from the discrete Nagumo equation, i.e., $b = 0$. Notice how the range of propagation failure in the discrete Nagumo equation limits tip of the (a, c) curve. The range of propagation failure limits the size of a for which there are one pulse solutions. Observe the larger values of a that are possible when b is small and the larger values of a obtained for the plot of the continuous (a, c) curve with $\gamma = 1$ as compared to the discrete for $d = 1$. In Figure 7.4 we vary the parameter $r > 0$. We set $b = 1$ and then have the requirement from Theorem 5.1 that $r^2 < 1$. In the plot for $r > 0$ there is an upper bound in c as well as a lower bound in c on the (a, c) curve.

7.2. Waveforms. In Figure 7.5 we plot one-front waveforms $\phi(\xi)$ and $\psi(\xi)$ fixing $(d, c, \gamma) = (1, 1, 0)$ and varying (b, r). In Figure 7.6 we plot one-pulse waveforms $\phi(\xi)$ and $\psi(\xi)$ fixing $(b, r) = (1, 0)$, setting $(d, \gamma) = (1, 0)$, and varying the wavespeed c. Refer to Figure 7.2 for the parameter values in the (a, c) curve. The plot for $c = 0.6$ does not satisfy our assumption of a one-pulse solution since it violates $\phi(\xi) < a$ for $\xi > \xi_1$. Note, however, that $c = 0.6$ is one of the smaller values of c obtained in the (a, c) curve in Figure 7.2. In Figure 7.7 we plot waveforms $\phi(\xi)$ and $\psi(\xi)$ fixing $(b, r) = (1, 0)$, setting $(d, \gamma) = (0, 1)$, and varying the wavespeed c. The waveforms for the continuous operator are smooth compared to the waveforms for the discrete operator especially for small wavespeeds. The two-pulse solution in Figure 7.8 is obtained by superimposing two identical one-pulse solutions, using the superimposed one-pulse solutions as an initial guess and then applying Newton's method. The one-pulse solution is obtained from $(d, \gamma, c, b, r) = (1, 0, 1, 1, 0)$ and the pulses are put at a distance (in ξ) of 40 units apart. The value of a is slightly perturbed from the value of a for the one-pulse as one might expect (see [41]). In Figure 7.9 we plot the dependence of the parameter a on the distance between the pulses, $\xi_2 - \xi_1$, and compare with the value of a obtained for the one-pulse solution.

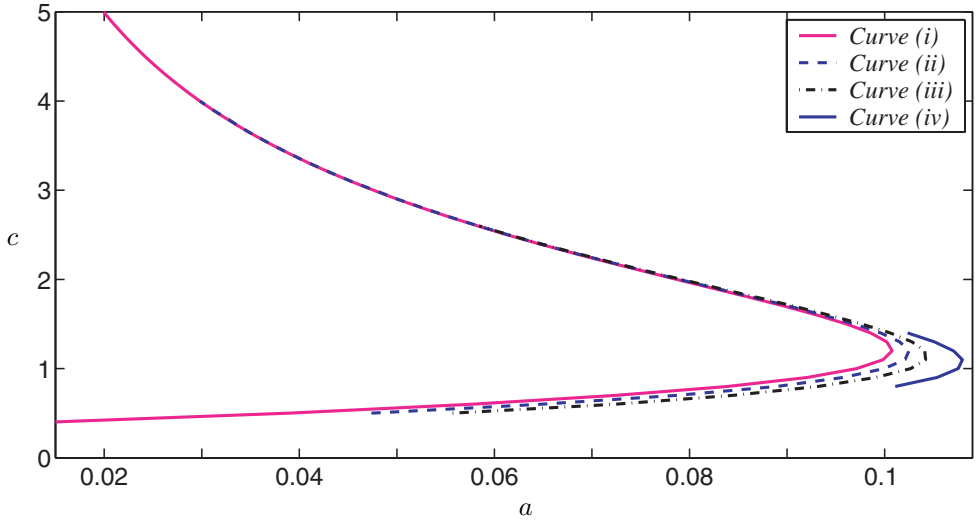


FIG. 7.4. *Plot of* (a, c) *curves (moving from left to right) for* (i) $(d, \gamma, b, r) = (1, 0, 1, 0)$, (ii) $(d, \gamma, b, r) = (1, 0, 1, 1/16)$, (iii) $(d, \gamma, b, r) = (1, 0, 1, 1/8)$, *and* (iv) $(d, \gamma, b, r) = (1, 0, 1, 1/4)$.

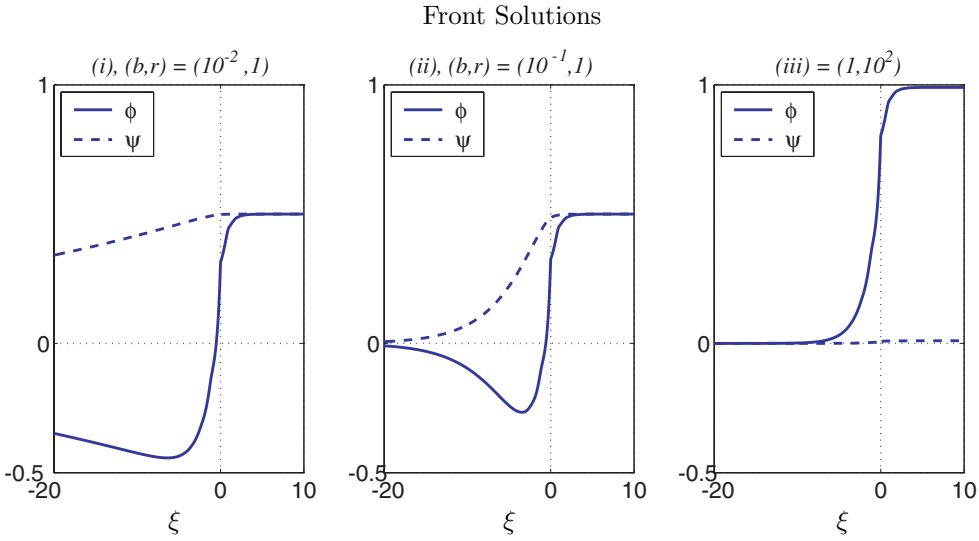


FIG. 7.5. *Plot of one-front waveforms* $\phi(\xi)$ *and* $\psi(\xi)$ *for fixed* $(d, c, \gamma) = (1, 1, 0)$, *and* (i) $(b, r) = (10^{-2}, 1)$, (ii) $(b, r) = (10^{-1}, 1)$, *and* (iii) $(b, r) = (1, 10^2)$.

7.3. Mixed continuous/discrete model. As an example of a model with continuous diffusion in one direction and discrete diffusion in the other coordinate direction we consider (2.2) with $N = 2$, $d_1 = d$, $d_2 = 0$, and $\gamma_1 = 0$, $\gamma_2 = \gamma$. Consider the direction of propagation $\sigma \in \mathbb{R}^2$ such that $||\sigma||_2 = 1$, and apply the traveling wave ansatz $u(\eta, t) = \phi(\eta \cdot \sigma - ct)$ and $w(\eta, t) = \psi(\eta \cdot \sigma - ct)$ to obtain (see (2.4))

$$(7.1)\quad \begin{cases} -c\phi'(\xi) = d(\phi(\xi + \sigma_1) - 2\phi(\xi) + \phi(\xi - \sigma_1)) + \gamma\sigma_2^2\phi''(\xi) - f(\phi(\xi)) - \psi(\xi), \\ -c\psi'(\xi) = b(\phi(\xi) - r\psi(\xi)). \end{cases}$$

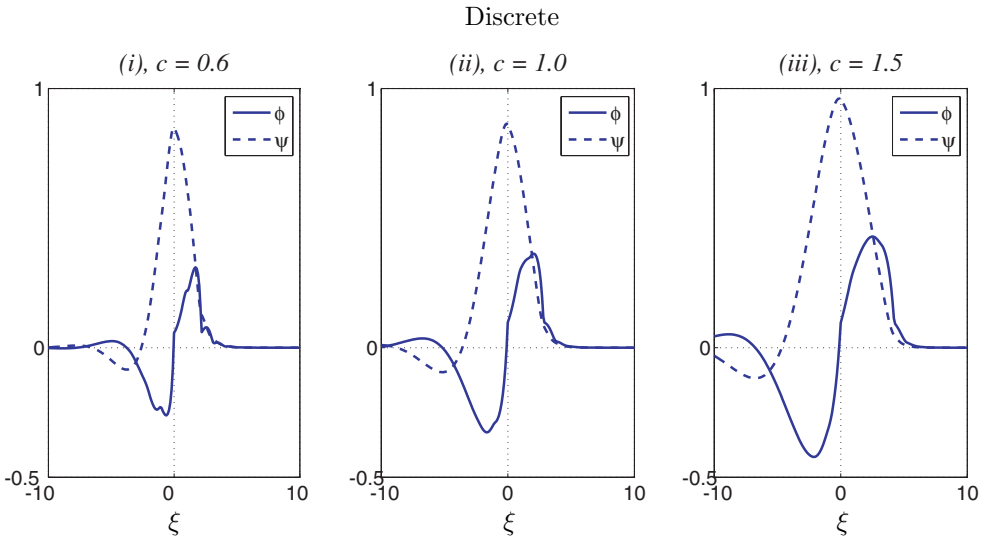


FIG. 7.6. *Discrete. Plot of waveforms for fixed* $(b, r) = (1, 0)$*, and* (i) $(d, \gamma, c) = (1, 0, 0.6)$, (ii) $(d, \gamma, c) = (1, 0, 1)$, (iii) $(d, \gamma, c) = (1, 0, 1.5)$.

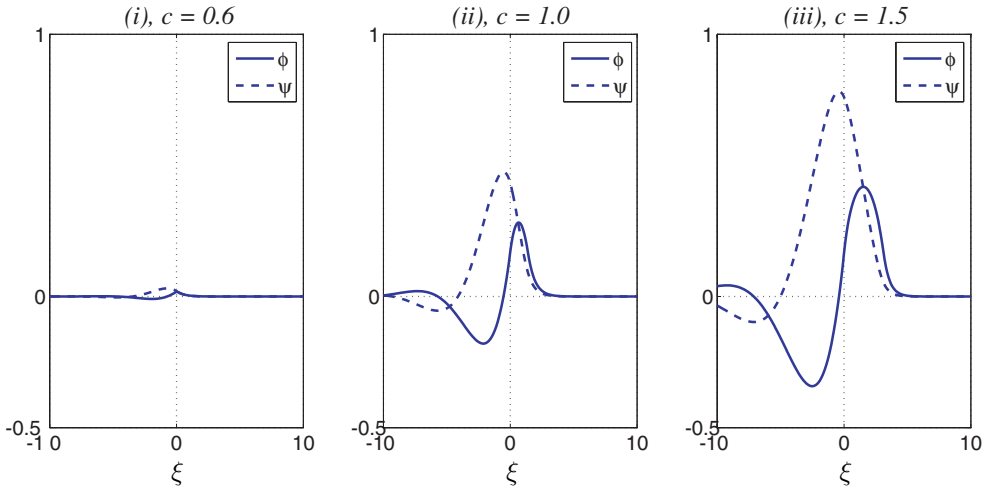


FIG. 7.7. *Continuous. Plot of waveforms for fixed* $(b, r) = (1, 0)$*, and* (i) $(d, \gamma, c) = (0, 1, 0.6)$, (ii) $(d, \gamma, c) = (0, 1, 1)$, (iii) $(d, \gamma, c) = (0, 1, 1.5)$.

If we rescale variables $x = \xi/\sigma_1$ and $\tilde{c} = c/\sigma_1$, then (7.1) becomes

$$\begin{cases} -\tilde{c}\phi'(x) = d(L\phi)(x) + g\phi''(x) - f(\phi(x)) - \psi(x), \\ -\tilde{c}\psi'(x) = b(\phi(x) - r\psi(x)), \end{cases} \tag{7.2}$$

where $g = \gamma\frac{\sigma_2^2}{\sigma_1^2}$ for $\sigma_1 \neq 0$ and $(L\phi)(x) = \phi(x+1) - 2\phi(x) + \phi(x-1)$.

In Figure 7.10 we set $(d, \gamma, b, r) = (1, 1, 1, 0)$ and $(d, \gamma, b, r) = (1, 0, 1, 0)$, set $c = 1$ and $c = 2$, let $\sigma_1 = \cos(\theta)$ and $\sigma_2 = \sin(\theta)$ for $0 \leq \theta \leq 2\pi$. The plot in Figure 7.10(i)

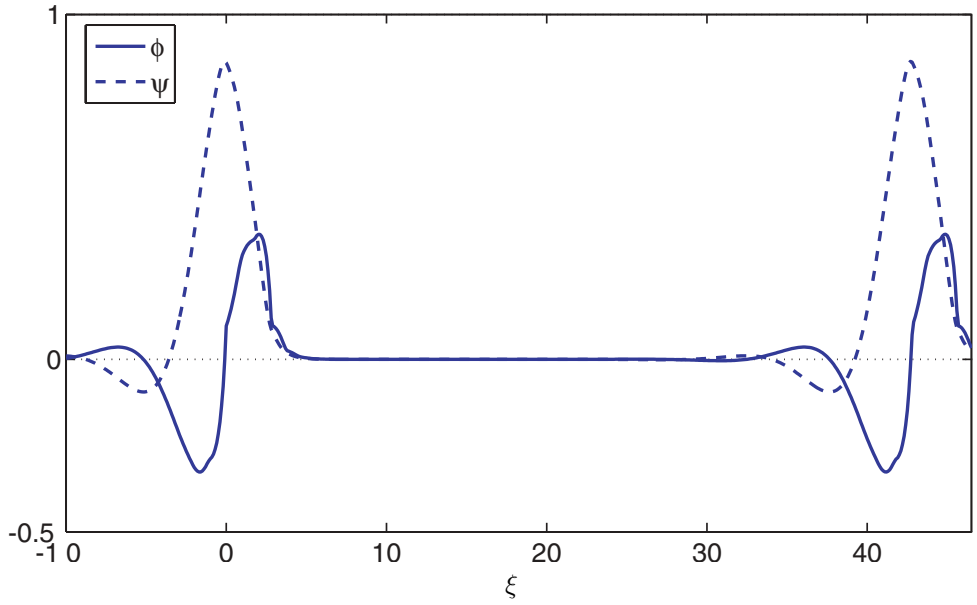


FIG. 7.8. *Plot of two-pulse waveforms for* $(d, c, \gamma, b, r) = (1, 1, 0, 1, 0)$ *obtained as a perturbation of a superposition of two identical one-pulse solutions.*

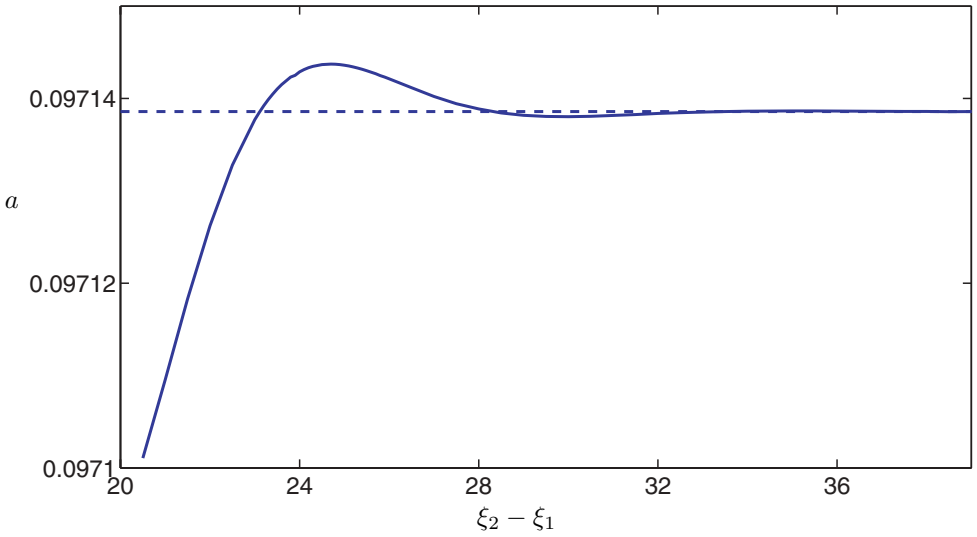


FIG. 7.9. *Plot of computed* a *obtained for* $(d, c, \gamma, b, r) = (1, 1, 0, 1, 0)$ *as a function the distance between the two pulses,* $\xi_2 - \xi_1$*, and compared with the horizontal line, the computed* a *value for the one-pulse solution.*

is a polar plot of a versus θ for the mixed continuous/discrete model and what we observed is that smaller values of a are obtained for $\sigma_1 \approx 1$ and compared with $\sigma_1 \approx 0$. This may be compared with the lack of anisotropy for the continuous model in Figure 7.10(ii) and the four-fold symmetry for the discrete model in Figure 7.10(iii).

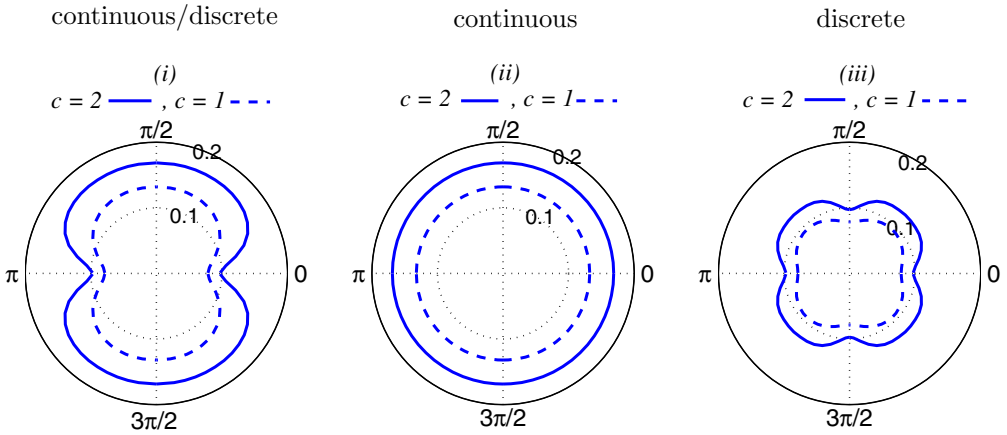


FIG. 7.10. *Polar plot of* $(\theta, a(\theta))$ *for* $(b, r) = (1, 0)$*; see* (7.2). *In plot* (i) $(d, g) = (1, 1)$ *with* $c = 2$ *and* $c = 1$ *for the mixed continuous/discrete version of the equation, in plot* (ii) $(d, g) = (0, 1)$ *with* $c = 2$ *and* $c = 1$ *for the continuous version of the equation, and in plot* (iii) $(d, g) = (1, 0)$ *with* $c = 2$ *and* $c = 1$ *for the discrete version of the equation.*

8. Conclusion. For a particular version of the FitzHugh–Nagumo equation (one which includes both spatially discrete and spatially continuous diffusive operators) with the McKean nonlinearity describing excitability, we have demonstrated how to construct candidate traveling wave solutions and have given tools for verifying if they are indeed solutions. This equation is meant to include models of action potential propagation on several length scales, from the internodal scale to the scale where a pulse becomes a spike. We begin by discussing the infinite number of eigenvalues obtained from linearizing. Leading edge behavior is governed by real eigenvalues, pulse trailing tails by complex ones (with nonzero real parts). Since solutions approach fixed points exponentially as we approach either plus or minus infinity, we can and do use the Fourier transform to derive candidate solutions. The one-front solutions (no recovery) we find have appeared in the Nagumo equation literature and their existence verification is relatively straightforward. The existence of pulse solutions, however, is more difficult. We have related existence to the pulse's relation to the "unstable root" of the reaction term, i.e., $\phi = a$, pointwise. At any point along the solution the solution is either above, below, or crossing a. The series of lemmas and theorems presented supply conditions, based on the derived candidate solutions, for verifying that these conditions are true, i.e., that a one-pulse solution crosses a exactly twice. Among the items that our numerical investigations of the solution behavior illustrate are

- that for single fronts more than one solution can exist;
- that for single pulses there is a range of a values such that there exists at least two distinct pairs of solution pulses;
- that the speed of front solutions can be seen as a bound for the speed of pulse solutions;
- that the distance between multiple pulse shows a dependence on the parameter a;
- that the spatially discrete diffusion operator retards propagation when compared to the spatially continuous diffusion operator.

Acknowledgments. The authors are grateful to Chris Lee and Paul Raff, who helped initiate this investigation as part of the REU site program DMS-9912293 at Colorado School of Mines.

REFERENCES

[1] A. R. A. Anderson and B. D. Sleeman, *Wave front propagation and its failure in coupled systems of discrete bistable cells modelled by FitzHugh-Nagumo dynamics*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 5 (1995), pp. 63–74.

[2] C. G. Aronson and H. F. Weinberger, *Multidimensional nonlinear diffusion arising from population genetics*, Adv. Math., 30 (1978), pp. 33–76.

[3] P. W. Bates, X. Chen, and A. J. J. Chmaj, *Traveling waves of bistable dynamics on a lattice*, SIAM J. Math. Anal., 35 (2003), pp. 520–546.

[4] J. Bell and C. Cosner, *Threshold behavior and propagation for nonlinear differential-difference systems motivated by modeling myelinated axons*, Quart. Appl. Math., 42 (1984), pp. 1–114.

[5] S. N. Chow, J. Mallet-Paret, and W. Shen, *Traveling waves in lattice dynamical systems*, J. Differential Equations, 149 (1998), pp. 248–291.

[6] J. W. Cahn, *Theory of crystal growth and interface motion in crystalline materials*, Acta Met., 6 (1960), pp. 554–561.

[7] J. W. Cahn, J. Mallet-Paret, and E. S. Van Vleck, *Traveling wave solutions for systems of ODEs on a two-dimensional spatial lattice*, SIAM J. Appl. Math., 59 (1998), pp. 455–493.

[8] A. Carpio and L. L. Bonilla, *Pulse propagation in discrete systems of coupled excitable cells*, SIAM J. Appl. Math., 63 (2002), pp. 619–635.

[9] S. Binczak, J. C. Eilbeck, and A. C. Scott, *Ephaptic coupling of myelinated nerve fibers*, Phys. D, 148 (2001), pp. 159–174.

[10] B. Deng, *The existence of infinitely many traveling front and back waves in the FitzHugh–Nagumo equations*, SIAM J. Math. Anal., 22 (1991), pp. 1631–1650.

[11] C. E. Elmer, *Multiple Planar Interfaces for Crystalline Materials with Spatially Discrete Gradient Energy*, preprint, 2005.

[12] C. E. Elmer and E. S. Van Vleck, *Analysis and computation of traveling wave solutions of bistable differential-difference equations*, Nonlinearity, 12 (1999), pp. 771–798.

[13] C. E. Elmer and E. S. Van Vleck, *Traveling wave solutions for bistable differential-difference equations with periodic diffusion*, SIAM J. Appl. Math., 61 (2001), pp. 1648–1679.

[14] T. Erneux and G. Nicolis, *Propagating waves in discrete bistable reaction-diffusion systems*, Phys. D, 67 (1993), pp. 237–244.

[15] J. Evans, *Nerve axon equations* I: *Linear approximations*, Indiana Univ. Math. J., 21 (1972), pp. 877–955.

[16] J. Evans, *Nerve axon equations* II: *Stability at rest*, Indiana Univ. Math. J., 22 (1972), pp. 75–90.

[17] J. Evans, *Nerve axon equations* III: *Stability of the nerve impulse*, Indiana Univ. Math. J., 22 (1972), pp. 577–594.

[18] J. Evans, *Nerve axon equations* IV: *The stable and unstable impulse*, Indiana Univ. Math. J., 24 (1975), pp. 1169–1190.

[19] G. Fath, *Propagation failure of traveling waves in discrete bistable medium*, Phys. D, 116 (1998), pp. 176–190.

[20] J. A. Feroe, *Existence and stability of multiple impulse solutions of a nerve equation*, SIAM J. Appl. Math., 42 (1982), pp. 235–246.

[21] P. C. Fife and J. B. McLeod, *The approach of solutions of nonlinear diffusion equations to travelling front solutions*, Arch. Ration. Mech. Anal., 65 (1977), pp. 335–361.

[22] R. FitzHugh, *Impulses and physiological states in theoretical models of nerve membrane*, Biophys. J., 1 (1961), pp. 445–466.

[23] W.-Z. Gao, *Threshold behavior and propagation for a differential-difference system*, SIAM J. Math. Anal., 24 (1993), pp. 89–115.

[24] S. Heinze, G. Papanicolaou, and A. Stevens, *Variational principles for propagation speeds in inhomogeneous media*, SIAM J. Appl. Math., 62 (2001), pp. 129–148.

[25] C. K. R. T. Jones, *Stability of the traveling wave solution of the FitzHugh-Nagumo system*, Trans. Amer. Math. Soc., 286 (1984), pp. 431–469.

[26] J. P. Keener, *Propagation and its failure in coupled systems of discrete excitable cells*, SIAM J. Appl. Math., 47 (1987), pp. 556–572.

[27] J. P. Keener, *The effects of discrete gap junction coupling on propagation in myocardium*, J. Theoret. Biol., 148 (1991), pp. 49–82.

[28] J. Keener and J. Sneyd, *Mathematical Physiology*, Springer-Verlag, New York, 1998.
[29] M. Krupa, B. Sandstede, and P. Szmolyan, *Fast and slow waves in the FitzHugh-Nagumo equation*, J. Differential Equations, 133 (1997), pp. 49–97.
[30] K. Maginu, *Geometrical characteristics associated with stability and bifurcations of periodic travelling waves in reaction-diffusion systems*, SIAM J. Appl. Math., 45 (1985), pp. 750–774.
[31] J. Mallet-Paret, *The Fredholm alternative for functional differential equations of mixed type*, J. Dynam. Differential Equations, 11 (1999), pp. 1–48.
[32] J. Mallet-Paret, *The global structure of traveling waves in spatially discrete dynamical systems*, J. Dynam. Differential Equations, 11 (1999), pp. 49–128.
[33] H. McKean, *Nagumo's equation*, Adv. Math., 4 (1970), pp. 209–223.
[34] J. Nagumo, S. Arimoto, and S. Yoshizawa, *An active pulse transmission line simulating nerve axon*, Proc. Inst. Radio Engrg., 50 (1964), pp. 2061–2070.
[35] J. Rinzel and J. B. Keller, *Traveling wave solutions of a nerve conduction equation*, Biophys. J., 13 (1973), pp. 1313–1337.
[36] A. Scott, *Neuroscience, A Mathematical Primer*, Springer-Verlag, New York, 2002.
[37] L. F. Shampine, R. C. Allen, and S. Pruess, *Fundamentals of Numerical Computing*, Wiley, New York, 1997.
[38] A. Tonnelier, *Wave propagation in discrete media*, J. Math. Biol., (2001), pp. 1–19.
[39] A. Tonnelier, *The McKean's caricature of the FitzHugh–Nagumo model* I. *The space-clamped system*, SIAM J. Appl. Math., 63 (2002), pp. 459–484.
[40] A. Tonnelier, *The McKean's caricature of the FitzHugh-Nagumo model: Traveling pulses in discrete diffusive medium*, Phys. Rev. E (3), 67 (2003), 036105.
[41] W.-P. Wang, *Multiple impulse solutions to McKean's caricature of the nerve equation.* I. *Existence*, Comm. Pure Appl. Math., 41 (1988), pp. 71–103.
[42] W.-P. Wang, *Multiple impulse solutions to McKean's caricature of the nerve equation.* II. *Stability*, Comm. Pure Appl. Math., 41 (1988), pp. 997–1025.
[43] E. Yanagida, *Stability of fast travelling pulse solutions of the FitzHugh-Nagumo equation*, J. Math. Biol., 22 (1985), pp. 81–104.
[44] B. Zinner, *Stability of traveling wavefronts for the discrete Nagumo equation*, SIAM J. Math. Anal., 22 (1991), pp. 1016–1020.
[45] B. Zinner, *Existence of traveling wavefront solutions for the discrete Nagumo equation*, J. Differential Equations, 96 (1992), pp. 1–27.

SIAM J. APPL. MATH.
Vol. 65, No. 4, pp. 1175–1199

COMBUSTION STABILIZATION BY FORCED OSCILLATIONS IN A DUCT*

ABRAM DORFMAN†

Abstract. The feasibility of stabilizing premixed combustion by forced oscillation is analytically demonstrated through the simulation of an active control input-output mechanism. The developed model is used for analysis of the interactions between an autonomous oscillation in a duct, a loudspeaker's input, and the unsteady heat release. We assume that the autonomous oscillations (at frequency ω_0) exist in a duct containing a flame with a loudspeaker at the input. At $t = 0$, the loudspeaker starts to generate oscillations at a different frequency ω. To find the resulting oscillations (the output), a mathematical technique is needed that takes into account (1) the pressure and velocity fields in the duct when the loudspeaker starts; (2) the variable amplitudes of the resulting oscillations, which depend on time and location; and (3) coupling of the fresh and burnt gas flows at the flame. Such a technique differs significantly from that used by previous authors for studying single oscillation/flame interactions. The mathematical development leads to an exact solution that gives a stability criterion in the form of a system of two integro-differential equations. Analysis shows that the stability domains of the time lag depend mainly on the flame location and the fresh/burnt gases temperature ratio. Numerical results are obtained for a centrally located flame and for the temperature ratio 1500 K/300 K.

Key words. combustion stability, perturbation, Laplace transform, integro-differential equations

AMS subject classifications. 80A25, 93C30, 93C73, 44A10, 45J05

DOI. 10.1137/S0036139902415579

1. Introduction.

Since the 1950s, when combustion instabilities caused numerous failures during the development of rocket motors, considerable effort has been spent on the investigation of this phenomenon. Since that time, a number of mathematical models have been developed to identify the mechanism of oscillations and to find ways to reduce their magnitude or suppress them. Crocco and Cheng [1] first studied this problem by using a simple (n-τ) model, where n is an interaction index and τ is a time lag.

The idea behind the time lag τ is that there is always an interval between the time of injection of the propellant and the time at which burning occurs. The interaction index, n, describes the intensity of the coupling between unsteady heat release and velocity fluctuations in the reaction zone. According to the Rayleigh criterion, this coupling process is responsible for combustion instabilities. The interaction index is usually taken as a constant of proportionality between the heat release and the velocity oscillation. The governing equations for this model are obtained from simplified forms of the conservation laws. As a version of the (n-τ) model will be used in this study, more details are given in subsection 3.3.

Crocco and Cheng also provide reviews of other early works [1, 2, 3]. Overviews of later results [4, 5, 6, 7], contemporary reviews [8, 9, 10, 11, 12, 13, 14, 15, 16], surveys of advanced numerical methods [17, 18], and contemporary models of turbulent flame dynamics [17, 19] are also available.

*Received by the editors October 4, 2002; accepted for publication (in revised form) June 2, 2004; published electronically April 14, 2005.

http://www.siam.org/journals/siap/65-4/41557.html

†Visiting Professor, Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122 (abram_dorfman@hotmail.com).

The time lag model was intensively used for both developing physical understanding and correlating experimental data by Crocco and associates in the 1960s [1, 2, 3, 4, 6]. Due to its simplicity, this approach is still of interest [9, 11, 20, 23, 24]. In particular, as described in a recent article [20], this model can be used for detailed analysis of the basic processes responsible for instabilities in premixed combustion. However, we note that this model is restricted to linear processes.

Later, Culick [10, 22] used the Galerkin method and presented the solution of the relevant wave equation as a series of normal modes. This model describes the linear and nonlinear behavior of the oscillations, but the heat release is still considered as a linear process. Culick's model was widely used for solving some basic problems of combustion instabilities [10, 21, 22, 25, 26, 27, 28, 29], and recently it was applied to model a closed-loop active control system with feedback. Two types of controller were considered, with fixed parameters [28] and with variable ones, self-tuned to the operating conditions [29]. The limitations of this model arise from the expansion in normal modes, which is based on the assumption that the studied problems differ only by a small amount from the unperturbed case. A second limitation of this approach results from the necessary truncation of the series expansion.

The kinematic laminar combustion model [30] is constructed using fundamental mechanistic principles. Poiseuille flow in a tube with a thin flame that moves at constant velocity was considered. Analysis of the moving flame front leads to differential equations for the flame surface area and for the unsteady heat release. These results allow for the derivation of a model describing the oscillation/flame interaction in (n-τ) terms. Using this model together with Culick's technique, different applications, including feedback control, have been addressed [31, 32, 33, 34, 35].

A second laminar model was developed by Merk [36]. To study combustion instabilities, he employed transfer functions and an analogy to electrical circuits. The developed theory was used to determine the stability domains for a duct with a thin flame. The stability analysis showed that the value of the time lag is very important because instability was found to occur in a certain interval of the time lag.

The basic ideas of this method have been applied in recent two papers [37, 38]. In the first one, in particular, a review of using transfer function to study premixed flames is given. In the second work, authors employed this method to predict and control instabilities in systems with inclined laminar flames. The studied system is described by a network of transfer functions that simulate acoustic elements in a combustor. To derive analytical expressions for the transfer functions, the transport G-equation is used. Conical and V-flames are considered. It is shown that in both cases, models containing one dimensionless parameter are a good approximation only for low frequencies. The limitations of such a model (considered, for example, in [30]) are discussed in [37, 38]. At higher frequencies, the flame dynamics are controlled by two independent parameters. Additional improvements could be achieved by taking into account convective effects.

Although laminar models are not applicable to devices with turbulent motion, they are important for developing a better understanding of the associated physics. They may also be applied to more complex models by fitting parameters to experimental data [39].

In contrast to Culick, Peracchio and Proscia [39] proposed a model that describes the coupling of a linear oscillation with turbulent nonlinear heat release. They adapted laminar model relations from [30]. By fitting model parameters to experimental data, an expression for the heat release was obtained. The dependence of the heat release

on the fuel/air ratio was also taken into account. The linear acoustics were described by Culick's equation, and the stability analysis was performed using the Laplace transform. On the basis on this model, the technique of identification of limit cycling systems has been developed [40].

Lang et al. [23] proposed a model to study the feedback active control. They used the (n-τ) approach and the solution of the wave equation for a duct with a thin flame. Autonomous oscillations with constant amplitudes were considered. Satisfying the boundary conditions leads to an expression for the stability criterion. Using this criterion and the transfer function, the authors analyzed stability with and without control. The results were compared with the authors' experimental data. Gulati and Mani [24] studied the effect of the equivalence ratio and the flow rate on the control performance using the same approach.

In these two works, specific cases were considered for comparison with experimental data. McManus, Poinsot, and Candel [9] applied the same model for a general analytical investigation of feedback active control in a duct. Ignoring the temperature jump across the flame, the authors analyze stability with and without control. As a result, the stability domains of frequency can be computed for the two first harmonic modes.

Two models with nonlinear heat release effects were proposed by Dowling. In the first [41], the idea of saturated heat release is used. According to that idea, a nonlinearity occurs when the linearly varying heat release becomes constant (i.e., saturates). In the second work [42], the Fliefil et al. [30] approach is extended to turbulent combustion. Unlike the authors of study [39], who fitted the laminar model parameters to the experimental data, Dowling applied the basic idea to a turbulent flame. By using the same assumption of neglecting expansion across the flame front, the relations are derived for the flame surface area and for the unsteady heat release. In both studies, the acoustics are treated as linear and are modeled by the well-known general solution of the wave equation. The solutions for the fresh and burnt gases are conjugated at the flame. To perform this conjugation, the conservation equations, written for the flame zone, and experimental data from [43] are employed. Applications have been subsequently investigated [44, 45, 46].

The brief review presented here considered contemporary analytical methods of modeling combustion instabilities and active control simulation. This part of the extensive literature on combustion instabilities has been reviewed since these studies are closely related to the subject of the current investigation. Experimental and numerical results may be found in the reviews mentioned above.

The literature review has shown that different analytical approaches can be used to study the interaction of oscillations and unsteady heat release, including the mechanism of this phenomenon. At the same time, the analytical models of active control are based on the input-output approach, without considering the processes inside the combustor.

In this article, we develop a model for simulating the active control input-output mechanism. The model describes the interaction of an existing oscillation in the combustor with a control input and with unsteady heat release. This case of the interaction of two oscillations in the combustor, one already existing and the other incoming, with unsteady heat release, is quite different from the well-known single oscillation/heat release coupling problem.

The proposed model for studying this type of interaction is based on the solution of the wave equation which takes into account the following facts: (1) There is a flow

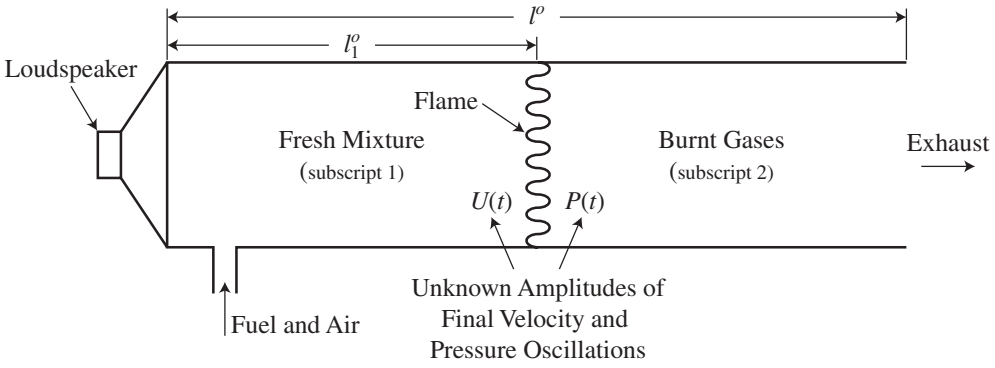


FIG. 1. *Schematic diagram of the modeled combustor as a duct.*

in the combustor when the control input enters. Therefore, the corresponding velocity and pressure fields should be used as initial conditions. (2) The result of interaction of the two different oscillations is complex, with unknown amplitude depending on time and location. (3) The resulting oscillation should be found by the conjugation of two wave equation solutions obtained separately for each part of combustor (i.e., the fresh and burnt gases). (4) While the wave equation is second order, there is only one boundary condition in each part of the combustor divided by the flame. Hence, two other conditions are needed.

For such additional conditions, we use two unknown variables defining, at the flame, the amplitudes of the velocity on the upstream side, and of the pressure on the downstream side. The corresponding mathematical technique leads to a system of two integro-differential equations that determine these unknown functions. A more detailed description of the model and mathematical technique is given in sections 3 and 4.

This system of two integro-differential equations can also be used to develop a stability criterion. Using the stability analysis, the feasibility of stabilizing the premixed combustion by forced oscillation is demonstrated in sections 5 and 6. The ranges of the time lag providing stability are derived and are determined to depend mainly on the flame location and burnt/fresh gases temperature ratio (section 5).

In this study, a combustor in the form of a duct with simple boundary conditions is considered and the (n-τ) approach is used to compute the unsteady heat release. In this case, the exact solution is obtained. The presented model of two oscillations/heat release interaction and mathematical technique may be applied to more complex combustor forms, boundary conditions, and unsteady heat release theories.

2. Problem formulation. The flame is located at $x = l_1$ in a duct of length l (Figure 1) with the loudspeaker at the input ($x = 0$). Assume that in such a duct, when the loudspeaker is off, autonomous oscillations at frequency ω_0 exist. At time $t = 0$, the loudspeaker starts to generate oscillations at frequency $\omega \neq \omega_0$.

Our goals are (1) a stability analysis of the autonomous oscillations when the loudspeaker is off; (2) computing the oscillation that results from the interaction of those existing in the duct autonomously and those from the loudspeaker; and (3) a stability analysis of the final flow and defining the combustion and forced oscillation characteristics that stabilize or suppress initially unstable oscillations.

Although this problem looks close to that of McManus, Poinsot, and Candel [9], they differ in essence. In [9] the active control is considered using the transfer function and the same simple waves with constant amplitudes as in the case without control. In this study, we simulate the mechanism of input–output control considering the interaction of two oscillations at different frequencies with unsteady heat release. The result of this interaction is a complex oscillation with variable amplitudes. We obtain this resulting oscillation by solving the wave equation with initial, boundary, and conjugation conditions.

3. Model. According to our goals, the model is intended to solve the following parts of the whole problem: (1) determining the pressure and velocity fields of the autonomous regime and a stability analysis when the loudspeaker is off; (2) determining the pressure and velocity fields in the two parts of the duct when the loudspeaker is on; (3) obtaining the resulting oscillation by conjugating the flows in the two parts of the duct at the flame; and (4) a stability analysis when the loudspeaker is on.

3.1. Basic assumptions. The proposed model is based on the same assumptions as the Lang, Poinsot, and Candel [23] model and others of this type. They are as follows [9, 23, 24]: (1) Flow in the combustor is a one-dimensional current of inviscid, non-heat-conducting gas with constant properties. (2) The effect of the mean flow and the mean heat transfer on the acoustic waves may be neglected, as well as the effect of pressure drop and heat losses. (3) The acoustic oscillations are plane longitudinal waves. (4) The flame is a thin sheet dividing the duct into two parts with fresh and burnt gases (subscripts 1 and 2, Figure 1).

3.2. Governing equations. According to assumption 1, the flow in the combustor is governed by one-dimensional conservation equations of mass, momentum, and energy:

$$\frac{\partial \rho}{\partial t}+\frac{\partial(\rho u)}{\partial x}=0, \quad \frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+\frac{1}{\rho}\frac{\partial p}{\partial x}=0, \quad \frac{\partial p}{\partial t}+u\frac{\partial p}{\partial x}+p\kappa\frac{\partial u}{\partial x}=(\kappa-1)\quad q_V, \tag{3.1}$$

where ρ, p, u, x, t, q_v, and κ are density, pressure, velocity, distance from loudspeaker, time, heat generation per unit volume, and the ratio of specific heats. The parameters in (3.1) are expanded into mean ($\bar{p}$) and fluctuating (p') components. Assuming that the mean parts are uniform, neglecting density fluctuations, and using $c^2=\kappa\bar{p}/\bar{\rho}$ for the speed of sound, one gets linearized momentum and energy equations,

$$\bar{\rho}\frac{\partial u'}{\partial t}+\frac{\partial p'}{\partial x}=0, \qquad \frac{\partial p'}{\partial t}+\rho c^2\frac{\partial u'}{\partial x}=(\kappa-1)\,q'_V. \tag{3.2}$$

Outside of the flame zone, $q'_V=0$. In such a case, eliminating one of the variables from (3.2) by differentiating leads to the two wave equations

$$\frac{\partial^2 u}{\partial t^2}-c^2\frac{\partial^2 u}{\partial x^2}=0, \qquad \frac{\partial^2 p}{\partial t^2}-c^2\frac{\partial^2 p}{\partial x^2}=0, \tag{3.3}$$

where the sign ($'$) is omitted because, in conformity with assumption 2, only fluctuating components need to be considered.

Using the mean density, $\bar{\rho}$, the length of the duct, l, and the speed of sound, c, as scales, we introduce nondimensional variables

$$x=\frac{x^\circ}{l}, \quad t=\frac{t^\circ c}{l}, \quad u=\frac{u^\circ}{c}, \quad p=\frac{p^\circ}{\bar{\rho}c^2}, \quad q_V=\frac{q_V^\circ l}{\bar{\rho}c^3}, \quad k=\frac{\omega^\circ l}{c}. \tag{3.4}$$

Here ω and k are frequency and wave number, and the superscript $^\circ$ is applied to distinguish dimensional from nondimensional variables. In the following, we use nondimensional quantities, except scales, and in some cases where dimensional variables are marked with sign ($^\circ$).

Since the duct is divided in two parts with fresh and burnt gases, there are two values of the speed of sound and corresponding wave numbers. However, if we neglect the slight temperature dependence of the specific heat ratio, we obtain the relationship $c_2/c_1 = (T_2^\circ/T_1^\circ)^{1/2} = \zeta$, where T_1° and T_2° are the temperatures of the fresh and burnt gases. Then, setting for the burnt gases $k_2 = k$, one gets $k_1 = k\zeta$ for the fresh gases. Since we use the speed of sound c_2 to obtain the dimensionless wave number, the nondimensional time is also determined by this speed of sound. With $c_1/c_2 = (T_1^\circ/T_2^\circ)^{1/2}$, the pressure scales for the fresh and burnt gases are equal ($c_1^2\bar\rho_1 = c_2^2\bar\rho_2$), because $\bar\rho_1/\bar\rho_2 = T_2^\circ/T_1^\circ$.

3.3. Boundary and conjugate conditions. Two boundary and two conjugate conditions are applied. One boundary condition defines the source of the oscillations. We suppose that the autonomous oscillations in the duct are caused by flame fluctuations, while the forced oscillations are produced by the loudspeaker. Assuming that both oscillations are harmonic, we have in the autonomous and forced cases, respectively,

$$u_1(l_1,t) = \alpha_0 \exp(-ik_0 t), \qquad u_1(0,t) = \alpha \exp(-ikt), \tag{3.5}$$

where $l_1 = l_1^\circ/l$ is the flame location, α is amplitude, and subscript 0 denotes the autonomous regime.

The second condition is as in [9, 23]: zero pressure fluctuation at the exhaust

$$p_2(1,t) = 0, \tag{3.6}$$

where the unit 1 indicates the nondimensional length of the duct.

Two conjugation conditions at the flame are adapted from [9, 23]. They are derived on the basis of the (n-τ) model and have the form

$$p_1(l_1,t) = p_2(l_1,t), \tag{3.7}$$

$$\zeta u_2(l_1,t) - u_1(l_1,t) = n u_1(l_1, t-\tau). \tag{3.8}$$

The first expression specifies the continuity of pressure across the flame. The second one defines the velocity jump caused by the high temperature of the burnt gases.

Relation (3.8) is obtained from the energy equation in (3.1) [9, 23]. The derivation starts by letting the flame zone width be Δx°. Then, the average value of the velocity derivative in the flame is $(u_2^\circ - u_1^\circ)/\Delta x^\circ$. The first two terms in the energy equation in (3.1) define the material pressure derivative. Since the pressure does not change across the flame, these two terms become zero in the flame zone. Omitting these terms, using the formula for the speed of sound $c^2 = \kappa p^\circ/\bar\rho$, and then introducing the dimensionless variables defined in (3.4), one gets the average velocity jump in the form $\zeta u_2 - u_1 = (\kappa - 1) q_V \Delta x$. Following Crooco and Cheng [1], the unsteady heat release is expressed using the interaction index, n, and the velocity fluctuation at the flame front, but with regard to time lag: $(\kappa - 1) q_V \Delta x = n u_1(l_1, t-\tau)$. These last two relations yield (3.8).

When the loudspeaker is off, the four boundary conditions (3.5)–(3.8) determine the pressure and velocity amplitudes of autonomous oscillations in the duct. However,

when the loudspeaker is on (i.e., when the interaction of two oscillations with different frequencies occurs), these four conditions are insufficient.

It is known that the result of such interactions is a complex oscillation with time and location dependent velocity and pressure amplitudes. As was mentioned above, we determine this oscillation by solving the wave equation separately for each part of the duct. To have two boundary conditions in each part, we introduce, in addition to (3.5)–(3.8), two unknown functions, $U(t)$ and $P(t)$. They define the velocity and pressure amplitudes on the upstream and downstream sides of the flame, respectively (see Figure 1), such that

$$|u_1(l_1,t)|_{\max} = \alpha U(t), \qquad |p_2(l_1,t)|_{\max} = \alpha P(t). \tag{3.9}$$

Conjugation of the solutions obtained for the two parts of the duct at the flame leads to a system that determines the functions $U(t)$ and $P(t)$.

3.4. Initial conditions. When the loudspeaker starts, autonomous oscillations with frequency ω_0 exist in the combustor. Hence, the pressure and velocity related to these oscillations at $t = 0$ are the initial conditions for solving the wave equation. They are found using conditions (3.5)–(3.8) in the next section.

4. Mathematical development. This section contains three parts. In the first part, the autonomous oscillations that exist when the loudspeaker is off are considered. The two goals of this part are (1) to find the pressure and velocity distributions in the duct when the loudspeaker starts—these are used as the initial conditions in solving the wave equation for the resulting oscillations; and (2) stability analysis of the autonomous regime—this information is needed to answer the main question: Are there conditions under which loudspeaker oscillations stabilize or suppress initially unstable autonomous oscillations?

The second part of this section includes solutions of the wave equations for the two parts of the duct, which give the pressure and velocity fields before and after the flame, in the fresh and burnt gases. The technique of conjugating these fields at the flame is presented in the third part of this section. The conjugation gives the final oscillation, i.e., the output of the interaction between the autonomous and forced oscillations. Subsequently, in section 5, this result is used in a stability analysis to determine if the final oscillations are stable.

4.1. Loudspeaker off. This problem is similar to that in the McManus, Poinsot, and Candel study [9], but in contrast to their solution, we take into account the temperature jump across the flame. We also assume that the flame fluctuations cause a harmonic velocity oscillation, given by (3.5), at the flame instead of imposing the condition $u_1(0,t) = 0$ at the input.

In (3.5), α_0 is an arbitrary complex amplitude. In study [9], the velocity oscillation at the flame front is also harmonic, but with a specific amplitude, which corresponds to the case of zero velocity at the inlet. At the same time, a specific velocity oscillation at the inlet corresponds to (3.5) at the flame in this study. As will be shown in subsection 4.1.1, such corresponding amplitudes at the flame in the first case and the velocity at the inlet in the second one are

$$\alpha_0 = 1-\exp(-2ik_0l_1\zeta), \qquad u(0,t) = \exp[-ik_0(l_1\zeta+t)]-(1-\alpha_0)\exp[ik_0(l_1\zeta-t)]. \tag{4.1}$$

Thus, the boundary conditions in both cases are similar. However, in the case of a closed inlet, it is natural to put zero velocity at $x = 0$, as well as the simple condition (3.5) at the flame instead of condition (4.1) at the inlet in the case with arbitrary amplitude α_0.

4.1.1. Pressure and velocity distributions. The oscillations in the duct are considered as waves with constant amplitudes A_i and B_i, similar to those in [9, 23]:

$$\begin{aligned} &A_1\exp\{ik_0[\zeta(x-l_1)-t]\}\pm B_1\exp\{-ik_0[\zeta(x-l_1)+t]\},\\ &A_2\exp\{ik_0[\zeta(x-l_1)-t]\}\pm B_2\exp\{-ik_0[\zeta(x-l_1)+t]\}. \end{aligned} \tag{4.2}$$

Here sums define the pressure, while differences give the velocity, $k_0=\omega_0^\circ l/c_2$, and $t=t^\circ c_2/l$.

Substituting (4.2) in the boundary and conjugate conditions (3.5)–(3.8) yields

$$\begin{aligned} &A_1-B_1=\alpha_0, \qquad A_2\exp[ik_0(1-l_1)]+B_2\exp[-ik_0(1-l_1)]=0,\\ &A_1+B_1=A_2+B_2, \qquad \zeta(A_2-B_2)-(A_1-B_1)[1+n_0\exp(ik_0\tau_0)]=0, \end{aligned} \tag{4.3}$$

where $1-l_1$ is the duct length behind the flame. The parameters (τ_0,n_0) are different from (τ,n) for the final oscillations since it is known that in general they depend on the oscillation frequency and combustion characteristics.

Because α_0 is arbitrary, it follows from (4.3) that one may set $A_1=1$. Then the first three equations in (4.3) determine the other amplitudes A_i and B_i. Using them together with (4.2), one obtains the desired pressure and velocity distributions:

$$\begin{aligned} p_{10}&=\exp\{ik_0[\zeta(x-l_1)-t]\}+(1-\alpha_0)\exp\{-ik_0[\zeta(x-l_1)+t]\},\\ u_{10}&=\exp\{ik_0[\zeta(x-l_1)-t]\}-(1-\alpha_0)\exp\{-ik_0[\zeta(x-l_1)+t]\},\\ p_{20}&=\frac{2-\alpha_0}{1-\exp[-2ik_0(1-l_1)]}\{\exp[-ik_0(x-l_1+t)]-\exp[ik_0(x+l_1-t-2)]\},\\ u_{20}&=\frac{2-\alpha_0}{1-\exp[-2ik_0(1-l_1)]}\{-\exp[-ik_0(x-l_1+t)]-\exp[ik_0(x+l_1-t-2)]\}. \end{aligned} \tag{4.4}$$

The expressions in (4.1) follow from the second equation of (4.4) if one puts $u_{10}(l_1,t)=0$ (to obtain the first expression) and $x=0$ (to obtain the second one).

4.1.2. Stability analysis. The last equation in (4.3) can be used to determine α_0. After substituting for the amplitudes A_i and B_i, this equation takes the form

$$\zeta(2-\alpha_0)\{1+\exp[-2ik_0(1-l_1)]\}+\alpha_0\{1-\exp[-2ik_0(1-l_1)]\}[1+n_0\exp(i\omega_0\tau_0)]=0. \tag{4.5}$$

This equation gives a generalized stability criterion for loudspeaker oscillations (3.5) with arbitrary amplitude α_0. The criterion obtained in [9, 23] with determinant may also be derived from the last equation of (4.3). This criterion follows also from (4.5) if one sets $\alpha_0=1-\exp(-2ik_0l_1\zeta)$, which corresponds to the case considered in [9].

Setting in (4.5) $\alpha_0=\alpha_{0R}+i\alpha_{0I}$ and $k_0=k_{0R}+ik_{0I}$ and separating the result into real and imaginary parts leads to

$$\begin{aligned} \alpha_{0R}&=\frac{2\zeta\{h_C\{1+\exp[2k_{0I}(1-l_1)\cos\phi]\}+h_S\sin2\phi\exp[2k_{0I}(1-l_1)]\}}{h_C^2+h_S^2},\\ \alpha_{0I}&=\frac{2\zeta\{h_S\{1+\exp[2k_{0I}(1-l_1)\cos\phi]\}-h_C\sin2\phi\exp[2k_{0I}(1-l_1)]\}}{h_C^2+h_S^2},\\ h_C&=f_C^+\exp[2k_{0I}(1-l_1)]+f_C^-, \qquad h_S=f_S^-\exp[2k_{0I}(1-l_1)]+f_S^+,\\ f_S^+&=n_0\sin\vartheta_0, \qquad f_S^-=(\zeta+1)\sin2\phi-n\sin(\vartheta_0-2\phi),\\ f_C^+&=(\zeta+1)\cos2\phi+n\cos(\vartheta_0-2\phi), \qquad f_C^-=(\zeta-1)-n_0\cos\vartheta_0. \end{aligned} \tag{4.6}$$

Here $\phi=k_{0R}(1-l_1)$, $\vartheta_0=2\pi(\tau_0/T_0)$, and T_0 is the autonomous oscillations period.

The neutral curves are obtained from (4.6) by setting $\exp[2k_{0I}(1-l_1)]=1$. To specify the stability domains, derivatives of k_{0I} with respect to α_{0R} and α_{0I} are needed:

$$
\begin{aligned}
2(1-l_1)\frac{\partial k_{0I}}{\partial \alpha_{0R}} &= \frac{(f_C^+ + f_C^-)^2 + (f_S^+ + f_S^-)^2}{g_C(f_C^+ + f_C^-) - g_S(f_S^+ + f_S^-)},\\
2(1-l_1)\frac{\partial k_{0I}}{\partial \alpha_{0I}} &= \frac{(f_C^+ + f_C^-)^2 + (f_S^+ + f_S^-)^2}{g_C(f_S^+ + f_S^-) + g_S(f_C^+ + f_C^-)},\\
g_C = 2\zeta\cos 2\phi - \alpha_{0R}f_C^+ - \alpha_{0I}f_S^-, &\qquad g_S = -2\zeta\sin 2\phi + \alpha_{0R}f_S^- - \alpha_{0I}f_C^+.
\end{aligned}
\tag{4.7}
$$

As an example, in Figure 2 the neutral curves $\alpha_0 = f(\tau_0/T_0)$ and the corresponding derivatives are given for the following data: $n_0 = 0.5, \phi = k_{0R}(1-l_1) = 0.4\pi, \zeta^2 = 5$. The derivative $\partial k_{0I}/\partial\alpha_{0I}$ is negative throughout the curve $\alpha_0 = f(\tau_0/T_0)$, while the other derivative, $\partial k_{0I}/\partial\alpha_{0R}$, is negative only in zone 3 (Figure 2(b)). Since on the neutral curves $k_{0I} = 0$, the sign of the derivative indicates the stability domain location. It lies above the neutral curve for negative derivatives and below the neutral curve for positive ones. Hence, in this case, the stability domains are located: outside of the curves in zones 1 and 4, inside of the curves in zone 2, and above both of the curves in zone 3. These regions are marked with hatching in Figure 2(a).

To determine the real and imaginary parts of α_0 for a given value of τ_0/T_0, one may choose from Figure 2(a) only α_{0R} or α_{0I}. The other should be found from the equality which is derived by solving (4.6) for $\exp[2k_{0I}(1-l_1)]$:

$$
\exp[2k_{0I}(1-l_1)] = \frac{-2\zeta + \alpha_{0R}f_C^- + \alpha_{0I}f_S^+}{2\zeta\cos 2\phi - \alpha_{0R}f_C^+ - \alpha_{0I}f_S^-} = \frac{\alpha_{0R}f_S^+ - \alpha_{0I}f_C^-}{2\zeta\sin 2\phi - \alpha_{0R}f_S^- + \alpha_{0I}f_C^+}.
\tag{4.8}
$$

For example, if for the unstable regime and $\tau_0/T_0 = 0.2$ one takes $\alpha_{0R} = 0.8$ from Figure 2(a), the equality (4.8) gives $\alpha_{0I} = 0.87$ and $\exp[2k_{0I}(1-l_1)] = 1.2$.

Thus, the analysis indicates that there are both stable and unstable autonomous oscillation regimes. We will show that unstable autonomous oscillations can be stabilized by the loudspeaker's input if the time lag and forced oscillation frequency correspond to specific domains. To show this, we start by determining the velocity and pressure fields that arise as an output of the interaction between the autonomous and loudspeaker oscillations.

4.2. Loudspeaker on. In this case the pressure and velocity fields are obtained by solving the wave equations (3.3) for each portion of the duct. The method of solution of such linear homogeneous equations with initial (4.4) and boundary ((3.5), (3.6), and (3.9)) conditions is well known [47]. After converting the inhomogeneous boundary condition (3.5) or (3.9) to a homogeneous one, the desired solution is presented as a sum of two others. One is a solution of the original equation (3.3) with a given initial, but homogeneous boundary condition. The other satisfies the zero initial and boundary conditions and an inhomogeneous equation that arises instead of the inhomogeneous boundary condition. In this case, such inhomogeneous wave equations contain the unknown functions $U(t)$ or $P(t)$ from the boundary conditions (3.9). Finally, this leads to a system of two integro-differential equations that determine these functions.

4.2.1. Velocity and pressure in the front of the flame. To find the velocity in the duct portion upstream of the flame, the first wave equation in (3.3) is solved. Let the required solution have the same form as the loudspeaker oscillation (3.5), such that

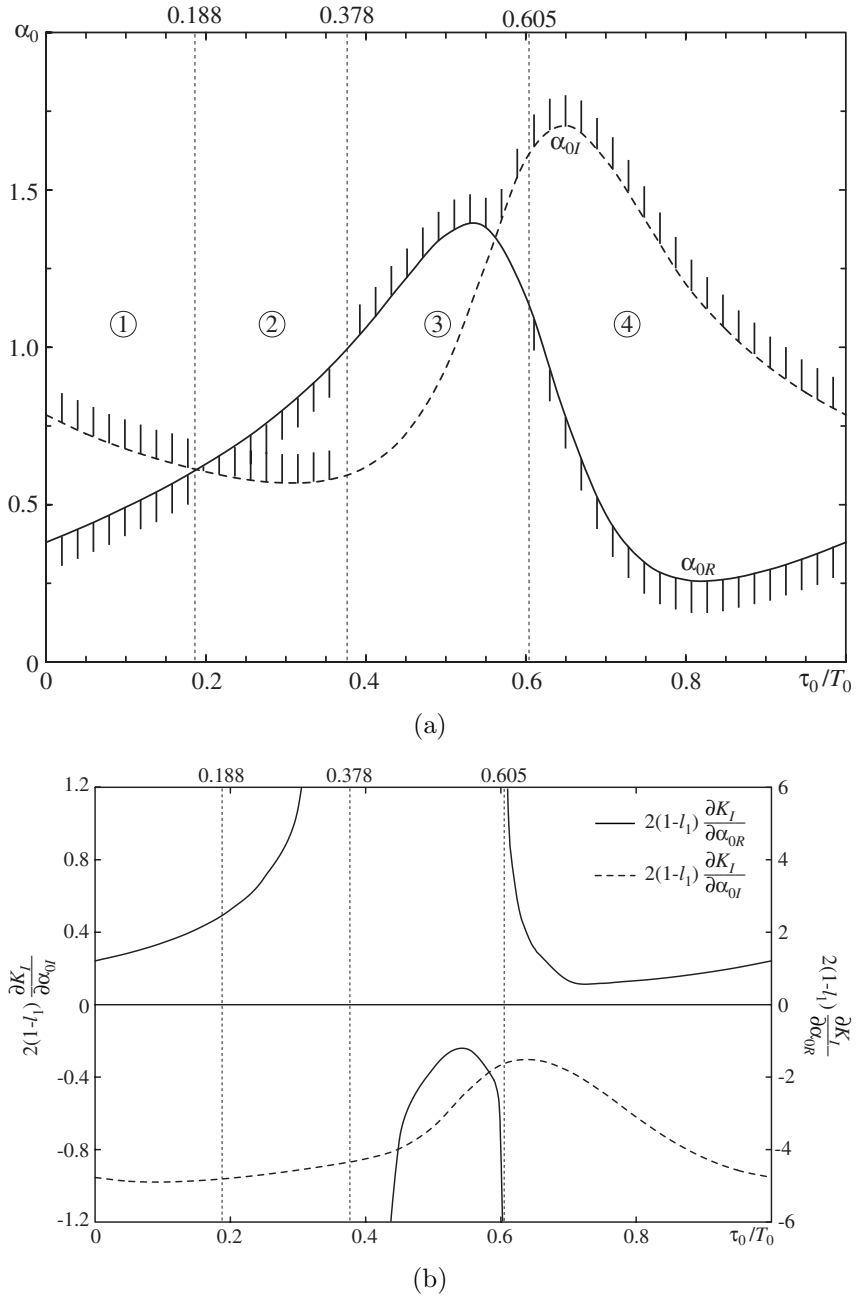


FIG. 2. *Neutral curves and stability domains for autonomous oscillations in a duct.* $n_0 = 0.5, \phi = 0.4\pi, T_2^0/T_1^0 = 5$. (a) *Neutral curves.* (b) *Derivatives of the neutral curves.* (τ_0/T_0): $0.188 =$ *the intersection of curves,* 0.378 *and* $0.605 =$ *changes of sign of the derivative* $\partial k_I/\partial\alpha_{0R}$.

$$u_1(x,t) = \alpha F(x,t)\exp(-ikt), \tag{4.9}$$

with the unknown function $F(x,t)$ defining the amplitude. The boundary condition for this function at the input follows from (3.5), while at the flame it is defined by (3.9):

$$F(0,t) = 1, \qquad F(l_1,t) = U(t). \tag{4.10}$$

Conditions (4.10) are transformed to homogeneous ones by the substitution

$$F(x,t) = \Phi(x,t) - (x/l_1)[1 - U(t)] + 1. \tag{4.11}$$

After using (4.10) and (4.11), the wave equation (3.3) takes an inhomogeneous form with the right-hand side containing the unknown function $U(t)$ and its first two derivatives. Replacing c_1 for c for fresh gases in (3.3), we present this equation in the form

$$\frac{\partial^2\Phi}{\partial t^2} - 2ik\frac{\partial\Phi}{\partial t} - k^2\Phi - \frac{1}{\zeta^2}\frac{\partial^2\Phi}{\partial x^2} = \frac{x}{l_1}Z(t) + k^2\left(1 - \frac{x}{l_1}\right), \tag{4.12}$$

$$Z(t) = -\frac{\partial^2 U}{\partial t^2} + 2ik\frac{\partial U}{\partial t} + k^2 U, \qquad \Phi(0,t) = \Phi(l_1,t) = 0. \tag{4.13}$$

Since $u_{10}(x,0)$, defined by (4.4), describes the velocity field in the combustor when the loudspeaker starts, the initial conditions are found by comparing (4.4), (4.9), and (4.11). Setting $t = 0$ in these expressions, one gets $\Phi(x,0)$ after solving the equation $u_{10}(x,0) = u_1(x,0)$. Similarly, the derivative is found after differentiating (4.4), (4.9), and (4.11):

$$\begin{aligned}\Phi(x,0) = {} & \frac{1}{\alpha}\{\exp[ik_0\zeta(x-l_1)] - \exp[-ik_0\zeta(x-l_1)]\} \\ & + \frac{\alpha_0}{\alpha}\exp[-ik_0\zeta(x-l_1)] + \frac{x}{l_1}\left(1 - \frac{\alpha}{\alpha_0}\right) - 1,\end{aligned} \tag{4.14}$$

$$\left(\frac{\partial\Phi}{\partial t}\right)_{t=0} = i(k-k_0)\left[\Phi(x,0) - \frac{x}{l_1} + 1\right]. \tag{4.15}$$

As indicated, the solution of (4.12) will be represented as a sum $\Phi = \Phi^{(1)} + \Phi^{(2)}$. The solution of the homogeneous part of (4.12), $\Phi^{(1)}$, satisfying the initial conditions (4.14) and (4.15) is obtained by a separation of variables and using a Fourier series:

$$\begin{aligned}\Phi^{(1)} = \sum_{m=1}^{\infty} \frac{l_1\zeta}{2m\pi}\sin\frac{m\pi x}{l_1}\Bigg\{ & -E_{+1}\left[\delta_{m1}(k_0 - \gamma_{m1}) - \frac{2}{m\pi}(k-k_0)\right] \\ & + E_{-1}\left[\delta_{m1}(k_0 + \gamma_{m1}) - \frac{2}{m\pi}(k-k_0)\right]\Bigg\}, \\ & + E_{\pm 1} = \exp[i(k \pm \gamma_{m1})t], \qquad \gamma_{m1} = \frac{m\pi}{l_1\zeta},\end{aligned}$$

$$\begin{aligned}\delta_{m1} = {} & \frac{2m\pi}{\alpha[(m\pi)^2 - (k_0 l_1\zeta)^2]}\{[\exp(-ik_0 l_1\zeta) - \exp(ik_0 l_1\zeta)] + \alpha_0[(-1)^{m+1} + \exp(ik_0 l_1\zeta)]\} \\ & - \frac{2}{m\pi}[(-1)^{m+1}(\alpha_0/\alpha) + 1].\end{aligned} \tag{4.16}$$

To get the solution of the inhomogeneous part of (4.12), $\Phi^{(2)}$, meeting the homogeneous conditions (4.13), Duhamel's principle is used [47]. The impulse function Φ^*

is defined as a solution of the homogeneous equation (4.12) with conditions given at $t = \eta$:

$$\Phi^*|_{t=\eta} = 0, \qquad \frac{\partial \Phi^*}{\partial t}|_{t=\eta} = \frac{x}{l_1} Z(\eta) + k^2 \left(1 - \frac{x}{l_1}\right), \tag{4.17}$$

where the derivative is equal to the right-hand side of (4.12). Then, according to Duhamel's principle, one gets the desired solution by integrating the impulse function

$$\begin{aligned} \Phi^{(2)} &= \int_0^t \Phi^* \left(\frac{x}{l_1}, t, \eta\right) d\eta \\ &= \sum_{m=1}^{\infty} \frac{l_1 \zeta}{(m\pi)^2} \sin \frac{m\pi x}{l_1} \left[(-1)^{m+1} i(E_{-1} J_{-1} + E_{+1} J_{+1}) + k^2 \right. \\ &\qquad \left. - \left(\frac{E_{-1}}{k - \gamma_{m1}} - \frac{E_{+1}}{(k + \gamma_{m1})} - \frac{2\gamma_{m1}}{k^2 - \gamma_{m1}^2}\right)\right], \\ J_{\pm 1} &= \int_0^t \frac{Z(\eta)}{E_{\pm 1}(\eta)}\, d\eta. \end{aligned} \tag{4.18}$$

The sum of (4.16) and (4.18) determines the required solution of (4.12):

$$\begin{aligned} \Phi(x,t) &= \sum_{m=1}^{\infty} \frac{l_1 \zeta}{(m\pi)^2} \sin \frac{m\pi x}{l_1} Y_{m1}, \\ Y_{m1} &= (-1)^{m+1} i(E_{-1} J_{-1} - E_{+1} J_{+1}) \\ &\quad + E_{-1} \left[\tfrac{m\pi\delta_{m1}}{2}(k_0 + \gamma_{m1}) - (k - k_0) + \frac{k^2}{k - \gamma_{m1}}\right] \\ &\quad - E_{+1} \left[\frac{m\pi\delta_{m1}}{2}(k_0 - \gamma_{m1}) - (k - k_0) + \frac{k^2}{k + \gamma_{m1}}\right] - \frac{2k^2\gamma_{m1}}{k^2 - \gamma_{m1}^2}. \end{aligned} \tag{4.19}$$

This expression, together with (4.9) and (4.11), specifies the velocity field $u_1(x,t)$ upstream of the flame. When the velocity is known, the pressure field is obtained by integrating the first equation in (3.2) from 0 to x:

$$\begin{aligned} p_1(x,t) = \alpha l_1 \zeta \exp(-ikt) &\left\{ \sum_{m=1}^{\infty} \frac{l_1 \zeta}{(m\pi)^3} \left[\cos \frac{m\pi x}{l_1} - 1\right] \left(\frac{dY_{m1}}{dt} - ikY_{m1}\right) \right. \\ &- \frac{1}{2} \left(\frac{x}{l_1}\right)^2 \left(\frac{dU}{dt} - ikU\right) \\ &\left. - ik \frac{x}{l_1} \left(\frac{x}{2l_1} - 1\right)\right\} + \exp(-ikt) p_{10}(0,0). \end{aligned} \tag{4.20}$$

It can be shown that when $t = 0$, (4.20) becomes the initial value $p_{10}(x,0)$ defined by (4.4).

4.2.2. Pressure and velocity behind the flame. The procedure of determining the pressure behind the flame is similar to that of the velocity computed in section 4.2.1. The solution is presented in the same form as (4.9) with the unknown function $F_p(x,t)$ as

$$p_2(x,t) = \alpha F_p(x,t) \exp(-ikt). \tag{4.21}$$

To have a homogeneous boundary condition that satisfies (3.6) and (3.9), a transformation, analogous to (4.11), is applied:

$$F_p(x,t) = \frac{1-x}{1-l_1}P(t) + \Phi_p(x,t). \tag{4.22}$$

Then one gets an equation similar to (4.12), but with a different right-hand side:

$$\frac{\partial^2\Phi_p}{\partial t^2} - 2ik\frac{\partial\Phi_p}{\partial t} - k^2\Phi_p - \frac{\partial^2\Phi_p}{\partial x^2} = \frac{1-x}{1-l_1}Z_p(t), \tag{4.23}$$

$$Z_p(t) = -\frac{d^2P}{dt} + 2ik\frac{dP}{dt} + k^2P, \qquad \Phi_p(l_1,t) = \Phi_p(1,t) = 0. \tag{4.24}$$

The initial conditions are found by the same way as conditions (4.14) and (4.15). Comparing (4.21) and (4.22) with $p_{20}(x,0)$, given by (4.4), yields $\Phi_p(x,0)$. The derivative is found after differentiating (4.4), (4.21), and (4.22):

$$\begin{aligned}\Phi_p(x,0) = \frac{2-\alpha_0}{\alpha}\Bigg\{&\frac{1}{1-\exp[-2ik_0(1-l_1)]}\{\exp[-ik_0(x-l_1)]\\ &-\exp[-ik_0(2-x-l_1)]\} - \frac{1-x}{1-l_1}\Bigg\},\end{aligned} \tag{4.25}$$

$$\left(\frac{\partial\Phi_p}{\partial t}\right)_{t=0} = i(k-k_0)\Phi_p(x,0). \tag{4.26}$$

The solution of (4.23) satisfying the boundary condition (4.24) and the initial conditions (4.25) and (4.26) is obtained in the same way as (4.19), giving

$$\begin{aligned}&\Phi_p(x,t) = \sum_{m=1}^{\infty}\frac{1-l_1}{(m\pi)^2}\sin\frac{m\pi(1-x)}{1-l_1}Y_{m2},\\ &Y_{m2} = (-1)^{m+1}i(E_{-2}J_{-2} - E_{+2}J_{+2}) - \frac{m\pi\delta_{m2}}{2}[E_{-2}(k_0+\gamma_{m2}) - E_{+2}(k_0-\gamma_{m2})],\\ &E_{\pm2} = \exp[i(kt\pm\gamma_{m2})t], \quad \gamma_{m2} = \frac{m\pi}{1-l_1}, \quad J_{\pm2} = \int_0^t\frac{Z_p(\eta)}{E_{\pm2}}\,d\eta,\\ &\delta_{m2} = (-1)^{m+1}\left\{\frac{2m\pi(2-\alpha_0)}{\alpha\{(m\pi)^2 - [k_0(1-l_1)]^2\}} - \frac{2\{1-\exp[-2ik_0(1-l_1)]\}}{m\pi}\right\}.\end{aligned} \tag{4.27}$$

Since the pressure is known, the velocity behind the flame is found by integrating the second of equations (3.2) from 1 to x, giving

$$\begin{aligned}u_2(x,t) = \alpha(1-l_1)\exp(-ikt)\Bigg\{&\sum_{m=1}^{\infty}\frac{1-l_1}{(m\pi)^3}\left[1-\cos\frac{m\pi(1-x)}{1-l_1}\right]\left(\frac{dY_{m2}}{dt} - ikY_{m2}\right)\\ &+\frac{1-x}{2(1-l_1)}\left(\frac{dP}{dt} - ikP\right)\Bigg\} + \exp(-ikt)\ u_2(0,0).\end{aligned} \tag{4.28}$$

The unknown functions $U(t)$ and $P(t)$, present in (4.20), (4.28) and in integrals $J_{\pm1}$ (4.19) and $J_{\pm2}$ (4.27), are found using the conjugation procedure described in the next subsection. Then the velocity and pressure fields may be calculated in each portion of the duct. These are determined upstream of the flame by (4.9), (4.11),

(4.19), and (4.20) and downstream by (4.21), (4.22), (4.27), and (4.28). Data related to the duct (l_1), the autonomous (α_0, k_0) and loudspeaker (α, k) oscillations, the fresh and burnt gases temperatures (T_1°, T_2°), and parameters (n, τ) should be known in order to perform the calculations.

In this study, we use the derived relations to estimate the characteristics of the combustion and loudspeaker oscillations that provide stabilization of initially unstable autonomous waves.

4.3. Conjugation of pressure and velocity at the flame. The conjugation conditions (3.7) and (3.8) contain the values of the pressure and velocity at the flame. To get these quantities for each part of the duct, we proceed as follows. Setting $x = l_1$ in (4.9), (4.11), (4.19), (4.20), (4.21), (4.22), (4.27), and (4.28), we substitute the results in (3.7) and (3.8). Then, taking into account that $F(l_1, t) = U(t)$ and that $F_p(l_1, t) = P(t)$, we get a system determining unknown functions $U(t)$ and $P(t)$:

$$\sum_{m=1,3,\dots}^{\infty} \frac{2l_1\zeta}{(m\pi)^3}\left(\frac{dY_{m1}}{dt} - ikY_{m1}\right) + \frac{1}{2}\left[\frac{dU}{dt} - ik(U+1)\right] + \frac{1}{l_1\zeta}\left[P - \frac{p_{10}(0,0)}{\alpha}\right] = 0,$$
$$\sum_{m=1,3,\dots}^{\infty} \frac{2(1-l_1)}{(m\pi)^3}\left(\frac{dY_{m2}}{dt} - ikY_{m2}\right) + \frac{1}{2}\left(\frac{dP}{dt} - ikP\right)$$
$$+ \frac{1}{(1-l_1)}\left[\frac{u_{20}(0,0)}{\alpha} - \frac{U + nU(t-\tau)\exp(ik\tau)}{\zeta}\right] = 0. \tag{4.29}$$

After using (4.19) and (4.27) for Y_{m1} and Y_{m2}, and two relations based on Euler's formula, $E_+ + E_- = 2\exp(ikt)\cos(\gamma_m t)$ and $E_+ - E_- = 2i\exp(ikt)\sin(\gamma_m t)$, the last system takes the form of two integro-differential equations:

$$2\exp(ikt)\int_\vartheta^t \left(\frac{d^2U}{dt^2} - 2ik\frac{dU}{dt} - k^2U\right)\exp(-ik\eta)\sum_{m=1,3,\dots}^{\infty}\frac{2}{(m\pi)^2}\cos[\gamma_{m1}(t-\eta)]\,d\eta$$
$$-\frac{1}{2}\left(\frac{dU}{dt} - ikU\right) - \frac{1}{l_1\zeta}P = G(t),$$
$$G(t) = -\sum_{m=1,3,\dots}^{\infty}\frac{4}{(m\pi)^2}\left\{i\exp(ikt)\cos(\gamma_{m1}t)\left[\frac{m\pi\delta_{m1}}{2}k_0 - (k-k_0) + \frac{k^3}{k^2-\gamma_{m1}^2}\right] - \frac{1}{2}ik\right.$$
$$\left. + \gamma_{m1}\exp(ikt)\sin(\gamma_{m1}t)\left(\frac{m\pi\delta_{m1}}{2} + \frac{k^2}{k^2-\gamma_{m1}^2}\right) - \frac{ik^3}{k^2-\gamma_{m1}^2}\right\} - \frac{1}{l_1\zeta\alpha}p_{10}(0,0), \tag{4.30}$$

$$2\exp(ikt)\int_\vartheta^t \left(\frac{d^2P}{dt^2} - 2ik\frac{dP}{dt} - k^2P\right)\exp(-ik\eta)\sum_{m=1,3,\dots}^{\infty}\frac{2}{(m\pi)^2}\cos[\gamma_{m2}(t-\eta)]\,d\eta$$
$$-\frac{1}{2}\left(\frac{dP}{dt} - ikP\right) + \frac{1}{\zeta(1-l_1)}[U + nU(t-\tau)\exp(ik\tau)] = G_p(t),$$
$$G_p(t) = -\sum_{m=1,3,\dots}^{\infty}\frac{2}{(m\pi)^2}\exp(ikt)m\pi\delta_{m2}[ik_0\cos(\gamma_{m2}t) + \gamma_{m2}\sin(\gamma_{m2}t)] + \frac{u_{20}(0,0)}{\alpha(1-l_1)}. \tag{4.31}$$

Here, the unknown functions $U(t)$ and $P(t)$ are placed on the left-hand sides, while others that are considered as known form $G(t)$ and $G_p(t)$ on the right-hand sides.

The system (4.30), (4.31) has two forms, according to different parts of the combustion process. There is no combustion during the first period, $0 \le t \le \tau$, after

injection but before burning of the propellant. Hence, in this case, the interaction index is zero ($n = 0$), and the lower limit ϑ of the integrals is the time when the injection starts, $\vartheta = 0$. For the second period, $t \geq \tau$, after the combustion starts, $n > 0$, and the lower limit ϑ is equal to the time lag ($\vartheta = \tau$). For our purpose, we use the second form of the system (4.30), (4.31) because our goal is to study the behavior of the final oscillation as $t \to \infty$.

5. Stability analysis. We perform a stability analysis on the system (4.30), (4.31). We need find out whether the functions $U(t)$ and $P(t)$, defining the velocity and pressure amplitudes, are bounded or grow infinitely as $t \to \infty$. The new variables

$$V = U\exp(-ikt), \quad V_p = P\exp(-ikt), \quad H = G\exp(-ikt), \quad H_p = G_p\exp(-ikt) \tag{5.1}$$

reduce the expressions with unknown functions $U(t)$ and $P(t)$ and their two derivatives to second derivatives of V and of V_p. Then the system (4.30), (4.31) simplifies to

$$\int_\tau^t \frac{d^2V}{dt^2} \sum_{m=1,3,\ldots}^{\infty} \frac{4}{(m\pi)^2}\cos[\gamma_{m1}(t-\eta)]\,d\eta - \frac{1}{2}\frac{dV}{dt} - \frac{V_p}{l_1\zeta} = H(t), \tag{5.2}$$

$$-\int_\tau^t \frac{d^2V_p}{dt^2} \sum_{m=1,3,\ldots}^{\infty} \frac{4}{(m\pi)^2}\cos[\gamma_{m2}(t-\eta)]\,d\eta + \frac{1}{2}\frac{dV_p}{dt} - \frac{1}{\zeta(1-l_1)}[V+nV(t-\tau)] = H_p(t). \tag{5.3}$$

5.1. Laplace transform. The Laplace transform of the system (5.2), (5.3) yields

$$[s^2V(s) - sV(\tau) - V'(\tau)] \sum_{m=1,3,\ldots}^{\infty} \frac{4}{(m\pi)^2}\frac{s}{s^2+\gamma_{m1}^2} - \frac{1}{2}[sV(s) - V(\tau)] - \frac{V_p(s)}{l_1\zeta} = H(s), \tag{5.4}$$

$$\begin{aligned} -[s^2V_p(s) - sV_p(\tau) - V_p'(\tau)] &\sum_{m=1,3,\ldots}^{\infty} \frac{4}{(m\pi)^2}\frac{s}{s^2+\gamma_{m2}^2} + \frac{1}{2}[sV_p(s) - V_p(\tau)] \\ &- \frac{1}{\zeta(1-l_1)}[V(s) + nV(s)\exp(-s\tau)] = H_p(s). \end{aligned} \tag{5.5}$$

Solving (5.4) and (5.5) for $V(s)$ and $V_p(s)$, and writing $V(s)$ as a fraction, one obtains

$$V(s) = \frac{M(s)}{N(s)}, \tag{5.6}$$

$$N(s) = \left[\sum_{m=1,3,\ldots}^{\infty} \frac{4s^3}{(m\pi)^2(s^2+\gamma_{m1}^2)} - \frac{s}{2}\right]\left[-\sum_{m=1,3,\ldots}^{\infty} \frac{4s^3}{(m\pi)^2(s^2+\gamma_{m2}^2)} + \frac{s}{2}\right] \tag{5.7}$$

$$\begin{aligned} M(s) = &\left\{\begin{array}{l} -\dfrac{1}{\zeta^2 l_1(1-l_1)}[1+n\exp(-s\tau)], \\ [sV(\tau)+V'(\tau)]\displaystyle\sum_{m=1,3,\ldots}^{\infty} \frac{4s}{(m\pi)^2(s^2+\gamma_{m1}^2)} - \frac{1}{2}V(\tau) + H(s) \end{array}\right\} \\ &\times\left[-\sum_{m=1,3,\ldots}^{\infty} \frac{4s^3}{(m\pi)^2} \times \frac{1}{s^2+\gamma_{m2}^2} + \frac{s}{2}\right] \\ &+ \frac{1}{l_1\zeta}\left\{-[sV_p(\tau) + V_p'(\tau)] \sum_{m=1,3,\ldots}^{\infty} \frac{4s}{(m\pi)^2(s^2+\gamma_{m2}^2)} + \frac{1}{2}V_p(\tau) + H_p(s)\right\}, \end{aligned} \tag{5.8}$$

$$V_p(s) = l_1\zeta\left\{V(s)\left[\sum_{m=1,3,\dots}^{\infty}\frac{4s^3}{(m\pi)^2(s^2+\gamma_{m1}^2)}-\frac{s}{2}\right]-[sV(\tau)+V'(\tau)]\right.$$
$$\left.\times\sum_{m=1,3,\dots}^{\infty}\frac{4s}{(m\pi)^2(s^2+\gamma_{m1}^2)}+\frac{1}{2}V(\tau)-H(s)\right\}. \tag{5.9}$$

It is known [48, 49] that this system is stable if all roots of the equation $N(s)=0$ are to the left of the imaginary axis in the complex s plane and if no roots on the imaginary axis are repeated. Since the last term of $N(s)$ contains the product $s\tau$, we should consider the complex variable $s\tau$ instead of s. Multiplying the equation $N(s)=0$ by τ, setting $s\tau=\sigma+i\psi$, and dividing the resulting expression into real and imaginary parts, we obtain the following two equations:

$$\left(-gX_{1R}-hX_{1I}+\frac{\sigma}{2}\right)\left(-gX_{2R}-hX_{2I}+\frac{\sigma}{2}\right)$$
$$+\left(-gX_{1I}+hX_{1R}-\frac{\psi}{2}\right)\left(gX_{2I}-hX_{2R}+\frac{\psi}{2}\right)$$
$$+\frac{\theta^2 l_1}{(1-l_1)}[1+n\exp(-\sigma)\cos\psi]=0,$$

$$X_{1R}=\sum_{m=1,3,\dots}^{\infty}\frac{4[\sigma^2-\psi^2+(m\pi\theta)^2]}{(m\pi)^2\{4\sigma^2\psi^2+[\sigma^2-\psi^2+(m\pi\theta)^2]^2\}},$$

$$g=\sigma^3-3\sigma\psi^2,\qquad X_{1I}=\sum_{m=1,3,\dots}^{\infty}\frac{8\sigma\psi}{(m\pi)^2\{4\sigma^2\psi^2+[\sigma^2-\psi^2+(m\pi\theta)^2]^2\}}, \tag{5.10}$$

$$\left(-gX_{1R}-hX_{1I}+\frac{\sigma}{2}\right)\left(gX_{2I}-hX_{2R}+\frac{\psi}{2}\right)$$
$$-\left(-gX_{1I}+hX_{1R}-\frac{\psi}{2}\right)\left(-gX_{2R}-hX_{2I}+\frac{\sigma}{2}\right)$$
$$-\frac{\theta^2 l_1}{(1-l_1)}n\exp(-\sigma)\sin\psi=0,$$

$$X_{2R}=\sum_{m=1,3,\dots}^{\infty}\frac{4\{\sigma^2-\psi^2+[m\pi l_1\theta\zeta/(1-l_1)]^2\}}{(m\pi)^2\{4\sigma^2\psi^2+\{\sigma^2-\psi^2+[m\pi l_1\theta\zeta/(1-l_1)]^2\}^2\}},$$

$$h=3\sigma^2\psi-\psi^3,\qquad X_{2I}=\sum_{m=1,3,\dots}^{\infty}\frac{8\sigma\psi}{(m\pi)^2\{4\sigma^2\psi^2+\{\sigma^2-\psi^2+[m\pi l_1\theta\zeta/(1-l_1)]^2\}^2\}}, \tag{5.11}$$

$$\theta=\frac{c_1\tau^\circ}{l_1^\circ}. \tag{5.12}$$

The characteristic system (5.10), (5.11) includes two unknowns, σ and ψ, and four nondimensional parameters. They define the flame location, l_1; the temperature ratio, $\zeta^2=T_2^\circ/T_1^\circ$; the interaction index, n; and the dimensionless time lag, θ. For each preset imaginary part, $\pm\psi$, of the complex root, the characteristic system (5.10), (5.11) allows for calculation of the real part, σ, of this root and one of four parameters if three others are known. This makes it possible to investigate the effect of these four parameters on stability. In this study, we investigate the effect of the nondimensional time lag, θ.

5.2. Stability domains (numerical results). The system (5.10), (5.11) does not have any purely real or imaginary roots. It follows from the fact that in the only meaningful case, $n > 0$, the first equation has no roots when $\psi = 0$, and the second one has no roots when $\sigma = 0$. Hence, the system (5.10), (5.11) may have only complex roots. In such a case, to determine the stability domains, one should estimate the range of parameters that corresponds to complex roots with only negative real parts.

The numerical results were obtained for the case of a centrally located flame, $l_1 = 1/2$, and a temperature ratio $\zeta^2 = 1500$ K/300 K. Since l_1 and ζ are given, the real part of the complex root, σ, and the dimensionless time lag, θ, depend on the interaction index, n, and the imaginary part, $\pm\psi$. However, (5.10) and (5.11) cannot be solved for either σ or θ. Therefore, we solve each equation for the interaction index n and then compute it by both equations. Since l_1 and ζ are given, each equation determines n as a function of the dimensionless time lag θ and of the complex root of the system $\sigma \pm i\psi$. Thus, in such a situation, the goal is to find, for each preset magnitude of $\pm\psi$, such values of θ and σ that provide identical results of n given by (5.10) and (5.11). Estimates were found by trial and error and then refined by iterations. Mathcad 8 Professional was utilized. The results obtained for $(\psi/\pi) = \pm(10/18, 14/18, 17/18)$ are given in Figure 3.

From the analysis of system (5.10), (5.11) and numerical results, we see the following.

(1) In the case of no combustion, when $n = 0$, the dimensionless time lag $\theta = 0$. Starting from this point, the dimensionless time lag increases as n grows until $n = \sim (6 - 10)$. After that point, it changes slowly, approaching an asymptote (Figure 3(a)). The asymptotic property of the function $\theta(n)$ is a result of the condition $T_2^\circ/T_1^\circ =$ constant that we used in the calculation. The data corresponding to the condition $\theta =$ constant and variable ζ (Figure 4), shows that the temperature of the burnt gases increases as the interaction index grows.

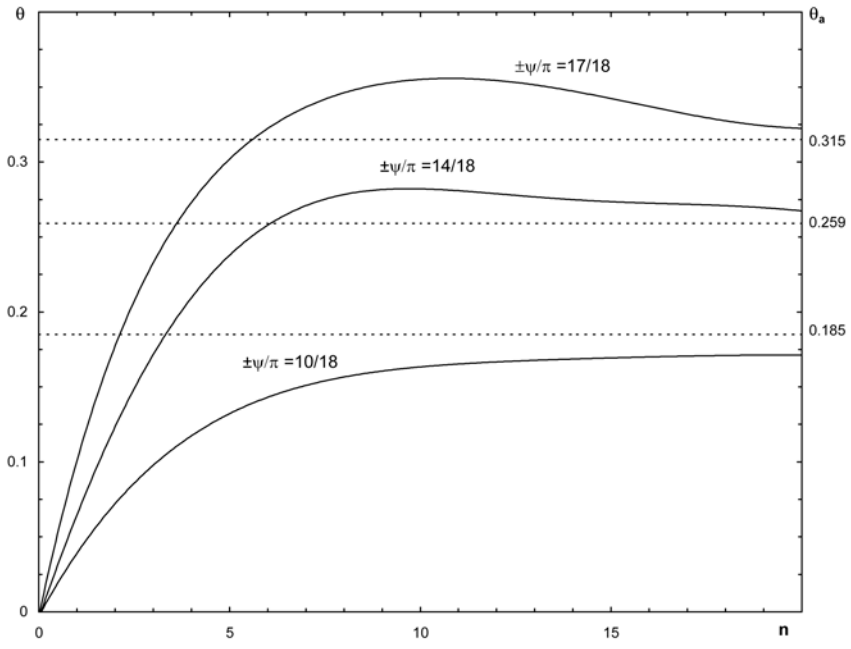
(2) As the time lag increases, the corresponding absolute value $|\sigma|$ increases as well (Figure 3(b)). After reaching a maximum over the range of $n = \sim (3 - 4)$, $|\sigma|$ rapidly decreases, approaching zero as $n \to \infty$. In this case, the product of the first two expressions in (5.10) vanishes because $\sigma = g = X_I = 0$. Then (5.10) takes the form

$$\left\{\sum_{m=1,3,\ldots}^{\infty} \frac{4\psi^3}{(m\pi)^2[\psi^2 - (m\pi\theta)^2]} - \frac{\psi}{2}\right\} \times \left\{\sum_{m=1,3,\ldots}^{\infty} \frac{4\psi^3}{(m\pi)^2 - \{\psi^2 - [m\pi l_1\theta\zeta/(1-l_1)]^2\}} - \frac{\psi}{2}\right\} - \frac{\theta^2 l_1}{(1-l_1)} = \theta^2 n\cos\psi. \tag{5.13}$$

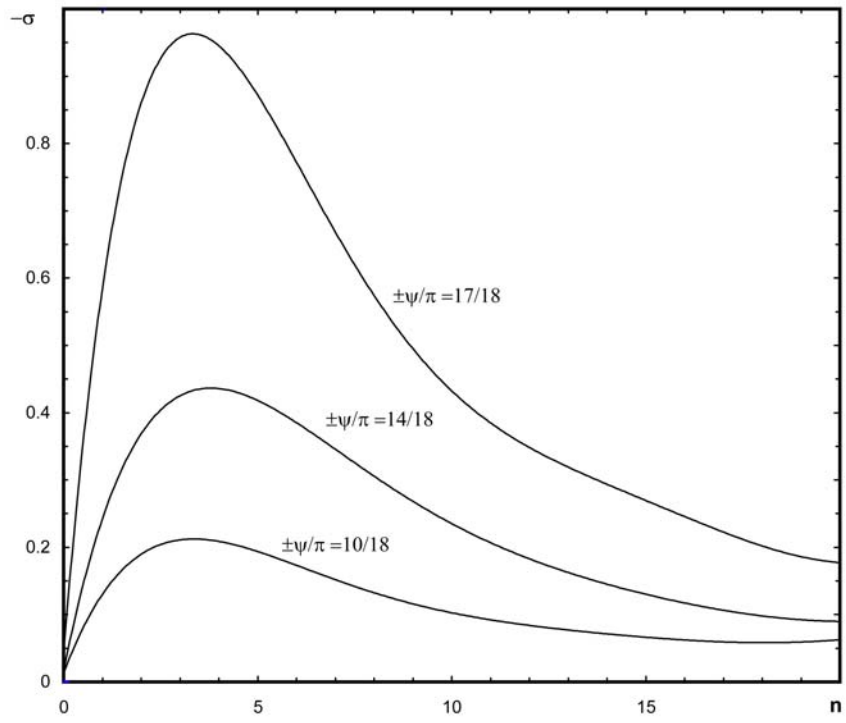
Since the right-hand side of this equation goes to infinity as $n \to \infty$, this equation is satisfied when the first sum in the left-hand side also approaches infinity. The first value of (θ/ψ) when this occurs is $0.106103292\ldots$. This value also satisfies (5.11) because this equation contains the same sum. Thus, the asymptotic value of the dimensionless time lag is

$$\theta_a = 0.106103292\ldots\psi. \tag{5.14}$$

(3) The calculation shows that in the domain of dimensionless time lag $0 < \theta < 0.36$ for all $n > 0$, the real parts, σ, of the complex roots are negative. Although



(a)



(b)

FIG. 3. *Nondimensional time lag* (a) *and real part of the complex roots of the characteristic system* (5.10), (5.11) (b) *for resulting oscillations for the case of centrally located flame* ($l_1 = 1/2$) *and temperature ratio* $T_2^0/T_1^0 = 5$.

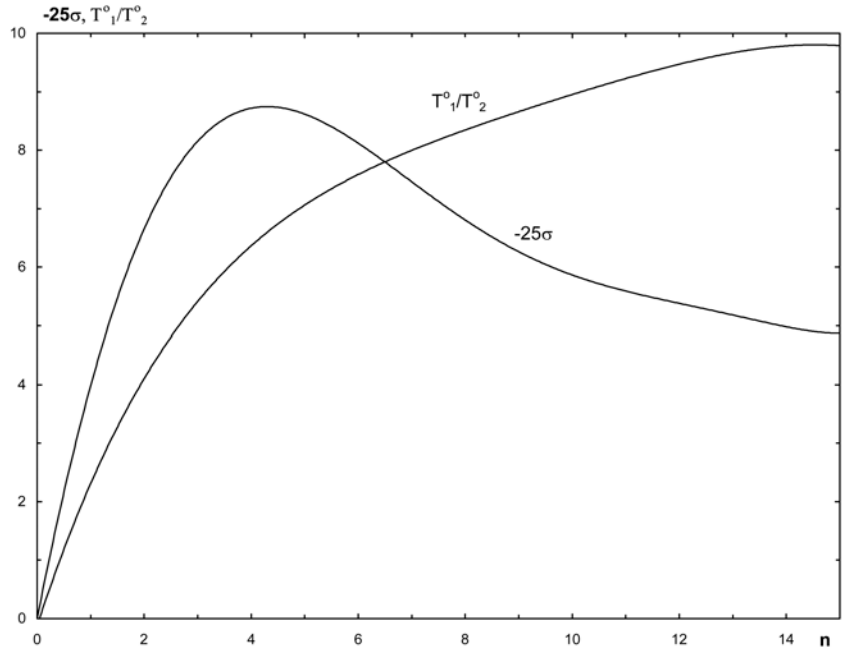


FIG. 4. *Temperature ratio and real part of the complex roots of the characteristic system* (5.10), (5.11) *for resulting oscillations for the case of centrally located flame* $l_1 = 1/2$, $\theta = 0.2$, *and* $\psi = (12/18)\pi$.

these roots correspond to stability conditions, it is not sufficient to say that the final oscillation is stable. Other roots in the obtained domains may not satisfy the stability conditions. Since we know that the system (5.10), (5.11) does not have purely real or imaginary roots, we need to show that this system does not have complex roots with positive real parts, σ, in these domains. To do it, we use two forms of (5.11).

(a) According to Figure 3(a), for small n, the stability dimensionless time lag, θ, is also small. For this case, we use the first nonzero term of the Taylor's series of (5.11),

$$\frac{2\sigma\psi}{4\sigma^2\psi^2 + (\sigma^2+\psi^2)^2} + n\exp(-\sigma)\sin\psi = 0. \tag{5.15}$$

It is clear that in the case of $\sigma > 0$ and $\psi > 0$ this equation does not have roots. Hence, there are no complex roots with positive real parts for small n and θ.

(b) For not small n, the function $\theta(n)$ has an asymptotic character. Here, the stability dimensionless time lag changes slowly so that the ratio θ/ψ remains almost constant, varying from 0.08 to 0.12. Note that, according to (5.14), this ratio is strongly constant as $n \to \infty$.

In view of this fact, we transform the two first terms in (5.11) to variables

$$\mu = \frac{\sigma}{\psi}, \qquad \nu = \frac{\theta}{\psi}. \tag{5.16}$$

After some algebra, $\sigma\psi$ appears as a factor, and (5.11) may be presented in the form

$$\sigma f(\mu,\nu) = \frac{\theta^2 l_1 n \exp(-\sigma)\sin\psi}{(1-l_1)\psi}, \tag{5.17}$$

$$f(\mu,\nu) = \left(-g^* X_{1R}^* - h^* X_{1I}^* + \frac{1}{2}\right)\left(g^* X_{2I}^* - h^* X_{2R}^* + \frac{1}{2}\right) - \left(-g^* X_{1I}^* + h^* X_{1R}^* - \frac{1}{2}\right)\left(-g^* X_{2R}^* - h^* X_{2I}^* + \frac{1}{2}\right).$$

Here, $*$ denotes the terms in (5.11) transformed to the variables defined in (5.16). Since $(\sin\psi)/\psi$ is positive, the right-hand side of (5.17) is always positive. Since the left-hand side of this equation is a product $\sigma f(\mu,\nu)$, the sign of σ coincides with that of function $f(\mu,\nu)$. Computations show that in the domain indicated above, $\nu = 0.08$–0.12, this function is negative for all positive μ. Then there are also no complex roots with positive real parts.

(4) Thus, the system (5.10), (5.11) has only complex roots with negative real parts. Hence, the corresponding values of the dimensionless time lag, $0 < \theta < 0.36$, given in Figure 3(a), determine the stability domain. Each of these θ defines, according to (5.12), a dimensional time lag $\tau^\circ = \theta(l_1^\circ/c_1)$ that theoretically provides the linear stability of the final oscillation in a duct with centrally located flame and temperature ratio 1500 K/300 K. This is true regardless of the stability of autonomous oscillations because no restrictions about this were used in the stability analysis. For other flame locations and temperature ratios, the time lag stability domains may be found by the same technique.

(5) In general, the nondimensional time lag depends on the flame location, the ratio of the burnt/fresh gas temperatures, and the interaction index. However, numerical results show that, except for relatively small n, the time lag depends only slightly on the interaction index (Figure 3(a)). Then the stability domains of the nondimensional time lag depend mainly on the flame location and the burnt/fresh gases temperature ratio if n is not small.

(6) Knowing θ, possible frequencies of forced oscillation may be estimated. Taking into account that the time lag usual is a fraction of the period of oscillation, one obtains

$$0 < \frac{\tau}{T} < 1, \qquad 0 < \omega^\circ < \frac{2\pi c_1}{l_1^\circ \theta}\left(\frac{\tau}{T}\right). \tag{5.18}$$

The performed above analysis shows that the dimensionless time lag from the stability domain $0 < \theta < 0.36$ (Figure 3(a)) and forced oscillation at frequency from the range (5.18) theoretically provide the linear stability of the final oscillation in a duct with centrally located flame and temperature ratio 1500 K/300 K no matter that initial oscillations is not stable. For other flame locations and temperature ratios, analogous stability analysis may be made by the same technique.

6. Asymptotic correlation as $t \to \infty$ for pressure and velocity behind flame. Since the problem is linear, to answer a question of stability means to find out whether the parameters of the final oscillation are bounded or become infinite as $t \to \infty$. To do this, we calculate the pressure and velocity behind the flame as $t \to \infty$. The pressure behind the flame is obtained from (4.21), while the velocity is easier to calculate by (3.8) and (4.9) than using (4.28). Considering relations (3.9) and (5.1), we have $F_p(l_1,t) = P(t) = V_p(t)\exp(ikt)$, $F(l_1,t) = U(t) = V(t)\exp(ikt)$,

and $F(l_1, t-\tau) = U(t-\tau) = V(t-\tau)\exp[ik(t-\tau)]$. Substituting these results into (4.21) and (3.8), one gets, after using (4.8),

$$\lim_{t\to\infty} p_2(l_1,t) = \alpha \lim_{t\to\infty} V_p(t), \tag{6.1}$$

$$\lim_{t\to\infty} u_2(l_1,t) = \frac{\alpha}{\zeta} \lim_{t\to\infty} [V(t) + V(t-\tau)\exp(ik\tau)]. \tag{6.2}$$

Thus, to find the parameters behind the flame as $t \to \infty$, one needs the limits of the functions $V(t)$ and $V_p(t)$. These may be determined by the inverse Laplace transform of expression (5.6) and (5.9). A second way is possible, in particular, when all roots of the denominator in (5.6) are to the left of the imaginary axis, which is simpler because it gives the result without using the inverse transform.

Since it was shown above that these conditions are satisfied, we first find the values of $V(t)$ and $V_p(t)$ as $t \to \infty$ by computing limits of $sV(s)$ and $sV_p(s)$ as $s \to 0$ [49]. The denominator in (5.6) does not become zero as $s \to 0$ because we have seen that $s = 0$ is not a root of the equation $N(s) = 0$. At the same time, it is clear that $sM(s)$, according to (5.8), approaches zero as $s \to 0$. Therefore, we have from (5.8)

$$\begin{aligned}
\lim_{t\to\infty} V(t) &= \lim_{s\to 0} sM(s) \\
&= \lim_{s\to 0} \left\{ \left\{ [sV(\tau) + V'(\tau)] \sum_{m=1,3,\ldots}^{\infty} \frac{4s^2}{(m\pi)^2(s^2+\gamma_{m1}^2)} - \frac{s}{2}V(\tau) + sH(s) \right\} \right. \\
&\times \left[\frac{s}{2} - \sum_{m=1,3,\ldots}^{\infty} \frac{4s^3}{(m\pi^2)(s^2+\gamma_{m1}^2)} \right] \\
&\left. - \frac{1}{l_1\zeta} \left\{ [sV_p(\tau) + V_p'(\tau)] \sum_{m=1,3,\ldots}^{\infty} \frac{4s^2}{(m\pi^2)(s^2+\gamma_{m2}^2)} - \frac{s}{2}V_p(\tau) - sH_p(s) \right\} \right\} \\
&= 0.
\end{aligned} \tag{6.3}$$

It follows from (5.9) that the other limit of interest also equals zero:

$$\begin{aligned}
\lim_{t\to\infty} V_p(t) &= \lim_{s\to 0} sV_p(s) \\
&= \lim_{s\to 0} l_1\zeta \left\{ sV(s) \left[\sum_{m=1,3,\ldots}^{\infty} \frac{4s^3}{(m\pi)^2(s^2+\gamma_{m1}^2)} - \frac{s}{2} \right] - [sV(\tau) + V'(\tau)] \right. \\
&\left. \times \sum_{m=1,3,\ldots} \frac{4s^2}{(m\pi)^2(s^2+\gamma_{m2}^2)} + \frac{s}{2}V(\tau) - sH(s) \right\} \\
&= 0.
\end{aligned} \tag{6.4}$$

In evaluating these limits, two others, $\lim[sH(s)] = 0$ and $\lim[sH_p(s)] = 0$, as $s \to 0$, are taken into account. They follow from (4.30), (4.31), and (5.1).

There are two types of oscillations in the t domain, which correspond to the zero limit in the s domain. They are: a suppressed oscillation and a steady-state one with zero mean value amplitude [49]. To clarify which oscillation is realized, we take the inverse Laplace transform of (5.6). This is a tiresome procedure because decomposition of the function $V(s)$ leads to sum of about 30 fractions of the types

$$\frac{1}{s^2+\gamma^2}, \quad \frac{s}{s^2+\gamma^2}, \quad \frac{1}{s+ik}, \quad \frac{1}{s^2+2\mu s+\mu^2+v^2}, \quad \frac{s}{s^2+2\mu s+\mu^2+v^2}. \tag{6.5}$$

The inverse transformation of these fractions yields the following expressions:

$$(6.6)\quad \frac{\sin\gamma t}{\gamma},\quad \cos\gamma t,\quad \exp(-ikt),\quad \frac{\exp(-\mu t)\sin\nu t}{\nu}\exp(-\mu t)\left(\cos\nu t-\frac{\mu}{\nu}\sin\nu t\right).$$

Since the two last functions become zero as $t\to\infty$, the result contains the three others:

$$(6.7)\quad \lim_{t\to\infty}V(t)=\sum_{m=1,3,\dots}^{\infty}\sum_{j=1,3,\dots}^{\infty}C_{1mj}\cos\gamma_{m1}t+D_{1mj}\sin\gamma_{m1}t+C_{2mj}\cos\gamma_{m2}t + D_{2mj}\sin\gamma_{m2}t+C_{3mj}\exp(-ikt).$$

The coefficients C_{mj} and D_{mj} depend on several others (from 4 to 9), which arise by decomposition of (5.6) into the fractions given in (6.5). These others depend on $\gamma_{m1}=m\pi/l_1\zeta,\gamma_{m2}=m\pi/(1-l_1),\mu=\sigma/\tau,v=\psi/\tau$, and k. Since the coefficients C_{mj} and D_{mj} in the double-sums (6.7) are proportional to $1/m^2$, or to $1/j^2$, or to their product, these sums converge. Hence, the function $V(t)$ is bounded as $t\to\infty$. An analogous expression for $V_p(t)$ is found by applying (5.9).

It is clear from (6.7) that the limit of the function $V(t)$ as well as the limit of the function $V_p(t)$ (as it is analogous) describe steady-state oscillatory motion with finite amplitudes and zero mean value for a period. Expressions (6.1) and (6.2) differ from functions $V(t)$ and $V_p(t)$ only by factors α and (α/ζ), respectively. Hence, the pressure and velocity behind the flame also perform steady-state oscillations with finite amplitudes and zero mean values as $t\to\infty$. Since the stability analysis was performed without any restriction on the initial oscillation stability (see subsection 5.2), this result is true no matter if the initial oscillation is stable or unstable.

Thus, the asymptotic analysis proves that the pressure and velocity behind the flame that form by interaction of initial and forced oscillations are bounded as $t\to\infty$. This issue confirms the stability analysis result: at the specific dimensionless time lags from the stability domains (Figure 3(a)), the forced oscillations at corresponding frequencies from the range (5.18) stabilize unstable initial autonomous oscillations in a duct. Hence, the active control by forced oscillation theoretically provides linear stability of combustion in the duct with initially unstable autonomous oscillations.

7. Conclusion. (1) A model and mathematical technique have been developed to simulate the input-output mechanism in an active control system. In this case, an existing oscillation in the combustor interacts with a control input and with a flame. Such a model and mathematical technique significantly differ from well-known ones for single oscillation/flame interactions.

(2) The model has been used to demonstrate the feasibility of stabilizing premixed combustion by forced oscillation. It is assumed that the oscillations at frequency ω generated by the loudspeaker are imposed on autonomous oscillations at frequency ω_0 in a duct.

(3) The problem of the interaction of two oscillations has been solved by conjugating wave equation solutions obtained separately for each portion of the duct. The whole solution takes into account (a) the initial conditions, i.e., the velocity and pressure fields existing in the combustor when the control input enters; (b) two boundary conditions in the duct; (c) conditions to conjugate the flows of fresh and burnt gases at the flame; (d) the fact that in each part of the duct only one boundary condition is known, and, hence, two others are needed. Because of that, two unknown functions are introduced which define the pressure and velocity amplitudes at the flame. The

stability analysis is reduced to a system of two integro-differential equations determining these functions.

(4) A stability analysis has been performed using the system of characteristic equations (5.10), (5.11). Two general results are derived: (a) The characteristic system does not have any purely real or imaginary roots, and (b) the nondimensional time lag (5.12) depends on the flame location, burnt/fresh gases temperature ratio, and interaction index.

(5) Numerical results and stability domains for the dimensionless time lag have been obtained for a centrally located flame and a burnt/fresh gases temperature ratio 1500 K/300 K. It has been proved that the characteristic system (5.10), (5.11) has only complex roots with negative real parts in the stability domains of the nondimensional time lag $0 < \theta < 0.36$. Knowing the time lag and the limits of the ratio τ/T, the corresponding range (5.18) of possible frequencies of forced oscillation has been estimated.

(6) It has been shown that behind the flame in the duct, the resulting pressure and velocity perform steady-state oscillations with finite amplitudes and zero mean values as $t \to \infty$. Thus, the forced oscillations at specific dimensionless time lags from the stability domains and corresponding frequencies stabilize unstable autonomous oscillations in the duct.

(7) The developed mathematical technique may be applied for different combustion models, configurations, boundary and initial conditions. In the case of the considered problem formulation for combustion in a duct, an exact solution has been obtained.

Acknowledgment. The author acknowledges Professor S. M. Meerkov of the University of Michigan, who suggested this problem and supported this work.

REFERENCES

[1] L. Crocco and S. L. Cheng, *Theory of Combustion Instability in Liquid Propellant Rocket Motors*, AGARDOGRAPH 8, Butterworths Science Publication, London, 1956.

[2] L. Crocco, J. Grey, and D. T. Harrje, *Theory of liquid propellant rocket combustion instability and its experimental verification*, J. Aero. Res. J., 30 (1960), pp. 159–168.

[3] L. Crocco, *Theoretical studies on liquid-propellant rocket instability*, in Tenth Symposium (International) on Combustion, The Combustion Institute, Pittsburgh, PA, 1965, pp. 1101–1128.

[4] L. Crocco, D. T. Harrje, and W. A. Sirignano, *Nonlinear Aspects of Combustion Instability in Liquid Propellant Rocket Motors*, NASA CR 72426, 1968.

[5] A. A. Putnam, *Combustion Driven Oscillations in Industry*, New York, Elsevier, 1971.

[6] D. T. Harrje and F. H. Reardon, *Liquid Propellant Rocket Instability*, NASA SP-194, 1972.

[7] F. E. C. Culick, *Combustion instabilities in liquid-fueled propulsion systems—an overview*, in AGARD 72B PEP Meeting, Bath, England, 1988.

[8] S. Candel, *Combustion instabilities coupled by pressure waves and their active control*, in 24th Symposium (International) on Combustion, The Combustion Institute, Pittsburgh, PA, 1992, pp. 1277–1296.

[9] K. R. McManus, T. Poinsot, and S. Candel, *A review of active control of combustion instability*, Prog. Energy Combust. Sci., 19 (1993), pp. 1–29.

[10] F. E. C. Culick, *Unsteady combustion*, in Proceedings of the NATO Advanced Study Institute on Unsteady Combustion, Praia de Granja, Portugal, 1993, F. Culick, M. V. Heitor, and J. H. Whitelaw, eds., Kluwer Academic Publishers, Dordrecht, Boston, London, 1996, pp. 173–241

[11] S. Candel, C. Huynh, and T. Poinsot, *Unsteady combustion*, in Proceedings of the NATO Advanced Study Institute on Unsteady Combustion Praia de Granja, Portugal, 1993, F. Culick, M. V. Heitor, and J. H. Whitelaw, eds., Kluwer Academic Publishers, Dordrecht, Boston, London, 1996, pp. 83–112.

[12] B. T. Zinn and Y. Neumeier, *An Overview of Active Control of Combustion Instabilities*, AIAA 97-0461, Reno, NV, 1997
[13] K. C. Schadow, V. Yang, F. E. C. Culick, T. J. Rosfjord, G. J. Sturgess, and B. T. Zinn, *Active Combustion Control for Propulsion Systems*, Report AGARD-R-820, 1997.
[14] B. S. Hong, V. Yang, and A. Ray, *Robust feedback control of combustion instability with modeling uncertainty*, Combust. Flame, 120 (2000), pp. 91–106.
[15] S. M. Kuo and D. R. Morgan, *Active Noise Control Systems*, Wiley Interscience, New York, 1996.
[16] S. Candel, *Combustion dynamics and control: Progress and challenges*, Proc. Combust. Inst., 29 (2003), pp. 1–28.
[17] S. Candel, D. Thevenin, N. Darabiha, and D. Veynante, *Progress in numerical combustion*, Combust. Sci. Tech., 149 (1999), pp. 297–337.
[18] S. Candel, D. Veynante, F. Lacas, N. Darabiha, and C. Rolon, *Current progress and future-trends in turbulent combustion*, Combust. Sci. Tech., 98 (1994), pp. 245–264.
[19] K. N. C. Bray, *Turbulent transport in flames*, Proc. R. Soc. Lond. Ser. A, 451 (1995), pp. 231–256.
[20] S. Ducruix, T. Schuller, D. Durox, and S. Candel, *Combustion dynamics and instabilities: Elementary coupling and driving mechanisms*, J. Propul. Power, 19 (2004), pp. 722–734.
[21] F. E. C. Culick, and V. Yang, *Overview of combustion instabilities in liquid-propellant rocket engines*, Liquid Rocket Engine Combustion Instability, V. Yang and W. Anderson, eds., Progress in Aeronautics and Astronautics, 169, 1995, pp. 3–37.
[22] F. E. C. Culick, *Nonlinear behavior of acoustic waves in combustion chambers*, Acta Astronaut, 3 (1976), pp. 714–757.
[23] W. Lang, T. Poinsot, and S. Candel, *Active control of combustion instability*, Combust. Flame, 70 (1987), pp. 281–289.
[24] A. Gulati and R. Mani, *Active control of unsteady combustion-induced oscillations*, J. Propul. Power, 8 (1992), pp. 1109–1115.
[25] F. E. C. Culick, W. H. Lin, C. C. Jahnke, and J. D. Sterling, *Modeling for active control of combustion and thermally driven oscillations*, in Proceedings of the American Control Conference, New York, 1991, pp. 2939–2948.
[26] V. Yang, S. I. Kim, and F. E. C. Culick, *Triggering of longitudinal pressure oscillations in combustion chambers.* 1. *Nonlinear gasdynamics*, Combust. Sci. Tech., 72 (1990), pp. 183–214.
[27] Y.-T. Fung and V. Yang, *Active control of nonlinear pressure oscillations in combustion chambers*, J. Propul. Power, 8 (1992), pp. 1282–1289.
[28] Y.-T. Fung and V. Yang, *Active control of combustion instabilities with distributed actuators*, Combust. Sci. Tech., 78 (1991), pp. 217–245.
[29] M. Krstic, A. Krupadanam, and C. Jacobson, *Self-tuning control of a nonlinear model of combustion instabilities*, IEEE Trans. Contr. Syst. Techn., 7 (1999), pp. 424–436.
[30] M. Fleifil, A. M. Annaswamy, Z. A. Ghoneim, and A. F. Ghoneim, *Response of a laminar premixed flame to flow oscillations: A kinematic model and thermoacoustic instability results*, Combust. Flame, 106 (1996), pp. 487–509.
[31] M. Fleifil, A. M. Annaswamy, Z. A. Ghoneim, and A. F. Ghoneim, *Active control of thermoacoustic instability in combustion systems*, in Proceedings of the 4th IEEE Conference on Control Applications, Albany, NY, 1995, pp. 685–690.
[32] J. P. Hathout, A. M. Annaswamy, M. Fleifil, and A. F. Ghoniem, *A model-based active control design for thermoacoustic instability*, Combust. Sci. Tech., 132 (1998), pp. 99–138.
[33] A. M. Annaswamy, O. M. El Rifai, M. Fleifil, J. P. Hathout, and A. F. Ghoniem, *A model-based self-tuning controller for thermoacoustic instability*, Combust. Sci. Tech., 135 (1998), pp. 213–240.
[34] A. M. Annaswamy and A. F. Ghoniem, *Active control of combustion instability: Theory and practice*, IEEE Contr. Syst. Mag., 22 (2002), pp. 37–54.
[35] J. P. Hathout, M. Fliefil, A. M. Annaswamy, and A. F. Ohoniem, *Combustion instability active control using periodic fuel injection*, J. Propul. Power, 18 (2002), pp. 390–399.
[36] H. J. Merk, *An analysis of unstable combustion of premixed gases*, in Proceedings of the Sixth Symposium (International) on Combustion, The Combustion Institute, Pittsburgh, PA, 1956, pp. 500–512.
[37] S. Ducruix, D. Durox, and S. Candel, *Theoretical and experimental determination of the transfer function of a laminarpremixed flame*, Proceedings of the Combustion Institute, 28 (2000), pp. 765–773.

[38] T. Schuller, D. Durox, and S. Candal, *A unified model for the prediction of laminar flame transfer function: Comparisons between conical and v-flame dynamics*, Combust. Flame, 134 (2003), pp. 21–34.

[39] A. A. Peracchio and W. M. Proscia, *Nonlinear heat-release/acoustic model for thermoacoustic instability in lean premixed combustors*, J. Eng. Gas. Turb. Power, 121 (1999), pp. 416–421.

[40] R. M. Murray, C. A. Jacobson, R. Casas, A. I. Khibnik, C. R. Johnson Jr., R. Bitmead, A. A. Peracchio, and W. M. Proscia, *System identification for limit cycling systems: A case study for combustion instabilities*, in Proceedings of the American Control Conference, Philadelphia, 1998, pp. 2004–2008.

[41] A. P. Dowling, *Nonlinear self-excited oscillations of a ducted flame*, J. Fluid Mech., 346 (1997), pp. 271–290.

[42] A. P. Dowling, *A kinematic model of a ducted flame*, J. Fluid Mech., 394 (1999), pp. 51–72.

[43] G. J. Bloxsidge, A. P. Dowling, and P. J. Langhorne, *Reheat buzz-an acoustically coupled combustion instability*, J. Fluid Mech., 193 (1988), pp. 445–473.

[44] S. Evesque and A. P. Dowling *LMS algorithm of adaptive control of combustion oscillations*, Combust. Sci. Tech., 164 (2001), pp. 65–93.

[45] A. P. Dowling and S. Hubbard, *Instability in lean premixed combustors*, Proceedings of Institution of Mechanical Engineers, Part A–J, Power Energy, 214 (2000), pp. 317–332.

[46] A. P. Dowling, *The* 1999 *Lanchester lecture: Vortices, sound and flame—a damaging combination*, Aeronaut. J., 104 (2000), pp. 105–116.

[47] R Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. 2, John Wiley & Sons, New York, 1962.

[48] H. S. Carslow and J.C. Jaeger, *Operational Methods in Applied Mathematics*, Dover Publications, New York, 1963.

[49] G. Fodor, *Laplace Transforms in Engineering*, Publishing House of the Hungarian Academy of Science, Budapest, 1965.

SIAM J. APPL. MATH.
Vol. 65, No. 4, pp. 1200–1226

ω-HARMONIC FUNCTIONS AND INVERSE CONDUCTIVITY PROBLEMS ON NETWORKS*

SOON-YEONG CHUNG† AND CARLOS A. BERENSTEIN‡

Abstract. In this paper, we discuss the inverse problem of identifying the connectivity and the conductivity of the links between adjacent pair of nodes in a network, in terms of an input-output map. To do this we deal with the weighted Laplacian Δ_ω and an ω-harmonic function on the graph, with its physical interpretation as a diffusion equation on the graph, which models an electric network. After deriving the basic properties of ω-harmonic functions, we prove the solvability of (direct) problems such as the Dirichlet and Neumann BVPs. Our main result is the global uniqueness of the inverse conductivity problem for a network under a suitable monotonicity condition.

Key words. discrete Laplacian, inverse conductivity problem, diffusion equation

AMS subject classifications. Primary, 05C40, 35R30; Secondary, 94C12

DOI. 10.1137/S0036139903432743

Introduction. A network represents a way of interconnecting any pair of users or nodes by means of some meaningful links. Thus, it is quite natural that its structure can be represented, at least in a simplified form, by a connected graph whose vertices represent nodes and whose edges represent their links.

The problem of discovering the detailed inner structure of the network from a collection of boundary measurements can be seen as a type of inverse problem, analogous to those arising in tomography. For example, problems of interest include checking connectivity, tracking data traffic, performance of software or hardware, security, reliability, and so on. In particular, when we have some problems on a part of the network or when we are in need of finding such a part having problems, it is almost impossible to investigate the whole network, since the network may be too vast and its structure or connectivity too complicated. For this reason, the study of the inverse problem to recover the whole network with partial data is becoming increasingly important for practical applications.

From the graph theoretical point of view, problems involving graph identification have been among the most important and famous open problems in graph theory [BH]. Most of the work on this subject has concentrated on spectral graph theory, on the realization of graphs with given distances, and on the reconstruction of graphs from vertex deleted subgraphs (see [B2], [B3], [Ch], [CY], [CGGS], [CO], [CL], [CvDGT], [CvDS], and [HY]). Thus far, spectral theory has been one of the most significant tools used in studying graphs, and it has led to noteworthy progress in the study of these questions. But, as is well known, graphs are not in general completely characterized by their spectra (see [CvDGT, p. 66]).

*Received by the editors August 7, 2003; accepted for publication (in revised form) July 1, 2004; published electronically April 14, 2005.

http://www.siam.org/journals/siap/65-4/43274.html

†Department of Mathematics, Sogang University, Seoul 121-742 (sychung@sogang.ac.kr). The work of this author was supported by the LG Foundation in 2002–2003 and the grant 1999-2-101-001-5 from KOSEF.

‡Department of Mathematics, University of Maryland, College Park, MD 20742 (carlos@math.umd.edu). The work of this author was supported by grants NSF-DMS0070044 and ARO-DAAD19-01-1-0494.

In this paper another method of studying the graph identification problem will be introduced—a discrete version of the inverse conductivity problem.

The original aim of the inverse conductivity problem was to identify the conductivity coefficient in continuous media from boundary measurements, such as Dirichlet data, Neumann data, or their appropriate combinations.

The discrete or finite nature of graphs makes working on graphs basically easier than investigating these problems in the continuous case. On the other hand, their discrete nature also gives rise to several disadvantages. For example, solutions of the Laplace equation (introduced in section 2) do not have the local uniqueness property, nor is their uniqueness guaranteed by the Cauchy data, contrary to the continuous case, where they are the most important mathematical tools used to study the inverse conductivity problem and related problems.

The purpose of this paper is to give a discrete analogue of the inverse conductivity problem as studied in a number of publications, such as [A], [Ca], [I], [IP], [KS], and [SU]. To do this we introduce an elliptic operator on the graph, the ω-Laplacian Δ_ω, and interpret it as a diffusion equation on the graph modeled by the electric network. Since little has been studied so far about partial differential equations on graphs, we will establish several useful properties of Δ_ω, which are essential to solve the inverse problem.

The inverse problem we study is to identify the connectivity of the nodes and the conductivity of the edges between each adjacent pair of nodes. The following global uniqueness result for the inverse conductivity problem in a network is the main result of this paper. We prove it in section 4.

THEOREM. *Let ω_1 and ω_2 be weights with $\omega_1 \le \omega_2$ on $\overline{S} \times \overline{S}$ and $f_1,\ f_2 : \overline{S} \to \mathbb{R}$ be functions satisfying, for $j = 1,\ 2$,*

$$\begin{cases} \Delta_{\omega_j} f_j(x) = 0, \ x \in S, \\ \frac{\partial f}{\partial_{\omega_j} n}(z) = \psi(z), \ z \in \partial S, \\ \int_S f_j \ d_{\omega_j} = K \end{cases}$$

for a given function $\psi : \partial S \to \mathbb{R}$ with $\int_{\partial S} \psi = 0$ and a given constant $K > 0$.

If we assume that

(i) $\omega_1(z, y) = \omega_2(z, y)$ *on* $\partial S \times \overset{\circ}{\partial} S$,

(ii) $f_1|_{\partial S} = f_2|_{\partial S}$,

then we have

$$f_1 = f_2 \text{ on } \quad \overline{S}$$

and

$$\omega_1 = \omega_2 \quad \text{on} \quad \overline{S} \times \overline{S}.$$

The second conclusion $\omega_1 = \omega_2$ above is exactly what we want to have. In fact, it shows not only whether or not each pair of nodes is connected by a link but also how nice the link is.

Both the monotonicity condition $\omega_1 \le \omega_2$ and the normalization condition $\int_S f_j \, d_{\omega_j} = K$ will be shown to be necessary by means of counterexamples. In fact, even in the continuous case, some form of monotonicity has also been considered (see [I], [Ca], and [A]).

To paraphrase the previous discussion in terms of communications networks, there are two clear ways such a network can be disrupted. One occurs when some nodes

fail or "cease to exist," in which case the structure of the network as a graph has changed. The other occurs when traffic among some nodes becomes so large that for practical purposes the network is also disrupted; this corresponds to the problem we consider here. The traffic may become extremely slow, but the structure of the underlying graph has not changed. This closely resembles the problem we consider here. Besides, there are also indications that there are possible applications to fault-testing in VLSI design and random walks with weighted transition probabilities. See also [ClMM], [AJB1], and [AJB2].

We organized this paper as follows: First, we discuss calculus on graphs in section 1, and in section 2 we introduce ω-harmonic functions on graphs and some good properties of them, which are useful later and for further study. In fact, those properties are interesting by themselves in the authors' opinion.

In section 3, we discuss the direct problems such as the Dirichlet BVP (DBVP) and Neumann BVP (NBVP) and give a physical interpretation of Δ_ω. Additional useful properties of ω-harmonic functions will be introduced.

Finally, in section 4, we prove the global uniqueness result for the inverse problem under the monotonicity condition. For its proof, we introduce a discrete version of the Dirichlet principle, which is an essential tool for the proof of the main theorem.

After the authors completed this paper, Professor Gunter Uhlmann informed the authors that Professor Morrow with his collaborators had published a series of papers (see [CM1], [CM2], [MMC], [IM] and [MIC]) on the inverse problem for networks. But their results were concentrated on the networks of special types such as circular networks or integer lattices. Moreover, their approaches do not seem to work for the networks of general type considered in this paper.

1. Calculus on weighted graphs. We shall begin with some definitions of graph theoretic notions frequently used throughout this paper.

By a *graph* $G = G(V, E)$ we mean a finite set V of *vertices* with a set E of two-element subsets of V (whose elements are called *edges*). The set of vertices and edges of a graph G are sometimes denoted by $V(G)$ and $E(G)$, or simply V and E, respectively. As conventionally used, we denote by either $x \in V$ or $x \in G$ the fact that x is a vertex in G.

A graph G is said to be *simple* if it has neither multiple edges nor loops, and G is said to be *connected* if for every pair of vertices x and y there exists a sequence (termed a *path*) of vertices $x = x_0, x_1, x_2, \dots, x_{n-1}, x_n = y$ such that x_{j-1} and x_j are connected by an edge (termed *adjacent*) for $j = 1, 2, \dots, n$.

A graph $S = S(V', E')$ is said to be a *subgraph* of $G(V, E)$ if $V' \subset V$ and $E' \subset E$. Then, we call G a *host graph* of S. If E' consists of all the edges from E which connect the vertices of V' in its host graph G, then S is called an *induced* subgraph. It is noted that an induced subgraph of a connected host graph may not be connected.

A *weighted (undirected) graph* is a graph $G(V, E)$ associated with a *weight* function $\omega : V \times V \to [0, \infty)$ satisfying

(i) $\omega(x, x) = 0, \quad x \in V$,

(ii) $\omega(x, y) = \omega(y, x) \quad$ if $x \sim y$,

(iii) $\omega(x, y) = 0 \quad$ if and only if $\quad \{x, y\} \notin E$.

Here, $x \sim y$ means that two vertices x and y are connected (adjacent) by an edge in E. In this case, $\{x, y\}$ denotes the edge connecting the vertices x and y.

In particular, a weight function ω satisfying

$$\omega(x, y) = 1 \quad \text{if } x \sim y$$

is called the *standard* weight on G. The physical meaning of the weight function will be discussed later in section 3.

The *degree* $d_\omega x$ of a vertex x in a weighted graph $G(V,E)$ with a weight ω is defined to be

$$d_\omega x := \sum_{y\in V} \omega(x,y).$$

Throughout this paper, all the subgraphs in our concern are assumed to be induced, simple, and connected subgraphs of a weighted graph. A function on a graph is understood to be a function defined just on the set of vertices.

The integration of a function $f : G \to \mathbb{R}$ on a graph $G = G(V,E)$ is defined by

$$\int_G f d_\omega \quad \left(\text{or simply } \int_G f\right) := \sum_{x\in V} f(x) d_\omega x.$$

We shall now define the directional derivative of a function $f : G \to \mathbb{R}$. For each x and $y \in V$ we define

$$D_{\omega,y} f(x) := [f(y) - f(x)] \sqrt{\frac{\omega(x,y)}{d_\omega x}}.$$

The gradient ∇_ω of function f is defined to be a vector

$$\nabla_\omega f(x) := \big(D_{\omega,y} f(x)\big)_{y\in V},$$

which is indexed by the vertices $y \in V$. Then it is easy to see that

$$\begin{aligned}\int_G |\nabla_\omega f(x)|^2 &= \sum_{x\in V} |\nabla_\omega f(x)|^2 d_\omega x \\ &= \sum_{x\in V}\sum_{y\in V} |f(y) - f(x)|^2 \omega(x,y) \\ &= 2 \sum_{\{x,y\}\in E} |f(y) - f(x)|^2 \omega(x,y).\end{aligned}$$

This integral is called the energy of f on G.

For a subgraph S of a graph $G = G(V,E)$, the (vertex) *boundary* ∂S of S is the set of all vertices $z \in V$ not in S but adjacent to some vertex in S, i.e.,

$$\partial S := \{z \in V - S | z \sim y \text{ for some } y \in S\},$$

and the *inner boundary* $\mathring{\partial} S$ is defined by

$$\mathring{\partial} S := \{y \in S | y \sim z \text{ for some } z \in \partial S\}.$$

Also, by $\overline{S}$ we denote a graph whose vertices and edges are in S and vertices in ∂S. We note here that by definition the boundary ∂S does not contain edges.

The (outward) *normal derivative* $\frac{\partial f}{\partial_\omega n}(z)$ at $z \in \partial S$ is defined to be

$$\frac{\partial f}{\partial_\omega n}(z) := \sum_{y\in S} [f(z) - f(y)] \cdot \frac{\omega(z,y)}{d'_\omega z},$$

where $d'_\omega z = \sum_{y\in S} \omega(z,y)$.

The ω-*Laplacian* Δ_ω of a function $f : G \to \mathbb{R}$ on a graph G is defined by

$$\Delta_\omega f(x) := \sum_{y \in V} [f(y) - f(x)] \cdot \frac{\omega(x,y)}{d_\omega x}, \quad x \in V. \tag{1.1}$$

For notation, notions, and conventions we refer the reader to [Ch] and [CvDS].

Remark 1.1.

(i) The discrete Laplacian on graphs can be found in several places, such as [Ch], [CvDS], and [B1]. But the ω-Laplacian defined above is not exactly the same as the one considered in those references. In fact, the definition used here will give us an advantage of a more consistent treatment in section 2.

(ii) The first derivatives and gradient in a discrete sense have not been introduced precisely so far in the literature, as far as the authors know. But the first derivative $D_{\omega,y}$ defined above may still be unsatisfactory in the sense that Leibniz's rule does not hold. In spite of this defect, it will be seen later that it has the appropriate physical meaning and works very well with respect to calculus on graphs.

In what follows, a function f defined on $\overline{S}$ may be understood as a function on its host graph G such that $f = 0$ on $G \setminus \overline{S}$ if necessary.

THEOREM 1.2. *Let S be a subgraph of a host graph G. Then for any pair of functions $f : \overline{S} \to \mathbb{R}$ and $h : \overline{S} \to \mathbb{R}$ we have*

$$2 \int_{\overline{S}} h(-\Delta_\omega f) = \int_{\overline{S}} \nabla_\omega h \cdot \nabla_\omega f. \tag{1.2}$$

Proof. A direct use of the definitions mentioned above gives

$$\begin{aligned}
2 \int_{\overline{S}} h(-\Delta_\omega f) &= 2 \sum_{x \in \overline{S}} h(x)[-\Delta_\omega f(x)] d_\omega x \\
&= -2 \sum_{x \in \overline{S}} h(x) \left\{ \sum_{y \in V(G)} [f(y) - f(x)] \omega(x,y) \right\} \\
&= 2 \sum_{x \in \overline{S}} \sum_{y \in \overline{S}} h(x)[f(x) - f(y)] \omega(x,y) \\
&= \sum_{x \in \overline{S}} \sum_{y \in \overline{S}} h(x)[f(x) - f(y)] \omega(x,y) + \sum_{x \in \overline{S}} \sum_{y \in \overline{S}} h(y)[f(y) - f(x)] \omega(x,y) \\
&= \sum_{x \in \overline{S}} \sum_{y \in \overline{S}} \left\{ [f(y) - f(x)] \sqrt{\omega(x,y)} \right\} \cdot \left\{ [h(y) - h(x)] \sqrt{\omega(x,y)} \right\} \\
&= \sum_{x \in \overline{S}} \left\{ \nabla_\omega f(x) \cdot \nabla_\omega h(x) \right\} d_\omega x \\
&= \int_{\overline{S}} \nabla_\omega f \cdot \nabla_\omega h. \qquad \square
\end{aligned}$$

The above theorem yields many useful formulas such as the graph version of the Green theorem.

COROLLARY 1.3. *Under the same hypotheses as above we have the following identities:*

(i)

$$2\int_{\overline{S}} f(-\Delta_\omega f) = \int_{\overline{S}} |\nabla_\omega f|^2 .$$

(ii)

$$\int_{\overline{S}} h\Delta_\omega f = \int_{\overline{S}} f\Delta_\omega h.$$

(iii) *(Green's formula)*

$$\int_S (f\Delta_\omega h - h\Delta_\omega f) = \int_{\partial S} \left(f\frac{\partial h}{\partial_\omega n} - h\frac{\partial f}{\partial_\omega n}\right).$$

Proof. (i) is trivial and (ii) can be easily obtained by the symmetry in (1.2). We prove (iii). In view of (ii) we have

$$\begin{aligned} 0 &= \int_{\overline{S}} [f\Delta_\omega h - h\Delta_\omega f] \\ &= \int_S [f\Delta_\omega h - h\Delta_\omega f] + \int_{\partial S} [f\Delta_\omega h - h\Delta_\omega f]. \end{aligned}$$

Then, since S is the induced subgraph, it follows that $\omega(z,y) = 0$ for all z and $y \in \partial S$ and

$$\begin{aligned} \int_S [f\Delta_\omega h - h\Delta_\omega f] &= \int_{\partial S} [h\Delta_\omega f - f\Delta_\omega h] \\ &= \sum_{z\in\partial S} [h(z)\Delta_\omega f(z) - f(z)\Delta_\omega h(z)]d_\omega z \\ &= \sum_{z\in\partial S}\sum_{y\in S} \Big\{ h(z)\big[f(y) - f(z)\big]\omega(z,y) - f(z)\big[h(y) - h(z)\big]\omega(z,y)\Big\} \\ &= \sum_{z\in\partial S} \left[h(z)\left\{ -\frac{\partial f}{\partial_\omega n}\right\} + f(z)\frac{\partial h}{\partial_\omega n}(z)\right] d_\omega z \\ &= \int_{\partial S} \left[f\frac{\partial h}{\partial_\omega n} - h\frac{\partial f}{\partial_\omega n}\right]. \qquad \square \end{aligned}$$

In the continuous case, the following are well-known formulas:

$$\Delta(fg) = f\Delta g + 2\nabla f\cdot\nabla g + g\Delta f,$$

$$\int_\Omega \nabla f\cdot\nabla g + \int_\Omega f\Delta g = \int_{\partial\Omega} f\frac{\partial g}{\partial n}.$$

Here we introduce a discrete analogue of these formulas.

THEOREM 1.4. *Under the same hypotheses as in Theorem* 1.2*, the following identities hold:*

(i)

$$\Delta_\omega(fh) = f\Delta_\omega h + \nabla_\omega f\cdot\nabla_\omega h + h\Delta_\omega f,$$

(ii)

$$\int_S \nabla_\omega f\cdot\nabla_\omega h + \int_S [f\Delta_\omega h + h\Delta_\omega f] = \int_{\partial S} \frac{\partial(fh)}{\partial_\omega n}.$$

Proof. (i) can be obtained by an elementary manipulation. Using now (i) and Theorem 1.2, (iii) with $h \equiv 1$ we obtain (ii). $\square$

2. ω-harmonic functions. In this section we will discuss the functional properties of functions which satisfy the equation

$$\Delta_\omega f(x) := \sum_{y\in\overline{S}} \big[f(y) - f(x)\big] \frac{\omega(x,y)}{d_\omega x} = 0. \tag{2.1}$$

For a subgraph S with boundary $\partial S \neq \phi$ of a host graph G with a weight ω we say that a function $f : \overline{S} \to \mathbb{R}$ is *ω-harmonic* on S if it satisfies (2.1) for all $x \in S$, i.e.,

$$f(x) = \frac{1}{d_\omega x} \sum_{y\in\overline{S}} f(y)\omega(x,y), \quad x \in S.$$

This implies that the value of f at x is given by a weighted average of the values of f at its neighboring vertices. From this point of view, we can clearly expect the following result to be true.

THEOREM 2.1 (minimum and maximum principle). *Let S be a subgraph of a host graph G with a weight ω and $f : \overline{S} \to \mathbb{R}$ be a function.*

(i) *If $\Delta_\omega f(x) \geq 0, x \in S$, and f has a maximum at a vertex in S, then f is constant.*
(ii) *If $\Delta_\omega f(x) \leq 0, x \in S$, and f has a minimum at a vertex in S, then f is constant.*
(iii) *If $\Delta_\omega f(x) = 0, x \in S$, and f has either a minimum or maximum in S, then f is constant.*
(iv) *If $\Delta_\omega f(x) = 0, x \in S$, and f is constant on the boundary ∂S, then f is constant.*

Proof. (ii) can be done in a similar way as in (i). (iii) and (iv) are easily obtained from (i) and (ii).

We prove (i). Assume that f has a maximum at a vertex $x_0 \in S$. Then

$$f(x_0) \geq f(y), \quad y \in \overline{S}, \tag{2.2}$$

and

$$f(x_0) \leq \sum_{y\in\overline{S}} f(y) \frac{\omega(x_0,y)}{d_\omega x_0}. \tag{2.3}$$

Suppose that there exists $y_0 \in \overline{S}$ such that $x_0 \sim y_0$ and $f(x_0) \neq f(y_0)$, i.e., $f(x_0) > f(y_0)$ in view of (2.2). Then it follows from (2.3) that

$$\begin{aligned} f(x_0) &\leq \sum_{\substack{y\in\overline{S}\\ y\neq y_0}} \frac{f(y)\omega(x_0,y)}{d_\omega x_0} + \frac{f(y_0)\omega(x_0,y_0)}{d_\omega x_0} \\ &< \sum_{\substack{y\in\overline{S}\\ y\neq y_0}} \frac{f(x_0)\omega(x_0,y)}{d_\omega x_0} + \frac{f(x_0)\omega(x_0,y_0)}{d_\omega x_0} \\ &= f(x_0), \end{aligned}$$

which implies that $f(x_0) = f(y)$ for all $y \in \overline{S}$ such that $y \sim x_0$. Now, for any $x \in \overline{S}$, there exists a path

$$x_0 \sim x_1 \sim x_2 \sim \cdots \sim x_{n-1} \sim x_n = x,$$

since S is connected. By applying the same argument as above inductively we see that $f(x_0) = f(x)$. □

The following is an easy consequence of the above theorem.

COROLLARY 2.2. *Under the same hypotheses as in Theorem* 2.1, *the following statements are true:*

(i) *If* $\Delta_\omega f \geq 0$ *on* S *and* $f|_{\partial S} \leq 0$ (< 0), *then* $f \leq 0$ (< 0) *on* S.

(ii) *If* $\Delta_\omega f \leq 0$ *on* S *and* $f|_{\partial S} \geq 0$ (> 0), *then* $f \geq 0$ (> 0) *on* S.

COROLLARY 2.3. (1) *If two functions* f *and* g *on* $\overline{S}$ *satisfy*

$$\Delta_\omega f = 0 \text{ and } \Delta_\omega g \geq 0$$

on S, *then* $g|_{\partial S} \leq f|_{\partial S}$ *implies* $g \leq f$ *on* S.

(2) *If a function* $f : \overline{S} \to \mathbb{R}$ *satisfies*

$$\Delta_\omega f(x) = 0, \quad x \in S,$$

and $|f|$ *has a maximum in* S, *then* f *is constant.*

In the continuous case, it is well known that a local maximum principle holds for a harmonic function in an open subset $\Omega \subset \mathbb{R}^n$. But it is not hard to see that the local maximum principle is no longer true in general in our case. Moreover, the local uniqueness principle does not hold in general. As a matter of fact, it is rather natural to expect that such discrepancies are caused by the discrete nature of graphs.

A nonempty subset Γ of vertices of a subgraph $\overline{S}$ is said to be a surface in $\overline{S}$ if $\Gamma = \partial T$ for a subgraph T whose vertices belong to S. In this case, we denote by $\overset{\circ}{\Gamma}$ the inner boundary $\overset{\circ}{\partial} T$. For each vertex $z \in \Gamma$ and $x \in \overset{\circ}{\Gamma}$ we define

$$d'_\omega z := \sum_{y \in \overset{\circ}{\Gamma}} \omega(y, z) \quad \text{(inward degree)}$$

and

$$d''_\omega x := \sum_{z \in \Gamma} \omega(x, z) \quad \text{(outward degree)}.$$

In addition, for a function f on $\overline{S}$ we write

$$\int_\Gamma f(z) d'_\omega z = \sum_{z \in \Gamma} f(z) d'_\omega z \quad \text{(inward integral)}$$

and

$$\int_{\overset{\circ}{\Gamma}} f(x) d''_\omega x = \sum_{x \in \overset{\circ}{\Gamma}} f(x) d''_\omega x \quad \text{(outward integral)}.$$

We use these notions to obtain the following interesting properties of ω-harmonic functions.

THEOREM 2.4. *Let* S *be a subgraph of a host graph with weight* ω *and let* $f : \overline{S} \to \mathbb{R}$. *Then* f *is* ω*-harmonic on* S, *i.e., for all* $x \in S$,

$$\Delta_\omega f(x) = 0, \tag{2.4}$$

if and only if for every surface Γ *in* $\overline{S}$

$$\int_{\Gamma} f(z)d'_{\omega}z = \int_{\overset{\circ}{\Gamma}} f(y)d''_{\omega}y. \tag{2.5}$$

Proof. Let $x \in S$ and $\Gamma_x = \{y \in \overline{S} | x \sim y\}$. Then Γ_x is a surface in $\overline{S}$ and $\overset{\circ}{\Gamma}_x = \{x\}$. Since $d_{\omega}x = d''_{\omega}x$ on $\overset{\circ}{\Gamma}_x$ and $d'_{\omega}z = \omega(x,z)$, (2.5) implies

$$f(x)d_{\omega}x = \sum_{z\in\Gamma_x} f(z)\omega(x,z),$$

which implies (2.4) immediately

Assume now that (2.4) holds and let Γ be a surface in $\overline{S}$ such that $\Gamma = \partial T$ for a subgraph $T \subset S$. We use Green's formula (Corollary 1.3, (iii)) to obtain

$$\begin{aligned} 0 &= \int_T \Delta_{\omega} f \\ &= \int_{\Gamma} \frac{\partial f}{\partial_{\omega} n} \\ &= \sum_{z\in\Gamma} \frac{\partial f}{\partial_{\omega} n}(z) d'_{\omega}z \\ &= \sum_{z\in\Gamma}\sum_{y\in\overset{\circ}{\Gamma}} \big[f(z) - f(y)\big]\omega(z,y). \end{aligned} \tag{2.6}$$

Then it follows that

$$\sum_{z\in\Gamma}\sum_{y\in\overset{\circ}{\Gamma}} f(z)\omega(z,y) = \sum_{z\in\Gamma}\sum_{y\in\overset{\circ}{\Gamma}} f(y)\omega(z,y)$$

or, equivalently,

$$\sum_{z\in\Gamma} f(z)\Big[\sum_{y\in\overset{\circ}{\Gamma}} \omega(z,y)\Big] = \sum_{y\in\overset{\circ}{\Gamma}} f(y)\Big[\sum_{z\in\Gamma} \omega(z,y)\Big],$$

which yields (2.5). □

In view of (2.6) we obtain the edge version of Theorem 2.4, the so-called dual theorem, as follows.

COROLLARY 2.5. *Under the same conditions as in Theorem* 2.4, *the formula* (2.5) *is equivalent to*

$$\sum_{\{z,y\}\in E(\Gamma,\overset{\circ}{\Gamma})} \big[f(z) - f(y)\big]\omega(z,y) = 0,$$

where $E(\Gamma, \overset{\circ}{\Gamma})$ *denotes the set of all edges joining a vertex in* Γ *and a vertex in* $\overset{\circ}{\Gamma}$.

For two vertices x and y in a connected graph, the distance $d(x,y)$ between x and y is the number of edges in a shortest path joining x and y.

For a vertex x_0 in a subgraph S we write

$$\Gamma_j(x_0) := \big\{y \in \overline{S} | \quad d(x_0,y) = j\big\}, \quad j = 0,1,2,\ldots,$$

which is called a neighborhood of x_0 with radius j.

Then the following is a variant of Theorem 2.4.

COROLLARY 2.6. *Let S and f be the same as in Theorem* 2.4. *Then f is ω-harmonic on S if and only if for every $x_0 \in S$*

$$\int_{\Gamma_j(x_0)} f(x) d''_\omega x = \int_{\Gamma_{j+1}(x_0)} f(x) d'_\omega x \tag{2.7}$$

for each j with $\Gamma_j(x_0) \subset S$.

Proof. Letting $j = 0$ in (2.7), we have the sufficiency. To prove the necessity, consider an induced subgraph T whose vertices are exactly those of $\bigcup_{k=0}^{j} \Gamma_k(x_0)$. Then it is easy to see that

$$\partial T = \Gamma_{j+1}(x_0) \text{ and } \overset{\circ}{\partial} T \subset \Gamma_j(x_0).$$

But a vertex x in $\Gamma_j(x_0)$, which does not belong to $\overset{\circ}{\partial} T$, does not make any contribution to the outer integral $\int_{\Gamma_j(x_0)} f(x) d''_\omega x$, since $d''_\omega x = 0$. Hence, condition (2.5) in Theorem 2.4 shows the condition is necessary. □

The following is the dual version of the above corollary.

COROLLARY 2.7. *Under the same conditions as in Corollary* 2.6 *the formula* (2.7) *is equivalent to*

$$\sum_{\{x,y\} \in E(\Gamma_j(x_0), \Gamma_{j+1}(x_0))} \big[f(x) - f(y)\big] \omega(x,y) = 0,$$

where $E(\Gamma_j(x_0), \Gamma_{j+1}(x_0))$ denotes the set of all edges joining a vertex in $\Gamma_j(x_0)$ and a vertex in $\Gamma_{j+1}(x_0)$.

3. The Dirichlet and Neumann BVPs: Direct problems. In this section, we discuss the direct problems such as the Dirichlet BVP (DBVP) and Neumann BVP (NBVP) (cf. [Ch], [CY], [CO], [BCE1], and [BCE2]).

We start this section with a physical interpretation of the ω-Laplace and ω-Poisson equations. Consider a host graph G with a weight ω and an (induced) subgraph S. For a surface Γ in $\overline{S}$ with $\Gamma = \partial T$ for some $T \subset S$ and $z \in \Gamma$, the flux of energy passing through z to its adjacent nodes in T is given by

$$-\sum_{y \sim z} \big[f(z) - f(y)\big] \cdot \frac{\omega(z,y)}{d'z}, \tag{3.1}$$

where $d'z = \sum_{y \sim z, y \in T} \omega(z,y)$ and f is a potential function in a diffusion field on a network (for example, an electrostatic field, a thermal field, or an elastic membrane). Here, the weight $\omega(z,y)$ plays the role of the conductivity of the diffusion along the edge $\{z,y\}$. In fact, (3.1) is exactly $-\frac{\partial f}{\partial_\omega n}(z)$ on Γ by definition (see section 1), and thus, by Green's formula we have

$$\int_T (-\Delta_\omega f) = \int_\Gamma \left(-\frac{\partial f}{\partial_\omega n}\right),$$

which is the flow across Γ.

On the other hand, assume that T gains (or loses) an amount of energy $\int_T g$, where g is the energy density. Then we have

$$\int_T (-\Delta_\omega f) = \int_T g.$$

Therefore, since T is arbitrary, by taking T to be any single vertex $x \in S$ we obtain the vertex equation

$$-\Delta_\omega f(x) = g(x), \quad x \in S. \tag{3.2}$$

Thus, it is reasonable to say that the conductivity equation on a graph can be represented as in (3.2), where $\omega(x, y)$ corresponds to the edge conductivity on the edge x, y.

Following the work of Fan Chung and her collaborators [Ch], [CY], and [CO], we will discuss first (3.2) on a graph $G = G(V, E)$ with a weight ω and no boundary. We consider the matrix

$$\Delta_\omega(x, y) = \begin{cases} -1 & \text{if } x = y, \\ \frac{\omega(x,y)}{d_\omega x} & \text{if } x \sim y, \\ 0 & \text{otherwise.} \end{cases}$$

We can consider the function f as a N-dimensional vector, where $N = |V|$ denotes the number of vertices of the graph G. Thus, (3.2) can be understood as a matrix linear equation. Let D denote the diagonal matrix with the (x, x)th entry having the value $d_\omega x$ for each x and $\mathcal{L}_\omega = D^{1/2}\Delta_\omega D^{-1/2}$. Then $(-\mathcal{L}_\omega)$ is a nonnegative definite symmetric matrix so that it has the eigenvalues

$$\lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{N-1}$$

and corresponding eigenfunctions

$$\Phi_0, \Phi_1, \Phi_2, \ldots, \Phi_{N-1}, \tag{3.3}$$

which are orthonormal in the sense that for each pair of distinct i and j

$$\sum_{x \in V} \Phi_i(x) \cdot \Phi_j(x) = 0,$$

while, for all j,

$$\sum_{x \in V} |\Phi_j(x)|^2 = 1.$$

It is easy to show that (see [Ch]) $\lambda_0 = 0, \lambda_1 > 0$ and $\Phi_0(x) = \frac{\sqrt{d_\omega x}}{\sqrt{\mathrm{vol}(G)}}$, $x \in V$, and $\mathrm{vol}(G) := \sum_{x \in V} d_\omega x$.

In what follows, we occasionally use the notation $\langle\ ,\ \rangle_X$, defined by $\langle f, g\rangle_X = \sum_{x \in X} f(x)g(x)$ for simplicity. Now we have the following solvability result for the Poisson equation.

THEOREM 3.1. *Let $G = G(V, E)$ be a graph with a weight ω and $f : G \to \mathbb{R}$ be a function. Then the equation*

$$\Delta_\omega f(x) = g(x), \quad x \in V, \tag{3.4}$$

has a solution if and only if $\int_G g = 0$. *In this case, the solution is given by*

$$f(x) = a_0 + \langle \Gamma_\omega(x,\cdot), g\rangle_V, \quad x \in V, \tag{3.5}$$

where a_0 *is an arbitrary constant and*

$$\Gamma_\omega(x,y) = \sum_{j=1}^{N-1} \left(-\frac{1}{\lambda_j}\right) \Phi_j(x)\Phi_j(y)\sqrt{\frac{d_\omega y}{d_\omega x}}, \quad x,y \in V. \tag{3.6}$$

Proof. Assume that $\int_G g = 0$. Then

$$\begin{aligned}\langle D^{1/2}g, \Phi_0\rangle &= \sum \sqrt{d_\omega x}\, g(x) \cdot \frac{\sqrt{d_\omega x}}{\sqrt{\mathrm{vol}G}} \\ &= \frac{1}{\sqrt{\mathrm{vol}G}} \int_G g \\ &= 0,\end{aligned}$$

where D is the diagonal matrix whose xth diagonal entry is $d_\omega x$.

Consider the orthogonal expansion

$$\left(D^{1/2}f\right)(x) = \sum_{j=0}^{N-1} a_j \Phi_j(x), \quad x \in V,$$

where $a_j = \langle D^{1/2}f, \Phi_j\rangle, j = 0,1,2,\ldots,N-1$. Then since $\mathcal{L}_\omega D^{1/2} = D^{1/2}\Delta_\omega$ and

$$\begin{aligned}-\lambda_j a_j &= \langle D^{1/2}f, \mathcal{L}_\omega \Phi_j\rangle \\ &= \langle \mathcal{L}_\omega D^{1/2} f, \Phi_j\rangle \\ &= \langle D^{1/2}g, \Phi_j\rangle,\end{aligned}$$

we have

$$a_j = \left(-\frac{1}{\lambda_j}\right)\langle D^{1/2}g, \Phi_j\rangle, \quad j = 1,2,\ldots,N-1,$$

and a_0 is an arbitrary constant. Hence

$$\sqrt{d_\omega x} f(x) = a_0 \frac{\sqrt{d_\omega x}}{\sqrt{\mathrm{vol}G}} + \sum_{j=1}^{N-1} \left(-\frac{1}{\lambda_j}\right)\left[\sum_{y\in V} g(y)\Phi_j(y)\sqrt{d_\omega y}\right]\Phi_j(x).$$

This is equivalent to

$$f(x) = \frac{a_0}{\sqrt{\mathrm{vol}G}} + \sum_{j=1}^{N-1}\left(-\frac{1}{\lambda_j}\right)\sum_{y\in V} g(y)\Phi_j(y)\frac{\sqrt{d_\omega y}}{\sqrt{d_\omega x}}\Phi_j(x),$$

which gives (3.5) with a different constant a_0. Conversely, a simple computation shows that

$$\Delta_\omega \cdot \Gamma_\omega\, g(x) = g(x) + \frac{1}{\mathrm{vol}G}\int_G g, \quad x \in V,$$

which implies that every function of the form (3.5) gives a solution to the equation (3.4).

The proof of the converse is easy. □

The matrix Γ_ω in (3.6) is called the Green function of Δ_ω. The following corollary is a Liouville-type theorem for ω-harmonic functions.

COROLLARY 3.2. *Under the same conditions as in Theorem* 3.1, *every solution* f *of*

$$\Delta_\omega f(x) = 0, \quad x \in V,$$

is constant.

The following corollary describes all functions which are ω-harmonic except possibly on a given (singularity) set T.

COROLLARY 3.3. *Under the same conditions as in Theorem* 3.1, *let* $T \subset V$. *Then every solution to*

$$\Delta_\omega f(x) = 0, \quad x \in V \setminus T,$$

can be represented as

$$f(x) = a_0 + \sum_{y \in T} \Gamma_\omega(x, y)\alpha(y), \quad x \in V, \tag{3.7}$$

where a_0 *is an arbitrary constant and*

$$\alpha(y) = \Delta_\omega f(y), \quad y \in T.$$

In particular, if $T = \{x_0\},\ x_0 \in V$, then (3.7) can be written as

$$f(x) = a_0 + \alpha_0 \Gamma_\omega(x, x_0), \quad x \in V,$$

where $\alpha_0 = \Delta_\omega f(x_0)$.

Let us now turn to BVPs and their eigenvalues. For a subgraph S of a host graph G with a weight ω, the *Dirichlet eigenvalues* of $-\mathcal{L}_\omega = -D^{1/2}\Delta_\omega D^{-1/2}$ are defined to be the eigenvalues

$$\nu_1 \le \nu_2 \le \cdots \le \nu_n$$

of the matrix $-\mathcal{L}_{\omega,S}$, where $\mathcal{L}_{\omega,S}$ is a submatrix of $\mathcal{L}_\omega$ with rows and columns restricted to those indexed by vertices in S and $n = |S|$. Let $\phi_1, \phi_2, \ldots, \phi_n$ be the linearly independent functions on $\overline{S}$ such that for each $j = 1, 2, \ldots, n$,

$$\mathcal{L}_{\omega,S}\phi_j(x) = (-\nu_j)\phi_j(x), \quad x \in S, \quad \text{and} \quad \phi_j|_{\partial S} = 0.$$

In fact, $\phi_1, \phi_2, \ldots, \phi_n$ are the eigenfunctions corresponding to $\nu_1 \le \nu_2 \le \cdots \le \nu_n$ and can be assumed to be orthonormal in the same sense as above, namely, that for each pair of distinct i and j

$$\sum_{x \in S} \phi_i(x) \cdot \phi_j(x) = 0,$$

while, for all j,

$$\sum_{x \in S} |\phi_j(x)|^2 = 1.$$

As usual, the first eigenvalue $\nu_1 > 0$. (See, for instance, [Ch].)

One can follow now the standard procedure to define Green functions $\gamma_{\omega,S}$ as follows:

$$\gamma_{\omega,S}(x,y)=\sum_{j=1}^{|S|}\left(-\frac{1}{\nu_j}\right)\phi_j(x)\phi_j(y)\frac{\sqrt{d_\omega y}}{\sqrt{d_\omega x}},\quad x,y\in S. \tag{3.8}$$

Letting D_S stand for the diagonal matrix whose xth entry is $d_\omega x$ for each $x\in S$ and setting $\Delta_{\omega,S}=D_S^{-1/2}\mathcal{L}_{\omega,S}D_S^{1/2}$, one can easily verify that

$$\gamma_{\omega,S}\Delta_{\omega,S}=\Delta_{\omega,S}\gamma_{\omega,S}=I \tag{3.9}$$

and

$$\Delta_{\omega,S}(x,y)=\sum_{j=1}^{|S|}(-\nu_j)\phi_j(x)\phi_j(y)\frac{\sqrt{d_\omega y}}{\sqrt{d_\omega x}},\quad x,y\in S, \tag{3.10}$$

where I denotes the $|S|$-dimensional identity matrix.

The DBVP was solved by Chung in [CY], when the graph has the standard weight. (For the interested reader, despite some minor errata, the proof given there is correct.) We prove now the solvability of the DBVP for graphs with arbitrary weights using a different method.

THEOREM 3.4. *Let S be a subgraph of a host graph with a weight ω and $\sigma:\partial S\to\mathbb{R}$ be a given function. Then the unique solution f to the Dirichlet boundary value problem (DBVP)*

$$\begin{cases}\Delta_\omega f(x)=0, & x\in S,\\ f|_{\partial S}=\sigma\end{cases}$$

can be represented as

$$f(x)=-\langle\gamma_\omega(x,\cdot),B_\sigma\rangle_{y\in S},\quad x\in S, \tag{3.11}$$

where

$$B_\sigma(y)=\sum_{z\in\partial S}\frac{\sigma(z)\omega(y,z)}{d_\omega y},\quad y\in S. \tag{3.12}$$

Proof. Let f be a solution of the DBVP. Then

$$\begin{aligned}0&=\sum_{y\in S}\gamma_{\omega,S}(x,y)\Delta_\omega f(y)\\ &=\sum_{j=1}^{|S|}\left(-\frac{1}{\nu_j}\right)\frac{\phi_j(x)}{\sqrt{d_\omega x}}\left[\sum_{y\in S}\phi_j(y)\sqrt{d_\omega y}\Delta_\omega f(y)\right]\\ &=\sum_{j=1}^{|S|}\left(-\frac{1}{\nu_j}\right)\frac{\phi_j(x)}{\sqrt{d_\omega x}}\left[\int_S\left(D_S^{-1/2}\phi_j\right)\cdot\Delta_\omega f\right]\\ &=\sum_{j=1}^{|S|}\left(-\frac{1}{\nu_j}\right)\frac{\phi_j(x)}{\sqrt{d_\omega x}}\left[\int_S f\cdot\Delta_\omega\left(D_S^{-1/2}\phi_j\right)\right.\\ &\quad\left.+\int_{\partial S}\left\{\left(D_S^{-1/2}\phi_j\right)\cdot\frac{\partial f}{\partial_\omega n}-f\cdot\frac{\partial}{\partial_\omega n}\left(D_S^{-1/2}\phi_j\right)\right\}\right].\end{aligned} \tag{3.13}$$

Here, we have used Green's formula from Corollary 1.3. On the other hand, one can show that

$$\Delta_\omega\big(D_S^{-1/2}\phi_j\big)(x) = (-\nu_j)\big(D_S^{-1/2}\phi_j\big)(x), \quad x \in S,$$

since $\phi_j = 0$ on ∂S. From this identity and orthonormality of ϕ_j we can conclude that

$$\begin{aligned}
&\sum_{j=1}^{|S|}\left(-\frac{1}{\nu_j}\right)\frac{\phi_j(x)}{\sqrt{d_\omega x}}\left[\int_S f\cdot\Delta_\omega\big(D_S^{-1/2}\phi_j\big)\right]\\
&=\sum_{j=1}^{|S|}\frac{\phi_j(x)}{\sqrt{d_\omega x}}\cdot\left[\sum_{y\in S} f(y)\cdot\frac{\phi_j(y)}{\sqrt{d_\omega y}}\cdot d_\omega y\right]\\
&=\sum_{y\in S} f(y)\left[\sum_{j=1}^{|S|}\phi_j(x)\phi_j(y)\sqrt{\frac{d_\omega y}{d_\omega x}}\,\right]\\
&= f(x).
\end{aligned}$$

Hence, from the equality (3.13) and the fact that $\phi_j = 0$ on ∂S, we have

$$\begin{aligned}
f(x) &= \sum_{j=1}^{|S|}\left(-\frac{1}{\nu_j}\right)\frac{\phi_j(x)}{\sqrt{d_\omega x}}\int_{\partial S}\left[f\cdot\frac{\partial}{\partial_\omega n}\big(D_S^{-1/2}\phi_j\big)\right]\\
&=\sum_{j=1}^{|S|}\left(-\frac{1}{\nu_j}\right)\frac{\phi_j(x)}{\sqrt{d_\omega x}}\left[\sum_{z\in\partial S} f(z)\cdot\frac{\partial}{\partial_\omega n}\big(D_S^{-1/2}\phi_j\big)(z)\cdot dz\right]\\
&=\sum_{j=1}^{|S|}\left(-\frac{1}{\nu_j}\right)\frac{\phi_j(x)}{\sqrt{d_\omega x}}\sum_{z\in\partial S}\sigma(z)dz\left[\sum_{y\in S}\left\{\frac{\phi_j(z)}{\sqrt{d_\omega z}}-\frac{\phi_j(y)}{\sqrt{d_\omega y}}\right\}\frac{\omega(z,y)}{d_\omega z}\right]\\
&=-\sum_{j=1}^{|S|}\left(-\frac{1}{\nu_j}\right)\frac{\phi_j(x)}{\sqrt{d_\omega x}}\sum_{y\in S}\phi_j(y)\sqrt{d_\omega y}\left(\sum_{z\in\partial S}\frac{\sigma(z)\omega(z,y)}{d_\omega y}\right)\\
&=-\sum_{y\in S}\gamma_{\omega,S}(x,y)B_\sigma(y)\\
&=-\langle\gamma_{\omega,S}(x,\cdot),B_\sigma\rangle_S
\end{aligned}$$

for each $x \in S$. Moreover, a simple calculation shows that every function of the form (3.11) gives a solution.

The desired uniqueness result now follows easily from Theorem 2.1. □

Remark 3.5.

(i) The identity (3.11) can be rewritten as

$$f(x)=\sum_{j=1}^{|S|}\frac{1}{\nu_j}\sum_{y\in S}\left[\sum_{z\in\partial S}\frac{\sigma(z)\omega(y,z)}{d_\omega y}\right]\phi_j(y)\phi_j(x)\sqrt{\frac{d_\omega y}{d_\omega x}},\quad x\in S.$$

In fact, B_σ is a function on S depending only on the value of σ on ∂S and $B_\sigma(y) = 0$ for $y \in S\setminus \overset{\circ}{\partial} S$. On the other hand, two different boundary conditions σ_1 and σ_2 may give rise to the same solution whenever $B_{\sigma_1} = B_{\sigma_2}$.

(ii) (3.11) can be understood as a matrix multiplication by

$$f = -\gamma_{\omega,S} \cdot B_\sigma \text{ on } S \tag{3.14}$$

or, equivalently,

$$\Delta_{\omega,S} f = -B_\sigma \text{ on } S \tag{3.15}$$

in view of (3.8). The relation (3.15) enables us to identify uniquely the boundary values from an ω-harmonic function f with $\Delta_\omega f = 0$ on S.

Now we characterize the ω-harmonic functions with a set of singularities in a subgraph with nonempty boundary.

THEOREM 3.6. *Let S be a subgraph of a graph with weight ω and $T \subset S$. Then every $f : \overline{S} \to \mathbb{C}$ satisfying*

$$\Delta_\omega f(x) = 0, \quad x \in S \setminus T,$$

can be uniquely represented as

$$f(x) = h(x) + \sum_{y \in T} \gamma_{\omega,S}(x,y)\beta(y), \quad x \in \overline{S}, \tag{3.16}$$

where h is an ω-harmonic function on S satisfying $h|_{\partial S} = f|_{\partial S}$ and $\beta(y) = \Delta_\omega f(y)$, $y \in T$.

Proof. The uniqueness is easy, by Theorem 2.1. Now let $\beta(y) := \Delta_\omega f(y), y \in T$. Then we have

$$\Delta_\omega f(x) = \begin{cases} 0, & x \in S \setminus T, \\ \beta(x), & x \in T. \end{cases}$$

Define, for $x \in \overline{S}$,

$$f_1(x) := \sum_{y \in T} \gamma_{\omega,S}(x,y)\beta(y)$$

and

$$h(x) := f(x) - f_1(x).$$

Then $h|_{\partial S} = f|_{\partial S}$ and, for each $x \in S$,

$$\begin{aligned} \Delta_\omega h(x) &= \Delta_\omega f(x) - \Delta_\omega \left[\sum_{y \in T} \sum_{j=1}^{|S|} \left(-\frac{1}{\nu_j} \frac{\phi_j(x)}{\sqrt{d_\omega x}} \cdot \phi_j(y) \sqrt{d_\omega y} \right) \beta(y) \right] \\ &= \Delta_\omega f(x) - \sum_{y \in T} \sum_{j=1}^{|S|} \frac{\phi_j(x)}{\sqrt{d_\omega x}} \cdot \left[\phi_j(y) \sqrt{d_\omega y} \beta(y) \right] \\ &= \Delta_\omega f(x) - \sum_{y \in T} \delta(x,y)\beta(y) \\ &= 0, \end{aligned}$$

which completes the proof. □

Remark 3.7.

(i) In particular, if $T = \{x_0\}$, $x_0 \in S$, then (3.16) can be written simply as

$$f(x) = h(x) + \gamma_{\omega,S}(x, x_0)\beta(x_0),$$

where $\beta(x_0) = \Delta_\omega f(x_0)$.

(ii) In fact, in view of (3.16) and Theorem 3.4, the solution to the nonhomogeneous DBVP

$$\begin{cases} \Delta_\omega f(x) = g(x), & x \in S, \\ f|_{\partial S} = \sigma \end{cases}$$

can be represented by

$$f(x) = -\langle \gamma_{\omega,S}(x, \cdot), B_\sigma \rangle_S + \langle \gamma_{\omega,S}(x, \cdot), g \rangle_S.$$

Now we will discuss the Neumann boundary value problem (NBVP). The solvability of the NBVP and its proof have not been seen yet in any literature, at least to the best of the authors' knowledge.

First, we recall Green's formula

$$\int_S \Delta_\omega f = \int_{\partial S} \frac{\partial f}{\partial_\omega n}.$$

Hence, if there exists a solution to

$$\begin{cases} \Delta_\omega f = g & \text{on } S, \\ \frac{\partial f}{\partial_\omega n} = \psi & \text{on } \partial S, \end{cases}$$

then by Green's formula it is necessary that $\int_S g = \int_{\partial S} \psi$.

THEOREM 3.8. *Let S be a subgraph of a host graph G with a weight ω and let $f : \overline{S} \to \mathbb{R}$, $g : S \to \mathbb{R}$, and $\psi : \partial S \to \mathbb{R}$ be functions with $\int_{\partial S} \psi = \int_S g$. Then the solution to the NBVP*

$$\begin{cases} \Delta_\omega f(x) = g(x), & x \in S, \\ \frac{\partial f}{\partial_\omega n}(z) = \psi(z), & z \in \partial S, \end{cases}$$

is given by

$$f(x) = a_0 + \langle \Gamma_\omega(x, \cdot), g \rangle_S - \langle \Gamma_\omega(x, \cdot), \psi \rangle_{\partial S},$$

where Γ_ω is the Green function of Δ_ω on the graph $\overline{S}$ as a new host graph of S and a_0 is an arbitrary constant.

Proof. We rewrite the NBVP as

$$\begin{cases} \sum_{y \in \overline{S}} \left[f(y) - f(x) \right] \frac{\omega(x,y)}{d_\omega x} = g(x), & x \in S, \\ \sum_{y \in S} \left[f(y) - f(z) \right] \frac{\omega(y,z)}{d'_\omega z} = -\psi(z), & z \in \partial S. \end{cases} \tag{3.17}$$

To solve the system (3.17), consider $\overline{S}$ as a new host graph with the weight ω and with no boundary. Then S is still a subgraph of $\overline{S}$. (In fact, we should note here that if we regard $\overline{S}$ as a subgraph of G, then its boundary $\partial\overline{S}$ may not be empty.) Then, for each $z \in \partial S$, the inner degree $d'_\omega z$ is equal to $d_\omega z$ in this new graph $\overline{S}$, since

the induced subgraph has no edges between the vertices on ∂S. Hence (3.17) can be written as

$$\begin{cases} \sum_{y\in V_0} \big[f(y)-f(x)\big]\frac{\omega(x,y)}{d_\omega x} = g(x), & x\in S,\\ \sum_{y\in V_0} \big[f(y)-f(z)\big]\frac{\omega(y,z)}{d'_\omega z} = -\psi(z), & z\in \partial S, \end{cases} \tag{3.18}$$

where V_0 is the set of vertices in $\overline{S}$. Hence (3.18) is equivalent to

$$\sum_{y\in V_0} \big[f(y)-f(x)\big]\frac{\omega(x,y)}{d_\omega x} = \Psi(x), \quad x\in \overline{S}, \tag{3.19}$$

where

$$\Psi(x) = \begin{cases} g(x), & x\in S,\\ -\psi(x), & x\in \partial S. \end{cases}$$

Therefore, the NBVP is equivalent to

$$\Delta_\omega f(x) = \Psi(x), \quad x\in \overline{S}.$$

Thus, it follows from Theorem 3.1 that

$$\begin{aligned} f(x) &= a_0 + \langle \Gamma_\omega(x,\cdot), \Psi\rangle \\ &= a_0 + \sum_{y\in V_0} \Gamma_\omega(x,y)\Psi(y) \\ &= a_0 + \sum_{y\in S} \Gamma_\omega(x,y)g(y) - \sum_{z\in \partial S} \Gamma_\omega(x,z)\psi(z) \\ &= a_0 + \langle \Gamma_\omega(x,\cdot), g\rangle_S - \langle \Gamma_\omega(x,\cdot), \psi\rangle_{\partial S}, \end{aligned}$$

where a_0 is an arbitrary constant. This completes the proof. □

Remark 3.9. The solution to the NBVP is uniquely determined by the Neumann data ψ on ∂S up to an additive constant. Thus, we get a unique solution if we prescribe the value of f at some vertex in S or, for example, if we seek the solution f with $\int_S f =$ (a given constant).

4. Inverse problems. In the previous section, we have seen that for a function $\psi : \partial S \to \mathbb{R}$ with $\int_{\partial S} \psi = 0$ the NBVP

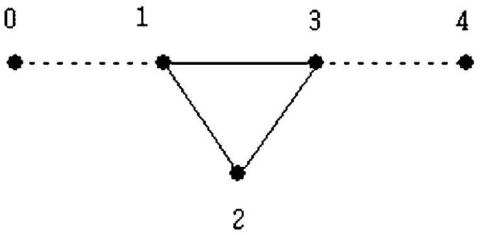
$$(NBVP) \begin{cases} \Delta_\omega f(x) = 0, & x\in S,\\ \frac{\partial f}{\partial_\omega n}(z) = \psi(z), & z\in \partial S, \end{cases}$$

has a unique solution up to an additive constant. Therefore, the Dirichlet data $f|_{\partial S}$ is well defined up to an additive constant.

In this section, we will discuss the inverse conductivity problem on the network (graph) S with nonempty boundary, which consists in recovering the conductivity (connectivity or weight) ω of the graph by using the so-called input-output map, for example, by using the Dirichlet data induced by the Neumann data (Neumann-to-Dirichlet map), with one boundary measurement.

In order to deal with this inverse problem , we need at least to know or be given the boundary data such as $f(z)$, $\frac{\partial f}{\partial_\omega n}(z)$ for $z\in\partial S$ and ω near the boundary. So it

is natural to assume that $f|_{\partial S}$, $\frac{\partial f}{\partial_\omega n}|_{\partial S}$, and $\omega|_{\partial S \times \overset{\circ}{\partial} S}$ are known (given or measured). But even though we are given all these data on the boundary, we are not guaranteed, in general, to be able to identify the conductivity ω uniquely. To illustrate this we consider a graph S whose vertices are $\{1, 2, 3\}$ and $\partial S = \{0, 4\}$ as follows:



with the weight

$$\omega(0,1) = 1,\ \omega(0,k) = 0 \quad (k = 2,\ 3,\ 4),$$

and

$$\omega(3,4) = 1,\ \omega(k,4) = 0 \quad (k = 0,\ 1,\ 2).$$

Let $f : \overline{S} \to \mathbb{R}$ be a function satisfying $\Delta_\omega f(k) = 0,\ k = 1, 2, 3$. Assume that

$$f(0) = 0,\ f(1) = 1,\ f(3) = 3,\ f(4) = 4,\ f(2) = (\text{unknown}).$$

Thus, since $\overset{\circ}{\partial} S = \{1, 3\}$, the boundary data $f|_{\partial S}$, $\frac{\partial f}{\partial_\omega n}|_{\partial S}$ and $\omega|_{\partial S \times \overset{\circ}{\partial} S}$ are known.

In fact,

$$\frac{\partial f}{\partial_\omega n}(0) = f(0) - f(1) = -1,$$
$$\frac{\partial f}{\partial_\omega n}(4) = f(4) - f(3) = 1.$$

The problem is to determine

$$\omega(1,2) = x,\ \ \omega(2,3) = y,\ \ \omega(1,3) = z, \quad \text{and} \quad f(2).$$

From $\Delta_\omega f(k) = 0,\ k = 1, 2, 3$, we have

$$f(1) = \frac{f(0) + xf(2) + zf(3)}{1 + x + z} = 1,$$
$$f(2) = \frac{xf(1) + yf(3)}{x + y},$$
$$f(3) = \frac{zf(1) + yf(2) + f(4)}{z + y + 1} = 3.$$

This system is equivalent to

$$\begin{cases} x(y-1) + y(x-1) + 2z(x+y) = 0, \\ f(2) = \frac{x+3y}{x+y}. \end{cases} \tag{4.1}$$

This system has infinitely many solutions. For instance, assume $z = 0$; that is, the two vertices 1 and 3 are not adjacent. Then (4.1) is reduced to

$$\begin{cases} \frac{1}{x}+\frac{1}{y}=2, \\ f(2)=\frac{x+3y}{x+y}. \end{cases} \tag{4.2}$$

For example, $(x,y,z)=(1,1,0)$ or $(2,2/3,0)$ satisfy the first equation. In fact, it is easy to see that there are infinitely many pairs (x,y) of nonnegative numbers satisfying the first equation in (4.2) so that $f(2)$ is undetermined as a result.

In view of the above example, in order to determine the weight ω uniquely we need some more information other than $f|_{\partial S}$, $\frac{\partial f}{\partial_\omega n}|_{\partial S}$, and $\omega|_{\partial S\times\mathring{\partial}S}$. To motivate the main theorem we impose in this example the additional constraints that

$$x\geq 1,\ y\geq 1,\ \text{and } z\geq 0 \tag{4.3}$$

in (4.1). Then (4.1) yields a unique triple of solution $x=1,\ y=1,\ z=0$, and $f(2)=2$.

As a matter of fact, even the inverse conductivity problem of the diffusion equation

$$P[a;u]:=\begin{cases} \operatorname{div}[a(x)\nabla u(x)]=0, & x\in\Omega, \\ u|_{\partial\Omega}=\sigma \end{cases} \tag{4.4}$$

in a bounded open subset $\Omega\subset\mathbb{R}^n$ has been studied under some additional constraints besides Dirichlet and Neumann data (see [A], [BF], [Ca], [I], [IP], and [SU]). In particular, in [A] and [I] it is shown that there is a global uniqueness result under the condition that

(i) $a_1=a_2$ near $\partial\Omega$, and $a_1\leq a_2$ in Ω,
(ii) $\frac{\partial u_1}{\partial n}=\frac{\partial u_2}{\partial n}$ on $\partial\Omega$,
(iii) $\int_\Omega u_1=\int_\Omega u_2=0$,

where $P[a_j;u_j]=0,\ j=1,2$ in (4.4).

Now we are in a position to state the first main theorem of this paper.

THEOREM 4.1. *Let ω_1 and ω_2 be weights with $\omega_1\leq\omega_2$ on $\overline{S}\times\overline{S}$ and $f_1,\ f_2:\overline{S}\to\mathbb{R}$ be functions satisfying that*

$$\begin{cases} \Delta_{\omega_j} f_j(x)=0, & x\in S, \\ \frac{\partial f}{\partial_{\omega_j} n}(z)=\psi(z), & z\in\partial S, \end{cases}$$

for a given function $\psi:\partial S\to\mathbb{R}$ with $\int_{\partial S}\psi=0$ and $j=1,\ 2$.

If we assume that

(i) $\omega_1(z,y)=\omega_2(z,y)$ *on* $\partial S\times\mathring{\partial}S$,
(ii) $f_1|_{\partial S}=f_2|_{\partial S}$,

then we have

(i) $f_1=f_2$ *on* $\overline{S}$,
(ii) $\omega_1(x,y)=\omega_2(x,y)$ *whenever* $f_1(x)\neq f_1(y)$, *or* $f_2(x)\neq f_2(y)$.

To prove this result we adapt the method of energy functionals, extensively used for theory of nonlinear partial differential equations. For functions $\sigma:\partial S\to\mathbb{R}$ and $g:\overline{S}\to\mathbb{R}$ we define a functional by

$$I_\omega[h]:=\int_{\overline{S}}\left[\frac{1}{4}|\nabla_\omega h|^2-hg\right] \tag{4.5}$$

for every function h in the set

$$A := \{h : \overline{S} \to \mathbb{R} | \quad h|_{\partial S} = \sigma\}, \tag{4.6}$$

which is called the admissible set. In the continuous case, the well-known Dirichlet principle states that the energy minimizer in the admissible set is a solution of the DBVP. We derive here the discrete version of Dirichlet's principle as follows.

THEOREM 4.2 (Dirichlet's principle). *Assume that $f : \overline{S} \to \mathbb{R}$ is a solution to*

$$\begin{cases} -\Delta_\omega f = g & \text{on } S, \\ f|_{\partial S} = \sigma. \end{cases} \tag{4.7}$$

Then

$$I_\omega[f] = \min_{h \in A} I_\omega[h]. \tag{4.8}$$

Conversely, if $f \in A$ satisfies (4.8)*, then f is the solution of* (4.7)*, and the only one.*

Proof. Let h be a function in A. Then, making use of (1.2) in Theorem 1.2, we have

$$\begin{aligned} 0 &= \int_{\overline{S}} (-\Delta_\omega f - g)(f - h) \\ &= \int_{\overline{S}} \big[(-\Delta_\omega f)(f - h) - g(f - h)\big] \\ &= \int_{\overline{S}} \left[\frac{1}{2}\nabla_\omega f \cdot \nabla_\omega (f - h) - g(f - h)\right] \\ &= \frac{1}{2}\int_{\overline{S}} |\nabla_\omega f|^2 - \frac{1}{2}\int_{\overline{S}} \nabla_\omega f \cdot \nabla_\omega h - \int_{\overline{S}} g(f - h). \end{aligned}$$

Hence

$$\begin{aligned} \int_{\overline{S}} \left[\frac{1}{2}|\nabla_\omega f|^2 - gf\right] &= \int_{\overline{S}} \left[\frac{1}{2}\nabla_\omega f \cdot \nabla_\omega h - gh\right] \\ &= \frac{1}{2}\sum_{x \in \overline{S}} \sum_{y \in \overline{S}} |\big[f(y) - f(x)\big] \cdot \big[h(y) - h(x)\big]| \cdot \omega(x, y) - \int_{\overline{S}} gh \\ &\le \frac{1}{2}\sum_{x \in \overline{S}} \sum_{y \in \overline{S}} \frac{\big[f(y) - f(x)\big]^2 + \big[h(y) - h(x)\big]^2}{2} \cdot \omega(x, y) - \int_{\overline{S}} gh \\ &= \frac{1}{4}\int_{\overline{S}} |\nabla_\omega f|^2 + \frac{1}{4}\int_{\overline{S}} |\nabla_\omega h|^2 - \int_{\overline{S}} gh, \end{aligned}$$

where we used the triangular inequality

$$|ab| \le \frac{a^2 + b^2}{2}, \quad a, b \in \mathbb{R}.$$

Thus, it follows that

$$\int_{\overline{S}} \left[\frac{1}{4}|\nabla_\omega f|^2 - gf\right] \le \int_{\overline{S}} \left[\frac{1}{4}|\nabla_\omega h|^2 - gh\right],$$

which implies

$$I_\omega[f] \le I_\omega[h], \quad h \in A.$$

Since $f \in A$, we have

$$\min_{h \in A} I_\omega[h] = I_\omega[f].$$

Now we prove the converse. Let T be a subset of vertices in S and

$$\chi_T(x) = \begin{cases} 1, & x \in T, \\ 0 & \text{otherwise.} \end{cases}$$

Then $f + \tau\chi_T \in A$ for each real number τ, since $\chi_T = 0$ on ∂S. Define

$$i(\tau) := I_\omega[f + \tau\chi_T], \quad \tau \in \mathbb{R}.$$

Then

$$\begin{aligned} i(\tau) &= \int_{\overline{S}} \left[\frac{1}{4} |\nabla_\omega f + \tau \nabla_\omega \chi_T|^2 - (f + \tau\chi_T) g \right] \\ &= \frac{1}{4} \int_{\overline{S}} |\nabla_\omega f|^2 + 2\tau \nabla_\omega f \cdot \nabla_\omega \chi_T + \tau^2 |\nabla_\omega \chi_T|^2 - \int_{\overline{S}} (f + \tau\chi_T) g. \end{aligned}$$

Note that the scalar function $i(\tau)$ has a minimum at $\tau = 0$ and thus $\frac{di}{d\tau}(0) = 0$. That is,

$$\begin{aligned} 0 &= \frac{1}{2} \int_{\overline{S}} \nabla_\omega f \cdot \nabla_\omega \chi_T - \int_{\overline{S}} \chi_T \cdot g \\ &= \int_{\overline{S}} \left[\chi_T (-\Delta_\omega f - g) \right] \\ &= \sum_{x \in T} \left[-\Delta_\omega f(x) - g(x) \right] d_\omega x. \end{aligned}$$

In particular, taking $T = \{x\}, x \in S$, we obtain

$$-\Delta_\omega f(x) - g(x) = 0,$$

which is the required result. The uniqueness follows from Theorem 3.4. □

Now we are ready to prove Theorem 4.1.

Proof of Theorem 4.1.

(i) Let $\sigma : \partial S \to \mathbb{R}$ be the function defined by

$$\sigma(z) = f_1(z) = f_2(z), \quad z \in \partial S,$$

using the hypothesis (ii). Define

$$I_{\omega_1}[h] := \frac{1}{4} \int_{\overline{S}} |\nabla_{\omega_1} h|^2 d_{\omega_1}$$

for every h in the admissible set

$$A = \left\{ h : \overline{S} \to \mathbb{R} | \quad h|_{\partial S} = \sigma \right\}.$$

Then, by virtue of Theorem 1.2 we have

$$\begin{aligned} I_{\omega_1}[h] &= \frac{1}{2} \int_{\overline{S}} h(-\Delta_{\omega_1} h) d_{\omega_1} \\ &= \frac{1}{2} \int_S h(-\Delta_{\omega_1} h) d_{\omega_1} + \frac{1}{2} \int_{\partial S} h(-\Delta_{\omega_1} h) d_{\omega_1}. \end{aligned}$$

Moreover, by the coincidence of the Dirichlet and Neumann data we can see that the boundary ∂S and the inner boundary $\mathring{\partial} S$ are well defined independently of the values of the weights $\omega_1,\ \omega_2$ and, moreover, for $z \in \partial S$,

$$d_{\omega_1} z = \sum_{y \in \mathring{\partial} S} \omega_1(z,y) = \sum_{y \in \mathring{\partial} S} \omega_2(z,y) = d_{\omega_2} z, \tag{4.9}$$

$$\begin{aligned} \Delta_{\omega_1} f_1(z) &= \sum_{y \in \mathring{\partial} S} \big[f_1(y) - f_1(z)\big] \frac{\omega_1(z,y)}{d_{\omega_1} z} \\ &= \sum_{y \in \mathring{\partial} S} \big[f_2(y) - f_2(z)\big] \frac{\omega_2(z,y)}{d_{\omega_2} z} \\ &= \Delta_{\omega_2} f_2(z). \end{aligned} \tag{4.10}$$

Then, it follows from the condition $\omega_1 \le \omega_2$ that

$$\begin{aligned} I_{\omega_1}[f_1] &= \frac{1}{2} \int_{\partial S} f_1(-\Delta_{\omega_1} f_1) d_{\omega_1} \\ &= \frac{1}{2} \int_{\partial S} f_2(-\Delta_{\omega_2} f_2) d_{\omega_1} \\ &= \frac{1}{2} \int_{S} f_2(-\Delta_{\omega_2} f_2) d_{\omega_2} + \frac{1}{2} \int_{\partial S} f_2(-\Delta_{\omega_2} f_2) d_{\omega_2} \\ &= \frac{1}{2} \int_{\overline{S}} f_2(-\Delta_{\omega_2} f_2) d_{\omega_2} \\ &= \frac{1}{4} \int_{\overline{S}} |\nabla_{\omega_2} f_2|^2 d_{\omega_2} \\ &= \frac{1}{4} \sum_{x \in \overline{S}} \sum_{y \in \overline{S}} \big[f_2(x) - f_2(y)\big]^2 \omega_2(x,y) \\ &\ge \frac{1}{4} \sum_{x \in \overline{S}} \sum_{y \in \overline{S}} \big[f_2(x) - f_2(y)\big]^2 \omega_1(x,y) \\ &= \frac{1}{4} \int_{\overline{S}} |\nabla_{\omega_1} f_2|^2 d_{\omega_1} \\ &= I_{\omega_1}[f_2]. \end{aligned}$$

Using Dirichlet's principle (Theorem 4.2), one sees that $f_1 = f_2$ on $\overline{S}$.

(ii) In the proof of (i) we actually have proved that $I_{\omega_1}[f_1] = I_{\omega_1}[f_2]$. In other words, taking $f := f_1 = f_2$ on $\overline{S}$,

$$\sum_{x \in \overline{S}} \sum_{y \in \overline{S}} \big[f(x) - f(y)\big]^2 \omega_2(x,y) = \sum_{x \in \overline{S}} \sum_{y \in \overline{S}} \big[f(x) - f(y)\big]^2 \omega_1(x,y),$$

or, equivalently,

$$\sum_{x \in \overline{S}} \sum_{y \in \overline{S}} \big[f(x) - f(y)\big]^2 \cdot \big[\omega_2(x,y) - \omega_1(x,y)\big] = 0.$$

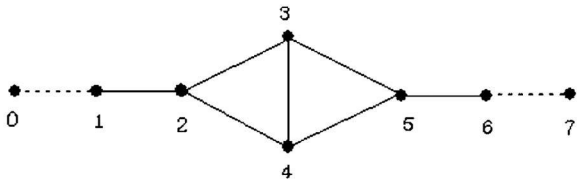
Therefore, we have

$$\big[f(x)-f(y)\big]^2\cdot\big[\omega_2(x,y)-\omega_1(x,y)\big]=0$$

for all $x\in\overline{S}$ and $y\in\overline{S}$. This gives (ii). □

Remark 4.3. In Theorem 4.1 above, if $f:=f_1=f_2$ is injective on S, then we are able to get $\omega_1=\omega_2$ on $\overline{S}\times\overline{S}$. For example, if S is the path P_n on n vertices with arbitrary weight ω, then it is not hard to see that every nonconstant ω-harmonic function f on P_n is strictly monotonic and hence all the weights are identified. But, in general, most graphs, even with the standard weight do not admit an injective solution to the DBVP or NBVP. Therefore, it will be quite interesting to figure out a pair of graphs and weights which admits an injective solution to the DBVP or NBVP.

To develop an idea to improve Theorem 4.1 we consider a graph $S=\{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$ with $\partial S=\{0,\ 7\}$ as follows:



Suppose that ω_1 is the standard weight and ω_2 is the weight given by $\omega_1=\omega_2$ except only $\omega_2(3,\ 4)=k,\ k\geq 1$. Then $\omega_1\leq\omega_2$ throughout the graph $\overline{S}$ and $\omega_1=\omega_2$ except on the edge $\{3,4\}$. Now define a function $f:\overline{S}\to\mathbb{R}$ as

$$f(0)=a,\ f(1)=a-\alpha,\ f(2)=a-2\alpha,\ f(3)=f(4)=\frac{2a-5\alpha}{2},$$

$$f(5)=a-3\alpha,\ f(6)=a-4\alpha,\ f(7)=a-5\alpha,$$

where a and α are arbitrary real numbers. Then it is easy to verify that f satisfies both the equations

$$\Delta_{\omega_1}f(x)=0=\Delta_{\omega_2}f(x),\quad x\in S.$$

Here, we note that f is uniquely determined by the Dirichlet data $f(0)=a,\ f(7)=a-5\alpha$ and the Neumann data

$$\frac{\partial f}{\partial_\omega n}(0)=f(0)-f(1)=\alpha,\quad \frac{\partial f}{\partial_\omega n}(7)=f(7)-f(6)=-\alpha,$$

and each value $f(x)$ is determined regardless of the weight $\omega_2(3,4)=k$. This implies that we cannot identify the weight $\omega_2(3,\ 4)=k$ even with all possible boundary data. To derive a key idea to overcome this difficulty, we take $a>0$ and α so that $f(0)>0$ and $f(7)>0$. By a direct calculation (or using Corollary 2.2) we see that

$$f(m)>0,\ m=0,1,2,\dots,7.$$

Suppose that f satisfies the relation

$$\int_S f d_{\omega_1}=\int_S f d_{\omega_2}. \tag{4.11}$$

Then, since

$$\int_S f d_{\omega_1} = 2f(1) + 3f(2) + 3f(3) + 3f(4) + 3f(5) + 2f(6)$$

and

$$\int_S f d_{\omega_2} = 2f(1) + 3f(2) + (2+k)f(3) + (2+k)f(4) + 3f(5) + 2f(6),$$

it follows that

$$k\big[f(3) + f(4)\big] = f(3) + f(4),$$

which gives $k = 1$. Therefore, in order to identify the weight over all edges we need to impose an additional condition such as (4.11).

Now we return to the general situation. We know that for a function $\psi : \partial S \to \mathbb{R}$ with $\int_{\partial S} \psi = 0$ and $j = 1,\ 2$, the equation

$$\begin{cases} \Delta_{\omega_j} h_j(x) = 0,\ x \in S, \\ \frac{\partial h}{\partial_{\omega_j} n}(z) = \psi(z),\ z \in \partial S, \\ \int_S h_j\ d_{\omega_j} = 0 \end{cases} \tag{4.12}$$

has a unique pair of solutions (h_1, h_2). Let

$$m_j = \min_{z \in \partial S}\ h_j(z), \quad j = 1, 2, \tag{4.13}$$

and

$$m_0 = \max_{j=1,2}\ |m_j| \cdot \mathrm{vol}(S, \omega_j), \tag{4.14}$$

where $\mathrm{vol}(S, \omega_j) = \sum_{x \in S} d_{\omega_j} x$.

Motivated by the above example, we refine Theorem 4.1 as follows.

THEOREM 4.4. *Let ω_1 and ω_2 be weights with $\omega_1 \le \omega_2$ on $\overline{S} \times \overline{S}$ and $f_1,\ f_2 : \overline{S} \to \mathbb{R}$ be functions satisfying that for each $j = 1,\ 2$,*

$$\begin{cases} \Delta_{\omega_j} f_j(x) = 0,\ x \in S, \\ \frac{\partial f}{\partial_{\omega_j} n}(z) = \psi(z),\ z \in \partial S, \\ \int_S f_j\ d_{\omega_j} = K \end{cases} \tag{4.15}$$

for a given function $\psi : \partial S \to \mathbb{R}$ with $\int_{\partial S} \psi = 0$ and a given constant K with $K > m_0$. (Here, m_0 is the constant in (4.14)*.)*

If we assume that

(i) $\omega_1(z, y) = \omega_2(z, y)$ *on* $\partial S \times \overset{\circ}{\partial} S$,

(ii) $f_1|_{\partial S} = f_2|_{\partial S}$,

then we have

$$f_1 \equiv f_2$$

and

$$\omega_1(x, y) = \omega_2(x, y)$$

for all x and y in $\overline{S}$.

Proof. We have already shown in Theorem 4.1 that $f_1 \equiv f_2$. Now, for each $j = 1, 2$, we choose a constant C_j so that $C_j \cdot \mathrm{vol}(S, \omega_j) = K$. Then, it follows that $C_j > |m_j|$ and hence $h_j(x) + C_j > 0,\ x \in S$, by the maximum principle (or Corollary 2.2). Moreover, the function $\tilde{h}(x) := h_j(x) + C_j$ satisfies (4.15). By the uniqueness of the solution we have

$$f_j(x) = \tilde{h}(x) = h_j(x) + C_j > 0,\ x \in S.$$

Let $f := f_1 = f_2$ on $\overline{S}$. Then it follows from the condition $\int_S f_1\ d_{\omega_1} = K = \int_S f_2\ d_{\omega_2}$ that

$$\sum_{x \in S} f(x) d_{\omega_1}(x) = \sum_{x \in S} f(x) d_{\omega_2}(x)$$

or, equivalently,

$$\sum_{x \in S} f(x) \big[d_{\omega_2}(x) - d_{\omega_1}(x) \big] = 0.$$

Since $f(x) > 0$ and $d_{\omega_1}(x) \geq d_{\omega_1}(x)$ for all $x \in S$, we have

$$\begin{aligned} 0 &= d_{\omega_2}(x) - d_{\omega_1}(x) \\ &= \sum_{y \in \overline{S}} \big[\omega_2(x, y) - \omega_2(x, y) \big]. \end{aligned}$$

Since $\omega_1(x, y) \leq \omega_2(x, y)$, we obtain

$$\omega_1(x, y) = \omega_2(x, y)$$

for all x and y in $\overline{S}$, as required. □

Remark 4.5. In the above proof, the condition $K > m_0$ was used only to guarantee that $f_j(x) > 0,\ x \in S$. Hence, if we replace this condition by $f|_{\partial S} > 0,\ j = 1, 2$, in Theorem 4.4, we arrive at the same conclusion. Practically the positive solutions are easily available by adjusting the boundary values.

As seen in the study of the inverse conductivity problem in the continuous case (see, for instance, [A], [Ca], [I], [IP], [SU]), it would be worthwhile to prove the uniqueness under a condition weaker than the monotonicity condition $\omega_1 \leq \omega_2$ imposed above. Moreover, it would also be interesting to consider a stability theorem for the same conductivity equation.

Acknowledgments. We take the opportunity to thank David Walnut and the anonymous referees for their thoughtful comments.

REFERENCES

[A] G. Alessandrini, *Remark on a paper by H. Bellout and A. Friedman*, Boll. Un. Mat. Ital. A(7), 3 (1989), pp. 243–249.

[AJB1] R. Albert, H. Jeong, and A.-L. Barabasi, *Internet: Diameter of the world-wide web*, Nature, 401 (1999), pp. 130–131.

[AJB2] R. Albert, H. Jeong, and A.-L. Barabasi, *The internet Archilles' heel: Error and attack tolerance of complex networks*, Nature, 406 (2000), pp. 378–382.

[B1] N. L. Biggs, *Algebraic Graph Theory*, 2nd ed., Cambridge University Press, Cambridge, UK, 1993.

[B2] N. L. Biggs, *Potential theory on distance-regular graphs*, Combin. Probab. Comput., 2 (1993), pp. 243–255.

[B3] N. L. Biggs, *Algebraic graph theory on graphs*, Bull. London Math. Soc., 29 (1997), pp. 641–682.

[BCE1] E. Bendito, A. Carmona, and A. M. Encinas, *Solving boundary value problems on networks using equilibrium measures*, J. Funct. Anal., 171 (2000), pp. 155–176.

[BCE2] E. Bendito, A. Carmona, and A. M. Encinas, *Solving Dirichlet and Poisson problems on graphs by means of equilibrium measures*, European J. Combin., 24 (2003), pp. 365–375.

[BF] H. Bellout and A. Friedman, *Identification problems in potential theory*, Arch. Rational Mech. Anal., 101 (1988), pp. 143–160.

[BH] A. Bondy and R. L. Hemminger, *Graph reconstruction—a survey*, J. Graph Theory, 1 (1977), pp. 227–168.

[Ca] A. P. Calderon, *On an inverse boundary value problem*, in Proceedings of the Seminar on Numerical Analysis and Its Applications to Continuum Physics, Soc. Braz. Math., Rio de Janeiro, Brazil, 1980, pp. 65–73.

[CGGS] F. R. K. Chung, M. Garrett, R. Graham, and D. Shallcross, *Distance realization problems with applications to internet tomography*, J. Comput. System Sci., 63 (2001), pp. 432–448.

[Ch] F. R. K. Chung, *Spectral Graph Theory*, CBMS Reg. Conf. Ser. Math. 92, AMS, Providence, RI, 1997.

[CL] F. R. K. Chung and R. P. Langlands, *A combinatorial Laplacian with vertex weights*, J. Combin. Theory Ser. A, 75 (1996), pp. 316–327.

[ClMM] K. C. Claffy, T. E. Monk, and D. McRobb, *Internet tomography*, Nature, Web Matters, http://helix.nature.com/webmatters/tomog.html, Jan. 1999.

[CM1] E. B. Curtis and J. A. Morrow, *Determining the resistors in a network*, SIAM J. Appl. Math., 50 (1990), pp. 918–930.

[CM2] E. B. Curtis and J. A. Morrow, *The Dirichlet to Neumann map for a resistor network*, SIAM J. Appl. Math., 51 (1991), pp. 1011–1029.

[CO] F. R. K. Chung and K. Oden, *Weighted graph Laplacians and isoperimetric inequalities*, Pacific J. Math., 192 (2000), pp. 257–274.

[CvDGT] D. M. Cvetkovic, M. Doob, I. Gutman, and A. Torgasev, *Recent Results in the Theory of Graph Spectra*, Annals of Discrete Mathematics 36, North–Holland, Amsterdam, 1988.

[CvDS] D. M. Cvetkovic, M. Doob, and H. Sachs, *Spectra of Graphs, Theory and Applications*, Academic Press, New York, 1980.

[CY] F. R. K. Chung and S.-T. Yau, *Discrete Green's functions*, J. Combin. Theory Ser. A, 91 (2000), pp. 191–214.

[HY] S. L. Hakimi and S. S. Yau, *Distance matrix of a graph and its realizability*, Quart. Appl. Math., 22 (1965), pp. 305–317.

[I] V. Isakov, *Inverse Problem for Partial Differential Equations*, Springer-Verlag, New York, 1998.

[IM] D. Ingerman and J. A. Morrow, *On a characterization of the kernel of the Dirichlet-to-Neumann map for a planar region*, SIAM J. Math. Anal., 29 (1998), pp. 106–115.

[IP] V. Isakov and J. Powell, *On the inverse conductivity problem with one measurement*, Inverse Problems, 6 (1990), pp. 311–318.

[KS] H. Kang and J. K. Seo, *Recent progress in the inverse conductivity problem with a single measurement*, in Inverse Problems and Related Topics (Kobe, 1998), Chapman & Hall/CRC Res. Notes Math. 419, Chapman & Hall/CRC, Boca Raton, FL, 2000, pp. 69–80.

[MIC] J. A. Morrow, D. Ingerman, and E. B. Curtis, *Circula planar graphs and resistor networks*, Linear Algebra Appl., 283 (1998), pp. 115–150.

[MMC] J. A. Morrow, E. Mooers, and E. B. Curtis, *Finding the conductors in circular networks from boundary measurements*, RAIRO Model. Math. Anal. Numer., 28 (1994), pp. 781–814.

[SU] J. Sylvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary value problem*, Ann. of Math. (2), 125 (1987), pp. 153–169.

SIAM J. APPL. MATH.
Vol. 65, No. 4, pp. 1227–1260

DISTANCE FUNCTIONS AND GEODESICS ON SUBMANIFOLDS OF $\mathbb{R}^d$ AND POINT CLOUDS*

FACUNDO MÉMOLI† AND GUILLERMO SAPIRO†

Abstract. A theoretical and computational framework for computing intrinsic distance functions and geodesics on submanifolds of $\mathbb{R}^d$ given by point clouds is introduced and developed in this paper. The basic idea is that, as shown here, intrinsic distance functions and geodesics on general co-dimension submanifolds of $\mathbb{R}^d$ can be accurately approximated by extrinsic Euclidean ones computed inside a thin offset band surrounding the manifold. This permits the use of computationally optimal algorithms for computing distance functions in Cartesian grids. We use these algorithms, modified to deal with spaces with boundaries, and obtain a computationally optimal approach also for the case of intrinsic distance functions on submanifolds of $\mathbb{R}^d$. For point clouds, the offset band is constructed without the need to explicitly find the underlying manifold, thereby computing intrinsic distance functions and geodesics on point clouds while skipping the manifold reconstruction step. The case of point clouds representing noisy samples of a submanifold of Euclidean space is studied as well. All the underlying theoretical results are presented along with experimental examples for diverse applications and comparisons to graph-based distance algorithms.

Key words. geodesic distance, point clouds, manifolds, high dimensions, eikonal equations, random coverings, fast marching

AMS subject classifications. 65D18, 57A07, 68U05, 60D05

DOI. 10.1137/S003613990342877X

1. Introduction.

One of the most popular sources of point clouds are three-dimensional (3D) shape acquisition devices, such as laser range scanners, with applications in geoscience, art (e.g., archival study), medicine (e.g., prosthetics), manufacturing (from cars to clothes), and security (e.g., recognition), among other disciplines. These scanners generally provide raw data in the form of (noisy) unorganized point clouds representing surface samples, and often produce very large numbers of points (tens of millions, for example, for the *David* model used in this paper). With the increasing popularity and very broad applications of this source of data, it is natural and important to work directly with such representations, without having to go through the intermediate step of fitting a surface to each (a step that can add computational complexity and introduce errors). See, for example, [11, 18, 20, 29, 33, 45, 46, 56, 58] for a few recent works with this type of data. Note that point clouds can also be used as primitives for visualization (e.g., [12, 33, 59]), as well as for editing [72].

Another important field where point clouds are found is in the representation of high-dimensional manifolds by samples (see, for example, [36, 44, 67]). This type of high-dimensional and general codimensional data appears in almost all disciplines, from computational biology to image analysis and financial data. Due to the extremely high number of dimensions in this case, it is impossible to perform manifold reconstruction, and the work needs to be done directly on the raw data, meaning the

*Received by the editors May 28, 2003; accepted for publication (in revised form) August 25, 2004, published electronically April 14, 2005. This work was supported by Office of Naval Research grant ONR-N00014-97-1-0509, the Presidential Early Career Award for Scientists and Engineers (PECASE), and a National Science Foundation CAREER Award.

http://www.siam.org/journals/siap/65-4/42877.html

†Electrical and Computer Engineering Department, University of Minnesota, Minneapolis, MN 55455, and Instituto de Ingeniería Eléctrica, Universidad de la República, Montevideo, Uruguay (memoli@ece.umn.edu, guille@ece.umn.edu). The research of the first author is also supported by CSIC-Uruguay.

point cloud. Also in this area, large amounts of data are becoming available, from neuroscience experiments with neural recording of millions of points to large image and protein databases.

Note that in general a point cloud representation is codimension free, in contrast with other popular representations such as triangular meshes. Some operations, such as the union of point clouds acquired from multiple views, are much easier when performed directly on the representations than when performed on the triangular meshes obtained from them. This paper addresses one of the most fundamental operations in the study and processing of submanifolds of Euclidean space, the computation of intrinsic distance functions and geodesics. We show that these computations can be made by working directly with the point cloud, without the need for reconstructing the underlying manifold. Even if possible (for example, at low dimensions), the meshing operation is avoided, saving computations and improving accuracy. The distance computation itself is performed in computationally optimal time. We present the corresponding theoretical results, experimental examples, and basic comparisons to mesh-based distance algorithms.[1] The results are valid for general dimensions and codimensions, and for (underlying) manifolds with or without boundary. These results include the analysis of noisy point clouds obtained from sampling the manifold. We provide bounds on the accuracy of the computations that depend on the sampling rate and pattern as well as on the noise, thereby addressing real manifold sampling scenarios.

A number of key building blocks are part of the framework introduced here. The first one is based on the fact that distance functions intrinsic to a given submanifold of $\mathbb{R}^d$ can be accurately approximated by Euclidean distance functions computed in a thin offset band that surrounds this manifold. This concept was first introduced in [49], where convergence results were given for hypersurfaces (codimension one submanifolds of $\mathbb{R}^d$) without boundary. This result is reviewed in section 2. In this paper, we first extend these results to general codimensions and deal with manifolds with or without boundary in section 3. Interestingly, we also show that the approximation is true not only for the intrinsic distance function but also for the intrinsic minimizing geodesic.

The approximation of intrinsic distance functions (and geodesics) by extrinsic Euclidean ones permits us to compute them using computationally optimal algorithms in Cartesian grids (as long as the discretization operation is permitted, memorywise;[2] see sections 7.1 and 8). These algorithms are based on the fact that the distance function satisfies a Hamilton–Jacobi partial differential equation (see section 2), for which consistent and fast algorithms have been developed in Cartesian grids [35, 62, 63, 69].[3] (See [40] for extensions to triangular meshes, and [68] for other Hamilton–Jacobi equations.) That is, due to these results, we can use computationally optimal algorithms in Cartesian grids (with boundaries) also to compute distance functions, and from them geodesics,[4] intrinsic to a given manifold, and in a computationally

[1] Theoretical results on the accuracy of the technique for 3D mesh-based computationally optimal distance computation proposed in [40] have not been reported to the best of our knowledge.

[2] This is of course just a limitation of a straightforward implementation that doesn't avoid allocating memory to empty grids and works in the embedding dimension, and not a limitation of the theoretical and computational frameworks here developed.

[3] Tsitsiklis first described an optimal-control type of approach to solving the Hamilton–Jacobi equation, while independently Sethian and Helmsen both developed techniques based on upwind numerical schemes.

[4] Geodesics are the integral curves corresponding to the gradient directions of the intrinsic distance function, and are obtained by back-propagating in this gradient direction from the target point to the source point.

optimal fashion. Note that, in contrast with the popular Dijkstra algorithm, these numerical techniques are consistent; they converge to the true distance when the grid is refined. Dijkstra's algorithm suffers from digitization bias due to metrication error when implemented on a grid (if no new graph edges are added to account for the new diagonals in each successive level of refinement of the grid); see [52, 53].

Once these basic results are available, we can then move on and deal with point clouds. The basic idea here is to construct the offset band directly from the point cloud, without the intermediate step of manifold reconstruction.[5] This is addressed in section 4 and section 5 for noise-free points and manifold samples, and in section 6 for points considered to be noisy samples of the manifold. In these cases, we explicitly compute the probability that the constructed offset band contains the underlying manifold. As we expect, this probability is a function of the number of point samples, the noise level, the size of the offset, and the basic geometric characteristics of the underlying manifold. This then covers the most realistic scenario, where the manifold is randomly sampled and the samples contain noise, thereby providing bounds that relate the error to the quality of the data. In the experimental section, section 7, we present a number of important applications. These applications are given to show the importance of this novel computational framework, and are by no means exhaustive. The data used in these examples were obtained from real acquisition devices, following laser scanning and photometric stereo. Concluding remarks are presented in section 8, where we also report the directions our research is taking.

To conclude this introduction, we should note that, to the best of our knowledge, the only additional work explicitly addressing the computation of distance functions and geodesics for point clouds is the one reported in [9, 67].[6] The comparison of performance in the presence of noise for our framework and the one proposed in [9, 67] is deferred to Appendix A.[7]

2. Preliminary results and notation. In this section we briefly review the main results in [49], where the idea of approximating intrinsic distances and geodesics by extrinsic ones was first introduced.

2.1. Notation. First, we introduce some basic notation that will be used throughout the article. For a compact and connected set $\Omega \in \mathbb{R}^d$, $d_\Omega(\cdot,\cdot)$ denotes the intrinsic distance between any two points of Ω, measured by paths constrained to be in Ω. We will also assume the convention that if $A \subset \mathbb{R}^d$ is compact, and x, y are not both in A, then $d_A(x,y) = D$ for some constant $D \gg \max_{x,y\in A} d_A(x,y)$. Given a k-dimensional submanifold $\mathcal{M}$ of $\mathbb{R}^d$, $\Omega^h_{\mathcal{M}}$ denotes the set $\{x \in \mathbb{R}^d : \ d(\mathcal{M},x) \le h\}$ (here the distance $d(\cdot,\cdot)$ is the Euclidean one). This is basically an h-offset of $\mathcal{M}$. To state that the sequence of functions $\{f_n(\cdot)\}_{n\in\mathbb{N}}$ uniformly converges to $f(\cdot)$ as $n \uparrow \infty$, we frequently write $f_n \overset{n}{\rightrightarrows} f$. For a given event $\mathcal{E}$, $\mathbb{P}(\mathcal{E})$ stands for its probability of occurring. For a random variable (R.V. from now on) X, its mean value is denoted by $\mathbb{E}(X)$. By

[5] Recent results such as those reported in [57] provide efficient techniques for constructing such bands for point cloud data.

[6] In addition to studying the computation of distance functions on point clouds, [9, 67] address the important combination of this with multidimensional scaling for manifold analysis. Prior work on using geodesics and multidimensional scaling can be found in [61].

[7] While concluding this paper, we learned of a recent extension to Isomap reported in [31]. This paper is also mesh-based, and follows the geodesics approach in Isomap with a novel neighborhood/connectivity approach and a number of interesting theoretical results and novel dimensionality estimation contributions. Further analysis of Isomap, as a dimensionality reduction technique, can be found in [19].

$X \sim \mathbf{U}[A]$ we mean that the R.V. X is *uniformly distributed* in the set A. For a function $f : \Omega \to \mathbb{R}$ and a subset A of Ω, $f|_A : A \to \mathbb{R}$ denotes the restriction of f to A. For a smooth function $f : \Omega \to \mathbb{R}$, Df, $D^2 f$, and $D^3 f$ stand for the first, second (Hessian matrix), and third differential, respectively, of f. Given a point x on the complete manifold $\mathcal{S}$, $B_{\mathcal{S}}(x, r)$ will denote the (intrinsic) open ball of radius $r > 0$ centered at x, and $B(y, r)$ will denote the *Euclidean* ball centered at y of radius r. Finally, $\log x$ will denote the natural logarithm of $x \in \mathbb{R}^+$.

2.2. Prelude. In [49], we presented a new approach for the computation of weighted intrinsic distance functions on hyper-surfaces. We proved convergence theorems and addressed the fast, computationally optimal, computation of such approximations; see comments after Theorem 1 below. The key starting idea is that distance functions satisfy the (intrinsic) Eikonal equation, a particular case of the general class of Hamilton–Jacobi partial differential equations. Given $p \in \mathcal{S}$ (a hypersurface in $\mathbb{R}^d$), we want to compute $d_{\mathcal{S}}(p, \cdot) : \mathcal{S} \to \mathbb{R}^+ \cup \{0\}$, the intrinsic distance function from every point on $\mathcal{S}$ to p. It is well known that the distance function $d_{\mathcal{S}}(p, \cdot)$ satisfies, in the viscosity sense (see [47]), the equation

$$\begin{cases} \|\nabla_{\mathcal{S}} d_{\mathcal{S}}(p, x)\| = 1 & \forall x \in \mathcal{S}, \\ d_{\mathcal{S}}(p, p) = 0, \end{cases}$$

where $\nabla_{\mathcal{S}}$ is the intrinsic differentiation (gradient). Instead of solving this intrinsic Eikonal equation on $\mathcal{S}$, we solve the corresponding extrinsic one in the offset band $\Omega^h_{\mathcal{S}}$:

$$\begin{cases} \|\nabla_x d_{\Omega^h_{\mathcal{S}}}(p, x)\| = 1 & \forall x \in \Omega^h_{\mathcal{S}}; \\ d_{\Omega^h_{\mathcal{S}}}(p, p) = 0, \end{cases}$$

where $d_{\Omega^h_{\mathcal{S}}}(p, \cdot)$ is the Euclidean distance and therefore now the differentiation is the usual one.

THEOREM 1 (see [49]). *Let p and q be any two points on the smooth (orientable, without boundary) hypersurface $\mathcal{S}$; then $\left| d_{\mathcal{S}}(p, q) - d_{\Omega^h_{\mathcal{S}}}(p, q) \right| \leq C_{\mathcal{S}} \sqrt{h}$ for small enough h,*[8] *where $C_{\mathcal{S}}$ is a constant depending on the geometry of $\mathcal{S}$.*

This simplification of the intrinsic problem into an extrinsic one permits the use of the computationally optimal algorithms mentioned in the introduction. This makes computing intrinsic distances, and from them geodesics, as simple and computationally efficient as computing them in Euclidean spaces. Moreover, as detailed in [49], the approximation of the intrinsic distance $d_{\mathcal{S}}$ by the extrinsic Euclidean one $d_{\Omega^h_{\mathcal{S}}}$ is never less accurate than the numerical error of these algorithms.

In [49], the result above was limited to hypersurfaces of $\mathbb{R}^d$ (codimension one submanifolds of $\mathbb{R}^d$) without boundary, and the theory was applied to implicit surfaces, where computing the offset band is straightforward. It is the purpose of the present work to extend Theorem 1 to deal with (1) submanifolds of $\mathbb{R}^d$ of any codimension and possibly with boundary,[9] (2) convergence of geodesic curves in addition to distance functions, (3) submanifolds of $\mathbb{R}^d$ represented as point clouds and (4) random sampling of submanifolds of $\mathbb{R}^d$ in the presence of noise. We should note that Theorem 1 holds even when the metric is not the one inherited from $\mathbb{R}^d$, obtaining weighted distance

[8] "Small enough h" means that $h < 1/\max_i \kappa_i(\mathcal{S})$, where $\kappa_i(\mathcal{S})$ is the ith principal curvature of $\mathcal{S}$. This guarantees having smoothness in $\partial\Omega^h_{\mathcal{S}}$; see [49].

[9] We will later impose some convexity conditions on the boundary in order to get rate of convergence estimates. However, the uniform convergence in itself doesn't require other hypotheses beyond smoothness.

functions; see [49]. Although we will not present these new results in such generality, this is a simple extension that will be reported elsewhere.

3. Submanifolds of $\mathbb{R}^d$ with boundary. We first extend Theorem 1 to more general manifolds, and we deal not only with distance functions but also with geodesics. The first extension is important for the learning of high-dimensional manifolds from samples and for scanned open volumes. The extension to geodesics is important for path planning on surfaces and for finding special curves such as crests and valleys; see [8, 49].

First we need to recall some results that will be key ingredients in our proofs below. All our results rest upon a certain degree of smoothness of geodesics in manifolds with boundary. We use "shortest path" and "minimizing geodesic" interchangeably.

THEOREM 2 (see [1]). *Let $\mathcal{M}$ be a C^3 Riemannian manifold with C^1 boundary $\partial\mathcal{M}$. Then any shortest path of $\partial\mathcal{M}$ is C^1.*

We will eventually need more regularity on the geodesics than simply C^1. This is achieved by requiring more regularity of the boundary.

THEOREM 3 ([48]). *Let $\mathcal{U} : \mathbb{R}^d \to \mathbb{R}$ be a C^3 function such that for some $h \in \mathbb{R}$*

(i) *the interior of $\{x \in \mathbb{R}^d | \, \mathcal{U}(x) = h\}$ is nonempty and there we have $D\mathcal{U}(x) \neq 0$.*

(ii) *the "obstacle" $\{x \in \mathbb{R}^d | \, \mathcal{U}(x) \geq h\}$ is compact.*

Let p and q be any two points in the same connected component of $\{x \in \mathbb{R}^d | \, \mathcal{U}(x) \leq h\}$; then the shortest (constrained) path joining both points is C^1 and has Lipschitz first derivative.

We now present the usual definition of length, as follows.

DEFINITION 1. *Let $\alpha : [a, b] \to \mathbb{R}^d$ be a curve, then we define its length $\mathbf{L}(\alpha)$ as*

$$\mathbf{L}(\alpha) \stackrel{\triangle}{=} \sup_{a=t_0<\cdots<t_N=b} \sum_{k=0}^{N-1} \|\alpha(t_{k+1}) - \alpha(t_k)\|.$$

Remark 1. Note that if α is Lipschitz with constant $\mathcal{L}_\alpha$, then $\mathbf{L}(\alpha) = \int_a^b \|\dot{\alpha}(t)\| dt$ and $\mathbf{L}(\alpha) \leq \mathcal{L}_\alpha (b - a)$.

PROPOSITION 1. *Let $\mathcal{S}$ be a smooth compact submanifold of $\mathbb{R}^d$ with boundary $\partial\mathcal{S}$. Let x, y be any two points in $\mathcal{S}$. Then $d_{\Omega_{\mathcal{S}}^h}(x, y)$ converges pointwise as $h \downarrow 0$.*

Proof. Since $\Omega_{\mathcal{S}}^h \subseteq \Omega_{\mathcal{S}}^{h'}$ if $h' \geq h$, we have that $d_{\Omega_{\mathcal{S}}^h}(x, y) \geq d_{\Omega_{\mathcal{S}}^{h'}}(x, y)$ Also, for any $h > 0$, $d_{\Omega_{\mathcal{S}}^h}(x, y) \leq d_{\mathcal{S}}(x, y) \leq \operatorname{diam}(\mathcal{S}) < +\infty$. Hence, the sequence $\{d_{\Omega_{\mathcal{S}}^h}(x, y)\}_{h>0}$ (for fixed x and y over $\mathcal{S}$) is bounded and nondecreasing, and therefore it converges to the supremum of its range. ∎

THEOREM 4. *Let $\mathcal{S}$ be a compact C^2 submanifold of $\mathbb{R}^d$ with (possibly empty) smooth boundary $\partial\mathcal{S}$. Let x, y be any two points in $\mathcal{S}$. Then we have*

1. *uniform convergence of the distances:*

$$d_{\Omega_{\mathcal{S}}^h}|_{\mathcal{S}\times\mathcal{S}}(\cdot, \cdot) \stackrel{h\downarrow 0}{\rightrightarrows} d_{\mathcal{S}}(\cdot, \cdot);$$

2. *convergence of the geodesics: Let x and y be joined by a unique minimizing geodesic $\gamma_{\mathcal{S}} : [0, 1] \to \mathcal{S}$ over $\mathcal{S}$, and let $\gamma_h : [0, 1] \to \Omega_{\mathcal{S}}^h$ be a $\Omega_{\mathcal{S}}^h$-minimizing geodesic; then*

$$\gamma_h \stackrel{h\downarrow 0}{\rightrightarrows} \gamma_{\mathcal{S}}.$$

Proof. Given our hypothesis on $\mathcal{S}$, and according to [26], there exists $H > 0$ such that $\partial\Omega_{\mathcal{S}}^h$ is $C^{1,1}$ for all $0 < h \leq H$. Then Theorem 2 guarantees that for $0 < h \leq H$, $\gamma_h : [0,1] \to \Omega_{\mathcal{S}}^H$, the $\Omega_{\mathcal{S}}^h$ length-minimizing geodesic joining x and y is of class C^1. Since $d_{\Omega_{\mathcal{S}}^h}(x,y) \leq d_{\mathcal{S}}(x,y) \leq \operatorname{diam}(\mathcal{S}) < +\infty$ for any $h \in (0,H]$, we see that we can admit our $\Omega_{\mathcal{S}}^h$-geodesics to have Lipschitz constant $\mathcal{L} \leq \operatorname{diam}(\mathcal{S})$. Obviously, the set $\Omega_{\mathcal{S}}^H$ is bounded, and then the family $\{\gamma_h\}_{0<h\leq H}$ is bounded and equicontinuous. Hence, by the Ascoli–Arzelá theorem, there exist a subsequence $\{\gamma_{h_k}\}_{k\in\mathbb{N}}$ and a curve $\gamma_0 \in C^0([0,1],\mathcal{S})$ such that $\max_{t\in[0,1]} \|\gamma_{h_k}(t) - \gamma_0(t)\| \overset{h_k\downarrow 0}{\longrightarrow} 0$.

Moreover, by writing $|\gamma_0(t)-\gamma_0(t')| \leq |\gamma_{h_k}(t)-\gamma_0(t)| + |\gamma_{h_k}(t')-\gamma_0(t')| + \mathcal{L}|t-t'|$ and using the (pointwise) convergence of γ_{h_k} towards γ_0, we find that $\mathcal{L}$ is also a Lipschitz constant for γ_0. Then we have $\gamma_0 \in C^{0,1}([0,1],\mathcal{S})$.

Now, since γ_0 lies on $\mathcal{S}$ but may not be a shortest path, we have that its (finite) length is greater than or equal to $d_{\mathcal{S}}(x,y)$. We also have the trivial inequality $d_{\mathcal{S}}(x,y) \geq d_{\Omega_{\mathcal{S}}^h}(x,y)$. Putting this all together, we obtain

$$\mathbf{L}(\gamma_h) = d_{\Omega_{\mathcal{S}}^h}(x,y) \leq d_{\mathcal{S}}(x,y) \leq \mathbf{L}(\gamma_0).$$

Therefore

$$\limsup_{h\downarrow 0} \mathbf{L}(\gamma_h) = \limsup_{h\downarrow 0} d_{\Omega_{\mathcal{S}}^h}(x,y) \leq d_{\mathcal{S}}(x,y) \leq \mathbf{L}(\gamma_0).$$

Note that $\mathbf{L}(\gamma_0) = \mathbf{L}(\lim_{h_k\downarrow 0}\gamma_{h_k}) \leq \liminf_{h_k\downarrow 0} \mathbf{L}(\gamma_{h_k})$. This is the semicontinuity of length, an immediate consequence of its definition; see [41].

Since $\liminf_{h_k\downarrow 0}(\cdot) \leq \limsup_{h_k\downarrow 0}(\cdot) \leq \limsup_{h\downarrow 0}(\cdot)$, we see that $\limsup_{h\downarrow 0} d_{\Omega_{\mathcal{S}}^h}(x,y)$ $= \limsup_{h\downarrow 0} \mathbf{L}(\gamma_h)$ equals $d_{\mathcal{S}}(x,y)$ for all x and y in $\mathcal{S}$. From Proposition 1, we find that in fact $\lim_{h\downarrow 0} d_{\Omega_{\mathcal{S}}^h}(x,y)$ exists and equals $d_{\mathcal{S}}(x,y)$. Then, we have that the function $d_{\Omega_{\mathcal{S}}^h}|_{\mathcal{S}\times\mathcal{S}}(\cdot,\cdot)$ satisfies the following:

(i) $d_{\Omega_{\mathcal{S}}^h}|_{\mathcal{S}\times\mathcal{S}} : \mathcal{S}\times\mathcal{S} \to \mathbb{R}\cup\{0\}$ is continuous for each $H > h > 0$;
(ii) for each $(x,y) \in \mathcal{S}\times\mathcal{S}$, $\{d_{\Omega_{\mathcal{S}}^h}|_{\mathcal{S}\times\mathcal{S}}(x,y)\}_h$ is nondecreasing;
(iii) $d_{\Omega_{\mathcal{S}}^h}|_{\mathcal{S}\times\mathcal{S}}(\cdot,\cdot)$ converges pointwise towards $d_{\mathcal{S}}(\cdot,\cdot)$, which is continuous.

Then by Dini's uniform convergence theorem (see [6]) we can conclude that the convergence is uniform.

We can also see that γ_0 must be a minimizing geodesic of $\mathcal{S}$ since from the above chain of equalities $\mathbf{L}(\gamma_0) = d_{\mathcal{S}}(x,y)$. Then, if there was only one such curve joining x with y, we would have uniform convergence (along any subsequence!) of γ_h towards γ_0.[10] □

Remark 2. In Theorem 4, the convergence (of distances) is uniform, but we will have forfeited rate of convergence estimates unless we impose additional conditions on $\partial\mathcal{S}$, as we do in Corollary 3. Note that the new setting is wider than the one considered in Theorem 1 since the codimension of the underlying manifold is not necessarily 1. This is very important for applications such as dimensionality reduction, where the dimension of the underlying manifold is unknown beforehand.

COROLLARY 1. *Let $\mathcal{S}$ and $\partial\mathcal{S}$ satisfy the hypotheses of Theorem* 4. *Let $\{\Sigma_i\}_{i\in\mathbb{N}}$ be a family of compact of sets in $\mathbb{R}^d$ such that $\mathcal{S} \subseteq \Sigma_i$ for all $i \in \mathbb{N}$ and $d_{\mathcal{H}}(\Sigma_i,\mathcal{S}) \overset{i\uparrow+\infty}{\longrightarrow} 0$.*

[10] This follows from the fact that uniform convergence of γ_h to γ_0 is equivalent to the statement that for any subsequence $\{\gamma_{h_i}\}$ there exists a further subsubsequence $\{\gamma_{h_{i_k}}\}$ uniformly converging to γ_0.

Then,

$$d_{\Sigma_i}(\cdot,\cdot)\,|_{\mathcal{S}\times\mathcal{S}} \overset{i\uparrow+\infty}{\rightrightarrows} d_{\mathcal{S}}(\cdot,\cdot),$$

where $d_{\mathcal{H}}$ stands for the Hausdorff distance between sets.

We now present a uniform rate of convergence result for the distance in the band in the case $\partial\mathcal{S}=\emptyset$, and from this we deduce Corollary 3 below, which deals with the case $\partial\mathcal{S}\neq\emptyset$. This result generalizes the one presented in [49] because it allows for any codimension.

THEOREM 5. *Under the hypotheses of Theorem 4, with $\partial\mathcal{S}=\emptyset$, we have that for small enough $h>0$,*

$$\max_{(x,y)\in\mathcal{S}\times\mathcal{S}} \left| d_{\Omega_{\mathcal{S}}^h}|_{\mathcal{S}\times\mathcal{S}}(x,y) - d_{\mathcal{S}}(x,y)\right| \leq C_{\mathcal{S}}\sqrt{h}, \tag{1}$$

where the constant $C_{\mathcal{S}}$ does not depend on h. Also, we have the "relative" rate of convergence bound

$$1 \leq \sup_{\substack{x,y\in\mathcal{S}\\ x\neq y}} \frac{d_{\mathcal{S}}(x,y)}{d_{\Omega_{\mathcal{S}}^h}(x,y)} \leq 1 + C_{\mathcal{S}}\sqrt{h}. \tag{2}$$

Proof. This is a remake of our proof of the main theorem in [49]; therefore we skip some technical details which can be found there. Throughout the proof we will sometimes write d_h instead of $d_{\Omega_{\mathcal{S}}^h}$ for the sake of notational simplicity. We will denote by k ($\leq n-1$) the dimension of $\mathcal{S}$.

Let γ_0 be the arc length parametrized $\mathcal{S}$-shortest path joining the points $x,\ y\in\mathcal{S}$; clearly, we have $\mathrm{trace}(\gamma_0)\subset\mathcal{S}$. Let γ_h be the $\Omega_{\mathcal{S}}^h$ arc length parametrized shortest path joining x and y, which, as we know from Theorem 4, uniformly converges toward γ_0. For a number H as in the proof of Theorem 4, we have $\gamma_h\in C^{1,1}([0,d_h],\mathcal{S})$, and also $\eta:\Omega_{\mathcal{S}}^H\to\mathbb{R}$ defined by $\eta(x)\overset{\triangle}{=}\frac{1}{2}d^2(x,\mathcal{S})$ is smooth; see Appendix B. We define the projection operator $\Pi_{\mathcal{S}}:\Omega_{\mathcal{S}}^H\to\mathcal{S}$ by $\Pi_{\mathcal{S}}(x)=x-D\eta(x)$. We refer the reader to Appendix B for properties of $\Pi_{\mathcal{S}}$ and η which we use below.

Now, $d_{\Omega_{\mathcal{S}}^h}(x,y)=\mathbf{L}(\gamma_h)\leq d_{\mathcal{S}}(x,y)\leq\mathbf{L}(\Pi_{\mathcal{S}}(\gamma_h))$; then

$$\begin{aligned}
d_{\mathcal{S}}(x,y)-d_{\Omega_{\mathcal{S}}^h}(x,y) &\leq |\mathbf{L}(\Pi_{\mathcal{S}}(\gamma_h))-\mathbf{L}(\gamma_h)|\\
&\leq \int_0^{d_h}\left\|\dot{\overline{\Pi_{\mathcal{S}}(\gamma_h(t))-\gamma_h(t)}}\right\|\,dt\\
&= \int_0^{d_h}\left\|\dot{\overline{D\eta(\gamma_h(t))}}\right\|\,dt\\
&\leq \sqrt{d_h\int_0^{d_h}\dot{V}(t)\cdot\dot{V}(t)\,dt} \qquad \text{(by Cauchy–Schwarz inequality)}\\
&\leq \sqrt{d_h\int_0^{d_h}V(t)\cdot\ddot{V}(t)\,dt} \quad \text{(integrating by parts; see below),}
\end{aligned}$$

where $V(t)\overset{\triangle}{=}D\eta(\gamma_h(t))$ and $V(0)=V(1)=0$; see Appendix B.

Also $\dot{V}(t)=D^2\eta(\gamma_h(t))\dot{\gamma}(t)$, and since $\dot{\gamma_h}$ is Lipschitz and η is smooth, $\ddot{V}(t)$ exists almost everywhere and $\ddot{V}(t)=D^3\eta(\gamma_h(t))[\dot{\gamma}_h(t),\dot{\gamma}_h(t)]+D^2\eta(\gamma_h(t))\ddot{\gamma}(t)$ at points of

existence. Then since $D^3\eta D\eta = D^2\eta(\boldsymbol{I} - D^2\eta)$ and $D^2\eta D\eta = D\eta$ (see Appendix B),

$$\begin{aligned} V \cdot \ddot{V} &= D^3\eta(\gamma_h)[D\eta(\gamma_h), \dot{\gamma_h}, \dot{\gamma_h}] + D^2\eta[\ddot{\gamma_h}, D\eta(\gamma_h)] \\ &= \left(D^2\eta(\gamma_h)\left(\boldsymbol{I} - D^2\eta(\gamma_h)\right)\right)[\dot{\gamma_h}, \dot{\gamma_h}] + \ddot{\gamma_h} \cdot D\eta(\gamma_h). \end{aligned}$$

The matrix $\Lambda(t) \stackrel{\triangle}{=} D^2\eta(\gamma_h(t))(\boldsymbol{I} - D^2\eta(\gamma_h(t)))$ filters out normal components and has eigenvalues associated with the tangential bundle given by

$$\lambda_i(t) = \frac{d(t)\lambda_i(0)}{(1 + d(t)\lambda_i(0))^2} \qquad \text{for } 1 \leq i \leq k,$$

where we let $d(t) = d(\gamma_h(t), \mathcal{S})$. Note that $\max_{1\leq i\leq k}|\lambda_i(t)|$ can be bounded by $d(t)$ times a certain finite constant K' independent of h.

On the other hand, we can bound $|\ddot{\gamma}_h(t)|$ almost anywhere by a finite constant, say K, which takes into account the maximal curvature of all the boundaries $\partial\Omega^h_{\mathcal{S}}$, $0 < h < H$, but does not depend on h.

Putting all this together, we find (recall that $\|D\eta(x)\| = \sqrt{2\eta(x)} = d(x, \mathcal{S})$; see Appendix B)

$$\begin{aligned} \left(d_{\mathcal{S}}(x,y) - d_{\Omega^h_{\mathcal{S}}}(x,y)\right)^2 &\leq d_h \int_0^{d_h} \Lambda(t)[\dot{\gamma}_h, \dot{\gamma}_h]dt \\ &\quad + d_h \int_0^{d_h} \|\ddot{\gamma}_h\| \, \|D\eta(\gamma_h)\|dt \\ &\leq K' \max_{t\in[0,d_h]} d(t)\, d_h^2 + K \max_{t\in[0,d_h]} d(t) d_h^2. \end{aligned}$$

Now, remembering that d_h stands for $d_{\Omega^h_{\mathcal{S}}}(x,y)$, that $\mathrm{trace}(\gamma_h) \subset \Omega^h_{\mathcal{S}}$, and defining $C = K + K'$, we arrive with only a little simple additional work, at the relations (1) or (2). ∎

Remark 3. Note that, as the simple case of a circle in the plane shows, the rate of convergence is at most $C \cdot h$.

We immediately obtain the following corollary, which will be useful ahead.

COROLLARY 2. *Let $p \in \mathcal{S}$ and $r \leq H$; then $B(p,r) \cap \mathcal{S} \subseteq B_{\mathcal{S}}(p, r(1 + C_{\mathcal{S}}\sqrt{r}))$.*

Proof. Let $q \in B(p,r) \cap \mathcal{S}$; then by (2), $d_{\mathcal{S}}(p,q) \leq d_{\Omega^r_{\mathcal{S}}}(p,q)(1 + C_{\mathcal{S}}\sqrt{r})$. However, $q \in B(p,r) \subset \Omega^r_{\mathcal{S}}$, and thus $d_{\Omega^h_{\mathcal{S}}}(p,q) = \|p - q\| \leq r$, which completes the proof. ∎

DEFINITION 2. (see [21]) *We say that the compact manifold $\mathcal{S}$ with boundary $\partial\mathcal{S}$ is* strongly convex *if for every pair of points x and y in $\mathcal{S}$ there exists a unique minimizing geodesic joining them whose interior is contained in the interior of $\mathcal{S}$.*

Using basically the same procedure as in Theorem 5 with the convexity hypotheses above, we can prove the following corollary, whose (sketched) proof is presented in Appendix C.

COROLLARY 3. *Under the hypotheses of Theorem 2, and assuming $\mathcal{S}$ to be strongly convex, we have for small enough $h > 0$ the same conclusions of Theorem 5 (rate of convergence).*

Remark 4. Note that in case $\partial\mathcal{S} \neq \emptyset$ is not strongly convex, then obviously the same statement of Corollary 3 remains valid for any strongly convex subset of $\mathcal{S}$.

To conclude, in this section we extended the results in [49] to geodesics and distance functions in general codimension manifolds with or without (smooth) boundary, thereby covering all possible manifolds in common shape, graphics, visualization, and

learning applications.[11] We are now ready to extend this to manifolds represented as point clouds.

4. Distance functions on point clouds. We are now interested in making computations on manifolds represented as point clouds, i.e., *sampled manifolds*. In the case of this paper we will restrict ourselves to the computation of intrinsic distances.[12] Let $\mathcal{P}_n \stackrel{\triangle}{=} \{p_1, \ldots, p_n\}$ be a set of n different points sampled from the *compact* submanifold $\mathcal{S}$ and define[13]

$$\Omega^h_{\mathcal{P}_n} \stackrel{\triangle}{=} \bigcup_{i=1}^{n} B(p_i, h).$$

Let h and $\mathcal{P}_n$ be such that $\mathcal{S} \subseteq \Omega^h_{\mathcal{P}_n}$. We then have $(\mathcal{S} \subseteq)\Omega^h_{\mathcal{P}_n} \subseteq \Omega^h_{\mathcal{S}}$. We now want to consider $d_{\Omega^h_{\mathcal{P}}}(p, q)$ for any pair of points $p, q \in \mathcal{S}$ and prove some kind of proximity to the real distance $d_{\mathcal{S}}(p, q)$. The argument carries over easily since

$$d_{\Omega^h_{\mathcal{S}}}(p, q) \leq d_{\Omega^h_{\mathcal{P}_n}}(p, q) \leq d_{\mathcal{S}}(p, q),$$

and hence

$$0 \leq d_{\mathcal{S}}(p, q) \; - \; d_{\Omega^h_{\mathcal{P}_n}}(p, q) \leq d_{\mathcal{S}}(p, q) \; - \; d_{\Omega^h_{\mathcal{S}}}(p, q), \tag{3}$$

and the rightmost quantity can be bounded by $C_{\mathcal{S}}\, h^{1/2}$ (see section 3) in the case that $\partial\mathcal{S}$ is either convex or void. In general, without hypotheses on $\partial\mathcal{S}$ other than some degree of smoothness, we can also work out uniform convergence since by virtue of Theorem 4 the upper bound in (3) uniformly converges to 0. The key condition is $\mathcal{S} \subset \Omega_{\mathcal{P}^h_n}$, something that can obviously be coped with using the compactness of $\mathcal{S}$.[14] We can then state the following claim.

THEOREM 6 (uniform convergence for point clouds). *Let $\mathcal{S}$ be a compact smooth submanifold of $\mathbb{R}^d$ possibly with boundary $\partial\mathcal{S}$. Then the following hold:*

1. General case: *Given $\varepsilon > 0$, there exists $h_\varepsilon > 0$ such that for all $0 < h \leq h_\varepsilon$ one can find finite $n(h)$ and a set of points $\mathcal{P}_{n(h)}(h) = \{p_1(h), \ldots, p_{n(h)}(h)\}$ sampled from $\mathcal{S}$ such that*

$$\max_{p,q \in \mathcal{S}} \left(d_{\mathcal{S}}(p, q) - d_{\Omega^h_{\mathcal{P}_{n(h)}(h)}}(p, q) \right) \leq \varepsilon.$$

2. $\partial\mathcal{S}$ is either void or convex: *For every sufficiently small $h > 0$ one can find finite $n(h)$ and a set of points $\mathcal{P}_{n(h)}(h) = \{p_1(h), \ldots, p_{n(h)}(h)\}$ sampled from $\mathcal{S}$ such that*

$$\max_{p,q \in \mathcal{S}} \left(d_{\mathcal{S}}(p, q) - d_{\Omega^h_{\mathcal{P}_{n(h)}(h)}}(p, q) \right) \leq C_{\mathcal{S}} \sqrt{h}.$$

[11] Although in this paper we consider only manifolds with constant codimension, many of the results are extendible to variable codimensions, and this will be reported elsewhere.

[12] Note that having the intrinsic distance allows us to compute basic intrinsic properties of the manifold; see e.g., [13].

[13] The balls now used are defined with respect to the metric of $\mathbb{R}^d$; they are not intrinsic.

[14] By compactness, given $h > 0$, we can find finite $N(h)$ and points $p_1, p_2, \ldots, p_{N(h)} \in \mathcal{S}$ such that $\mathcal{S} = \cup_{i=1}^{N(h)} B_{\mathcal{S}}(p_i, h)$. But since for $p \in \mathcal{S}$, $B_{\mathcal{S}}(p, h) \subset B(p, h) \cap \mathcal{S}$, we also get $\mathcal{S} \subset \cup_{i=1}^{N(h)} B(p_i, h)$.

In practice, one must worry about both the number of points and the radii of the balls. Obviously, there is a tradeoff between both quantities. If we want to use few points, in order to cover $\mathcal{S}$ with the balls we have to increase the value of the radius. Clearly, there exists a value H such that for values of h smaller than H we do not change the topology; see [3, 4, 5]. This implies that the number of points must be larger than a certain lower bound. This result can be generalized to ellipsoids which can be locally adapted to the geometry of the point cloud [15], or from minimal spanning trees. Note that we are interested in the smallest possible offset of the point cloud that covers $\mathcal{S}$. Further comments on this are presented below and are also the subject of current efforts to be reported elsewhere.

The practical significance of the previous Theorem is clear. Part 1 says that in general, given a desired precision for the computation of the distance, we have a maximum *nonzero* value for the radius of all the balls, below which we can always find a *finite* number of points sampled from the manifold for which the "Ω-set" formed by those points achieves the desired accuracy;[15] that is, we can choose the radius at our convenience *within* a certain range which depends on this level of accuracy. Part 2 says more, since it actually links ε to h_ε. It basically says that the radius of the balls must be of the order of the square of the desired error.

5. Extension to random sampling of manifolds. In practice, we really do not have too much control over the way in which points are sampled by the acquisition device (e.g., scanner) or given by the learned sampled data. Therefore it is more realistic to make a probabilistic model of the situation and then try to conveniently estimate the probability of achieving a prescribed level of accuracy as a function of the number of points and the radii of the balls. It will be interesting to see how geometric quantities of $\mathcal{S}$ enter in those bounds we will establish. However, since the bounds are based in local volume computations and all manifolds are locally Euclidean, those curvature dependent quantities will be asymptotically negligible.

We now present a simple model for the current setting, while results for other models can be developed from the derivations below. Here we assume that the points in $\mathcal{P}_n$ are independently and identically sampled on the submanifold $\mathcal{S}$ with the uniform probability law;[16] we will write this as $p_i \sim \mathbf{U}[\mathcal{S}]$. For simplicity of exposition, we will restrict ourselves to the case when $\mathcal{S}$ has no boundary.[17] Also, we deal only with uniform independently and identically distributed (i.i.d.) sampling; results for other sampling models, including those adapted to the manifold geometry, can be easily obtained following the developments below and will be reported elsewhere.

We have to define the way in which we are going to measure accuracy. A possibility for such a measure is (for each $\varepsilon > 0$)

$$\mathbb{P}\left(\max_{p,q\in\mathcal{S}}\left(d_{\mathcal{S}}(p,q) - d_{\Omega^h_{\mathcal{P}_n}}(p,q)\right) > \varepsilon\right). \tag{4}$$

There is a potential problem with this way of testing accuracy, since we are assuming that when we use the approximate distance, we will be evaluating it on $\mathcal{S}$. This might seem a bit awkward since we don't exactly know all the surface but just

[15] We are considering the case when all the balls have the same radii.

[16] This means that for any subset $A \subseteq \mathcal{S}$ and any $p_i \in \mathcal{P}_n$, $\mathbb{P}(p_i \in A) = \frac{\mu(A)}{\mu(S)}$, where $\mu(\cdot)$ stands for the measure (area/volume) of the set.

[17] In order to extend the results in this section to the case $\partial\mathcal{S} \neq \emptyset$, the same considerations discussed in [9] remain valid in our case.

some points on it. Moreover, a more natural and real-problem-motivated approach would be to measure the discrepancy over $\mathcal{P}_n$ itself (see section 7 ahead), over part of this set, or over another *trial* set of points $\mathcal{Q}_m$.

However, since for any set of points $\mathcal{Q}_m \subset \mathcal{S}$ we have that the following inclusion of events,

$$\left\{ \max_{p,q\in\mathcal{Q}_m} \left(d_{\mathcal{S}}(p,q) - d_{\Omega^h_{\mathcal{P}_n}}(p,q) \right) > \varepsilon \right\} \subseteq \left\{ \max_{p,q\in\mathcal{S}} \left(d_{\mathcal{S}}(p,q) - d_{\Omega^h_{\mathcal{P}_n}}(p,q) \right) > \varepsilon \right\},$$

holds, bounding (4) suffices for dealing with any of the possibilities mentioned above. Note that we are somehow considering $d_{\Omega^h_{\mathcal{P}_n}}$ defined for all pairs of points in $\mathcal{S} \times \mathcal{S}$, even if it might happen that $\mathcal{S} \cap \Omega^h_{\mathcal{P}_n} \neq \mathcal{S}$. In any case we extend $d_{\Omega^h_{\mathcal{P}_n}}$ to all $\Omega^h_{\mathcal{S}} \times \Omega_{\mathcal{S}}$ by a large constant, say $k \operatorname{diam}(\mathcal{S})$, $k \gg 1$.

Let us spell out a few definitions so as to avoid an overload of notation:

$$\mathcal{E}_\varepsilon \triangleq \left\{ \max_{p,q\in\mathcal{S}} \left(d_{\mathcal{S}}(p,q) - d_{\Omega^h_{\mathcal{P}_n}}(p,q) \right) > \varepsilon \right\}, \tag{5}$$

$$\mathcal{I}_{h,n} \triangleq \left\{ \mathcal{S} \subseteq \Omega^h_{\mathcal{P}_n} \right\}. \tag{6}$$

Now, since $\mathcal{E}_\varepsilon = (\mathcal{E}_\varepsilon \cap \mathcal{I}_{h,n}) \cup (\mathcal{E}_\varepsilon \cap \mathcal{I}^c_{h,n})$, using the union bound and then Bayes rule, we have

$$\begin{aligned} \mathbb{P}(\mathcal{E}_\varepsilon) &\leq \mathbb{P}(\mathcal{E}_\varepsilon \cap \mathcal{I}_{h,n}) + \mathbb{P}\left(\mathcal{E}_\varepsilon \cap \mathcal{I}^c_{h,n}\right) \\ &= \mathbb{P}(\mathcal{E}_\varepsilon \mid \mathcal{I}_{h,n})\, \mathbb{P}(\mathcal{I}_{h,n}) + \mathbb{P}\left(\mathcal{E}_\varepsilon \mid \mathcal{I}^c_{h,n}\right) \mathbb{P}\left(\mathcal{I}^c_{h,n}\right) \end{aligned}$$

$$\Downarrow$$

$$\mathbb{P}(\mathcal{E}_\varepsilon) \leq \mathbb{P}(\mathcal{E}_\varepsilon \mid \mathcal{I}_{h,n}) + \mathbb{P}\left(\mathcal{I}^c_{h,n}\right). \tag{7}$$

It is clear now that we must find a convenient lower bound for the second term in the previous expression, the probability of covering all $\mathcal{S}$ with the union of balls. (The first term will be dealt with using the convergence theorems presented in previous sections.) For this we need a few lemmas.

LEMMA 1. *Let K be an upper bound for the sectional curvatures of $\mathcal{S}$ ($\operatorname{diam}(\mathcal{S}) = k$) and $x \in \mathcal{S}$ be a fixed point. Then, under the hypotheses on $\mathcal{P}_n$ described above, there exist a constant $\omega_k > 0$ and a function $\theta_{\mathcal{S}}(\cdot)$ with $\lim_{h\downarrow 0} \frac{\theta_{\mathcal{S}}(h)}{h^{k+1}} = 0$ such that for small enough $h > 0$*

$$\mathbb{P}\left(\left\{x \notin \Omega^h_{\mathcal{P}_n} \cap \mathcal{S}\right\}\right) \leq \left(1 - \frac{\omega_k\, h^k + \theta_{\mathcal{S}}(h)}{\mu(\mathcal{S})}\right)^n. \tag{8}$$

Moreover, one can further expand the right-hand side of (8) *as*

$$\left(1 - \frac{\omega_k h^k (1 - K c_k h^2) + \phi_{\mathcal{S}}(h)}{\mu(\mathcal{S})}\right)^n$$

for some c_k depending only on the dimension k of $\mathcal{S}$ and a function $\phi_{\mathcal{S}}$ such that $\frac{\phi_{\mathcal{S}}(h)}{h^{k+2}} \to 0$ as $h \downarrow 0$.

Proof.

$$\mathbb{P}\left(\left\{x \notin \Omega^h_{\mathcal{P}_n} \cap \mathcal{S}\right\}\right) = \mathbb{P}\left(\left\{\bigcap_{i=1}^n \{x \notin B(p_i,h) \cap \mathcal{S}\}\right\}\right) \tag{9}$$

$$= \mathbb{P}\left(\left\{\bigcap_{i=1}^n \{p_i \notin B(x,h) \cap \mathcal{S}\}\right\}\right) \tag{10}$$

$$= \prod_{i=1}^n \mathbb{P}\left(\{p_i \notin B(x,h) \cap \mathcal{S}\}\right) \tag{11}$$

$$= \prod_{i=1}^n \left(1 - \mathbb{P}\left(\{p_i \in B(x,h) \cap \mathcal{S}\}\right)\right). \tag{12}$$

Since $B_{\mathcal{S}}(x,h) \subseteq B(x,h) \cap S$,[18] then $\mu\left(\mathcal{S} \cap B(x,h)\right) \geq \mu\left(B_{\mathcal{S}}(x,h)\right)$. On the other hand, note that

$$\begin{aligned}\mathbb{P}\left(\{p_i \in B(x,h) \cap \mathcal{S}\}\right) &= \frac{\mu\left(\mathcal{S} \cap B(x,h)\right)}{\mu\left(\mathcal{S}\right)} \\ &\geq \frac{\mu\left(B_{\mathcal{S}}(x,h)\right)}{\mu\left(\mathcal{S}\right)}.\end{aligned}$$

Finally, as shown in Appendix D, one can lower bound $\mu\left(B_{\mathcal{S}}(x,h)\right)$ using information on the curvatures of $\mathcal{S}$, by means of the Bishop–Günther volume comparison theorem. More precisely, we can write

$$\mu\left(B_{\mathcal{S}}(x,h)\right) \geq \min_{\zeta \in \mathcal{S}} \mu\left(B_{\mathcal{S}}(\zeta,h)\right) \geq \omega_k h^k + \theta_{\mathcal{S}}(h),$$

where $\frac{\theta_{\mathcal{S}}(h)}{h^q} \to 0$ when $h \to 0$ for $q \leq k+1$. Therefore, from (9) we obtain

$$\mathbb{P}\left(\left\{x \notin \Omega^h_{\mathcal{P}_n} \cap \mathcal{S}\right\}\right) \leq \left(1 - \frac{\omega_k\, h^k + \theta_{\mathcal{S}}(h)}{\mu\left(\mathcal{S}\right)}\right)^n.$$

The last assertion follows from Proposition 3. ∎

Remark 5. Note that we cannot, however, from (8), conclude that $\mathbb{P}(\mathcal{S} \nsubseteq \Omega^h_{\mathcal{P}_n}) \leq \left(1 - \frac{\omega_k\, h^k + \theta_{\mathcal{S}}(h)}{\mu(\mathcal{S})}\right)^n$. In order to upper bound $\mathbb{P}(\mathcal{S} \nsubseteq \Omega^h_{\mathcal{P}_n})$ we will first estimate $\mathbb{P}(B_{\mathcal{S}}(x,\delta) \nsubseteq \Omega^h_{\mathcal{P}_n})$ for any $x \in \mathcal{S}$ and small $\delta > 0$. Then we will use the compactness of $\mathcal{S}$ by covering it with a finite δ-net consisting of $\mathcal{N}(\mathcal{S},\delta)$ points, and conclude by using the union bound. Yet another intermediate step will therefore be to estimate the covering number $\mathcal{N}(\mathcal{S},\delta)$.

LEMMA 2. *Under the hypotheses of the previous lemma, let $\delta \in (0,h)$; then*

$$\mathbb{P}\left(B_{\mathcal{S}}(x,\delta) \nsubseteq \Omega^h_{\mathcal{P}_n}\right) \leq \left(1 - \frac{\omega_k\,(h-\delta)^k + \theta_{\mathcal{S}}(h-\delta)}{\mu\left(\mathcal{S}\right)}\right)^n. \tag{13}$$

Proof. We find α and β such that $\{B_{\mathcal{S}}(q,\delta) \subseteq \Omega^h_{\mathcal{P}_n}\} \supseteq \{q \in \Omega^{\alpha h+\beta\delta}_{\mathcal{P}_n}\}$. Note first that for any $x \in B_{\mathcal{S}}(q,\delta)$, $|x-q| \leq d_{\mathcal{S}}(x,q) \leq \delta$. Assume that the event $\{q \in \Omega^{\alpha h+\beta\delta}_{\mathcal{P}_n}\}$

[18] Consider $z \in B_{\mathcal{S}}(x,h)$; then $d_{\mathcal{S}}(x,z) \leq h$, but always $d(x,z) \leq d_{\mathcal{S}}(x,z)$, and thus $d(x,z) \leq h$, which implies $z \in B(x,h) \cap \mathcal{S}$.

holds. Then for some $p_r \in \mathcal{P}_n$, $q \in B(p_r, \alpha h + \beta\delta)$; that is, $|q - p_r| \leq \alpha h + \beta\delta$. Now, note that

$$|x - p_r| \leq |x - q| + |q - p_r| \leq \alpha h + (\beta + 1)\delta.$$

If we force the rightmost number to be h, we find that we must have $(1+\beta)\delta = (1-\alpha)h$, and then $\alpha h + \beta\delta = h - \delta$. Then we have found $B_{\mathcal{S}}(q, \delta) \subseteq B(p_r, h-\delta) \subset \Omega^h_{\mathcal{P}_n}$. Hence (using (8)), $\mathbb{P}(B_{\mathcal{S}}(q,\delta) \subseteq \Omega^h_{\mathcal{P}_n}) \geq \mathbb{P}(q \in \Omega^{h-\delta}_{\mathcal{P}_n} \geq 1 - (1 - \frac{\omega_k\,(h-\delta)^k + \theta_{\mathcal{S}}\,(h-\delta)}{\mu(\mathcal{S})})^n$. ∎

We also need the next lemma, whose proof is deferred to Appendix C.

LEMMA 3 (bounding the covering number). *Under the hypotheses of Lemma* 2 *and further assuming* $\mathcal{S}$ *to be compact, we have that for any small enough* $\delta > 0$ *there exists a* δ*-covering of* $\mathcal{S}$ *with cardinality*

$$\mathcal{N}(\mathcal{S}, \delta) \leq \frac{\mu\,(\mathcal{S})}{\omega_k(\delta/2)^k + \theta_{\mathcal{S}}(\delta/2)}. \tag{14}$$

PROPOSITION 2. *Let the set of hypotheses sustaining all of the previous lemmas hold. Let also* $([0,1) \ni)\, x_h \triangleq \frac{\omega_k(h/2)^k + \theta_{\mathcal{S}}\,(h/2)}{\mu(\mathcal{S})}$, *where* ω_k *and* $\theta_{\mathcal{S}}$ *are given as in the proof of Lemma* 1. *Then*

$$\mathbb{P}\left(\mathcal{S} \nsubseteq \Omega^h_{\mathcal{P}_n}\right) \leq \frac{e^{-nx_h}}{x_h}. \tag{15}$$

Proof. Consider a finite $\frac{h}{2}$-net covering $\mathcal{S}$ given by Lemma 3, that is, $\mathcal{S} = \bigcup_{i=1}^{\mathcal{N}\left(\mathcal{S}, \frac{h}{2}\right)} B_{\mathcal{S}}(q_i, \frac{h}{2})$; then

$$\begin{aligned}
\mathbb{P}\left(\mathcal{S} \nsubseteq \Omega^h_{\mathcal{P}_n}\right) &= \mathbb{P}\left(\bigcup_{x \in \mathcal{S}} \{x \notin \Omega^h_{\mathcal{P}_n}\}\right) \\
&= \mathbb{P}\left(\bigcup_{i=1}^{\mathcal{N}(\mathcal{S}, \frac{h}{2})} \bigcup_{x \,\in\, B_{\mathcal{S}}(q_i, \frac{h}{2})} \{x \notin \Omega^h_{\mathcal{P}_n}\}\right) \\
&\leq \mathcal{N}\left(\mathcal{S}, \frac{h}{2}\right) \max_{1 \leq i \leq \mathcal{N}\left(\mathcal{S}, \frac{h}{2}\right)} \mathbb{P}\left(\bigcup_{x \,\in\, B_{\mathcal{S}}(q_i, \frac{h}{2})} \{x \notin \Omega^h_{\mathcal{P}_n}\}\right) \\
&= \mathcal{N}\left(\mathcal{S}, \frac{h}{2}\right) \max_{1 \leq i \leq \mathcal{N}\left(\mathcal{S}, \frac{h}{2}\right)} \mathbb{P}\left(B_{\mathcal{S}}\left(q_i, \frac{h}{2}\right) \nsubseteq \Omega^h_{\mathcal{P}_n}\right) \\
&= \mathcal{N}\left(\mathcal{S}, \frac{h}{2}\right) \left(1 - \min_{1 \leq i \leq \mathcal{N}\left(\mathcal{S}, \frac{h}{2}\right)} \mathbb{P}\left(B_{\mathcal{S}}\left(q_i, \frac{h}{2}\right) \subseteq \Omega^h_{\mathcal{P}_n}\right)\right).
\end{aligned}$$

Using the lemmas above, we obtain

$$\mathbb{P}\left(\mathcal{S} \nsubseteq \Omega^h_{\mathcal{P}_n}\right) \leq \frac{(1 - x_h)^n}{x_h},$$

and we conclude by using the inequality $1 - x \leq e^{-x}$, valid for $x \geq 0$. ∎

It is both interesting and useful to find a relation between n (the number of points in the cloud), h (the radii of the balls), and k (the dimension of the manifold) which guarantees $\lim_{n\uparrow+\infty,\, h\downarrow 0} \mathbb{P}(\mathcal{S} \nsubseteq \Omega^h_{\mathcal{P}_n}) = 0$. For this purpose we will use Proposition 2.

Note that $h > 0$ will be small, and also, if we are attempting to approximate $d_{\mathcal{S}}$, h should tend to 0.[19]

Remark 6. Note that for $\{a_m\}_{m\in\mathbb{N}}$, $a_m \downarrow 0$, $\frac{e^{-ma_m}}{a_m}$ goes to zero as $m \uparrow \infty$ if a_m is asymptotically greater than or equal to $\frac{\log m}{m}$. Then, in order to have the right-hand side of (15) tend to zero, we should have $x_h \gtrsim \frac{\log n}{n}$, and the condition relating h, k, and n should then be[20]

$$h^k \gtrsim \left(\mu(\mathcal{S})\frac{2^k}{\omega_k}\right)\frac{\log n}{n}. \tag{16}$$

Also, under this condition we can estimate the rate at which $\frac{e^{-nx_h}}{x_h}$ approaches zero as $n \uparrow \infty$. For example, with $x_h \simeq \frac{\log n}{n}$, $\frac{e^{-nx_h}}{x_h} \simeq \frac{1}{\log n}$ as $n \uparrow \infty$. Note that, of course, we can speed up the convergence towards zero by choosing slower variations of x_{h_n} with n; for instance, with $x_{h_n} \simeq \frac{\log n^\gamma}{n}$, $\gamma \geq 1$, we have $\frac{e^{-nx_h}}{x_h} \simeq \frac{1}{\gamma(\log n)n^{\gamma-1}}$ as $n \uparrow \infty$. Bounds for $\mathbb{P}(\mathcal{S} \not\subseteq \Omega^h_{\mathcal{P}_n})$ similar to ours can be found in [27]. It can be seen that our bounds are better than the ones reported in [27] for a certain range of k, the dimension of $\mathcal{S}$. We should point out that with our bounds we can obtain rates of convergence comparable to the optimal ones. Let us elaborate on this: In the case of the unit circle S^1 it is known (see [66]) that

$$p_1(n,h) \stackrel{\triangle}{=} 2ne^{-n\frac{h}{\pi}} \simeq \mathbb{P}\left(S^1 \not\subseteq \Omega^h_{\mathcal{P}_n}\right) \tag{17}$$

for n large and $\frac{h}{\pi} \ll 1$, whereas our bound is $p_2(n,h) \stackrel{\triangle}{=} \frac{e^{-n}h/2\pi}{h/2\pi} \gtrsim \mathbb{P}(S^1 \not\subseteq \Omega^h_{\mathcal{P}_n})$. Choose for p_1, $h_n^{(1)} = \gamma_1\pi\frac{\log n}{n}$ and for p_2, $h_n^{(2)} = \gamma_2\pi\frac{\log n}{n}$. Plugging these expressions into the formulas for p_1 and p_2, we find $p_1 = 2n^{1-\gamma_1}$ and $p_2 = \frac{2}{\gamma_2(\log n)}n^{1-\frac{\gamma_2}{2}}$. Hence, by letting $(2 >)\gamma_2 = 2\gamma_1$ (which is equivalent to $\frac{h_n^{(2)}}{h_n^{(1)}} = 2$), we obtain $p_2 \lesssim p_1$. The *optimal* bound (17) for the case of S^1 is derived using direct knowledge of the distribution of the minimal number of random arcs (of a certain fixed size) needed to cover S^1 completely. This distribution is unknown for all nontrivial cases [66, 34]. In the case of the sphere S^2, also in [66], a bound of the type $\mathbb{P}(S^2 \not\subseteq \Omega^h_{\mathcal{P}_n} \leq CN^2e^{-DNh^2})$ is reported (for certain constants C and D); however, the proof seems to use properties of symmetry of the sphere in a fundamental way. Other interesting bounds which could be used in this situation are those in [38].

We should finally point out that the problem of covering a certain domain (usually S^1) with balls centered at random points sampled from this domain has been studied by many authors [66, 27, 28, 37, 65, 39, 34] and even by Shannon in [64].

We have the following interesting corollary, whose proof can be found in Appendix C.

COROLLARY 4. *Let $\mathcal{S}$ be a smooth compact submanifold of $\mathbb{R}^d$ without boundary. We have that if* (16) *holds, then for any $\varepsilon > 0$*

$$\lim_{h,n}\mathbb{P}\left(d_{\mathcal{H}}(\mathcal{S},\Omega^h_{\mathcal{P}_n}) > \varepsilon\right) = 0,$$

where $d_{\mathcal{H}}$ is the Hausdorff distance between sets.

[19] For constant $h > 0$, by definition $0 < x_h < 1$, and then obviously $\frac{e^{-nx_h}}{x_h} \to 0$ as $n \uparrow \infty$.

[20] This kind of condition is commonplace in the literature of random coverings; see, e.g., [25, 65, 22].

We are now ready to state and prove the following convergence theorem.

THEOREM 7. *Let $\mathcal{S}$ be a k-dimensional smooth compact submanifold of $\mathbb{R}^d$. Let $\mathcal{P}_n = \{p_1, \dots, p_n\} \subseteq \mathcal{S}$ be such that $p_i \sim \mathbf{U}[\mathcal{S}]$ for $1 \le i \le n$. Then if $h = h_n$ is such that $h_n \downarrow 0$ and (16) holds as $n \uparrow \infty$, we have that for any $\varepsilon > 0$,*

$$\mathbb{P}\left(\max_{p,q\in\mathcal{S}} \left(d_{\mathcal{S}}(p,q) - d_{\Omega^h_{\mathcal{P}_n}}(p,q)\right) > \varepsilon\right) \overset{n\uparrow\infty}{\longrightarrow} 0.$$

Proof. We base our proof on (7). We first note that $\mathbb{P}(\mathcal{E}_\varepsilon \mid \mathcal{I}_{h,n}) = 0$ for n large enough because, from considerations at the beginning of section 4, $\max_{p,q\in\mathcal{S}}(d_{\mathcal{S}}(p,q) - d_{\Omega^{h_n}_{\mathcal{P}_n}}(p,q)) \le C_{\mathcal{S}}\sqrt{h_n}$ whenever $\mathcal{S} \subseteq \Omega_{\mathcal{P}^{h_n}_n}$ holds. Let $N = N(\varepsilon) \in \mathbb{N}$ be such that $h_n < (\frac{\varepsilon}{C_{\mathcal{S}}})^2$ for all $n \ge N(\varepsilon)$. Then, for $n \ge N(\varepsilon)$, $\mathbb{P}(\mathcal{E}_\varepsilon) = \mathbb{P}\big(\mathcal{I}^c_{h,n} \le \frac{e^{-nx_{h_n}}}{x_{h_n}}\big)$, and since by assumption (16) holds, the right-hand side goes to 0 as $n \uparrow \infty$. ∎

Remark 7.

1. As can be gathered from the preceding proof, for fixed $\varepsilon > 0$ and large $n \in \mathbb{N}$,

$$\mathbb{P}\left(\max_{p,q\in\mathcal{S}} \left(d_{\mathcal{S}}(p,q) - d_{\Omega^h_{\mathcal{P}_n}}(p,q)\right) > \varepsilon\right)$$

can be upper bounded by $\frac{e^{-nx_{h_n}}}{x_{h_n}}$. For example, setting $x_{h_n} = \gamma\frac{\log n}{n}$ for $\gamma \ge 1$ yields (given n big enough)

$$\mathbb{P}\left(\max_{p,q\in\mathcal{S}} \left(d_{\mathcal{S}}(p,q) - d_{\Omega^h_{\mathcal{P}_n}}(p,q)\right) > \varepsilon\right) \le \frac{1}{\gamma\, n^{\gamma-1}\log n}.$$

2. Then we see that by requiring $\sum_{n\ge1} \frac{e^{-nx_{h_n}}}{x_{h_n}} < \infty$ and using the Borel–Cantelli lemma, we obtain *almost sure convergence*, namely,

$$\mathbb{P}\left(\lim_{n\uparrow\infty} \max_{p,q\in\mathcal{S}} \left(d_{\mathcal{S}}(p,q) - d_{\Omega^h_{\mathcal{P}_n}}(p,q)\right) = 0\right) = 1.$$

This can be guaranteed (for example) by setting $x_{h_n} = \gamma\frac{\log n}{n}$ for $\gamma > 2$.

This concludes our study of distance functions on (noiseless) point clouds (sampled manifolds). We now turn to the even more realistic scenario where the points are considered to be *noisy* samples.

6. Noisy sampling of manifolds. We assume that we have some uncertainty on the actual position of the surface, and we model this as if each point in the set of sampled points is modified by a (not yet random) perturbation of magnitude smaller than Δ. More explicitly, each p_i is given as $p_i = p + \zeta \times \vec{v}$ for some $\vec{v} \in S^{d-1}$, some p in $\mathcal{S}$, and $\Delta \ge \zeta \ge 0$. Then we can guarantee that the point p from which p_i comes can be found inside $B(p_i, \Delta) \cap \mathcal{S}$. We are again interested in comparing $d_{\Omega^h_{\mathcal{P}_n}} : \Omega^h_{\mathcal{P}_n} \to \mathbb{R}^+ \cup \{0\}$ with $d_{\mathcal{S}} : \mathcal{S} \to \mathbb{R}^+ \cup \{0\}$, but now these functions have different domains; therefore we must be careful in defining a meaningful way of relating them. If we consider

$$\mathcal{F}^\Delta_{\mathcal{S}} \triangleq \{f \mid f : \Omega^\Delta_{\mathcal{S}} \to \mathcal{S},\ f(p) \in B(p,\Delta) \cap \mathcal{S}\},$$

we can compare, for some $f \in \mathcal{F}^\Delta_{\mathcal{S}}$ and $1 \le i, j \le n$, $d_{\Omega^h_{\mathcal{P}_n}}(p_i, p_j)$ with $d_{\mathcal{S}}(f(p_i), f(p_j))$. Note that as the perturbation's magnitude goes to zero, $\mathcal{F}^\Delta_{\mathcal{S}} \ni f(p) \overset{\Delta\downarrow0}{\longrightarrow} p$, for $p \in \Omega^\Delta_{\mathcal{S}}$. The next step is to write $\max_{1\le i,j\le n} \|d^h_{\Omega_{\mathcal{P}_n}}(p_i,p_j) - d_{\mathcal{S}}(f(p_i), f(p_j))\|$, the biggest

error we have for our set of points. And finally, the next logical step is to look at the worst possible choice for f:

$$\mathcal{L}_{\mathbb{S}}(\mathcal{P}_n;\Delta,h) \triangleq \sup_{f\in\mathcal{F}_{\mathbb{S}}^{\Delta}} \max_{1\leq i,j\leq n} \left| d_{\mathbb{S}}(f(p_i),f(p_j)) - d_{\Omega^h_{\mathcal{P}_n}}(p_i,p_j)\right|. \tag{18}$$

We start by presenting deterministic bounds for the expression in (18), and only later will we be more (randomly) greedy and, in the spirit of Theorem 7, prove for $\varepsilon>0$ a result of the form ($\mathcal{L}_{\mathbb{S}}(\mathcal{P}_n;\Delta,h)$ will be a R.V.)

$$\mathbb{P}\left(\mathcal{L}_{\mathbb{S}}(\mathcal{P}_n;\Delta,h)>\varepsilon\right) \xrightarrow{n\uparrow\infty} 0.$$

6.1. Deterministic setting. The idea is to prove that for some convenient function $\widehat{f}\in\mathcal{F}_{\mathbb{S}}^{\Delta}$ we can write

$$\mathcal{L}_{\mathbb{S}}(\mathcal{P}_n;\Delta,h) \leq \max_{1\leq i,j\leq n} \left| d_{\mathbb{S}}(\widehat{f}(p_i),\widehat{f}(p_j)) - d_{\Omega^h_{\mathcal{P}_n}}(p_i,p_j)\right| + \lambda(h,\Delta),$$

where $0\leq\lambda(x,y)\xrightarrow{x,y\downarrow 0}0$. The natural candidate for $\widehat{f}$ is the orthogonal projection onto $\mathbb{S}$, $\Pi_{\mathbb{S}}:\Omega^H_{\mathbb{S}}\to\mathbb{S}$, whose properties are discussed in Appendix B. Then we see that we can reduce everything to bounding $\max_{p,q\in\mathbb{S}}\|d_{\mathbb{S}}(p,q)-d_{\Omega^h_{\mathcal{P}_n}}(p,q)\|$. This is simple since if $\mathcal{P}_n\subset\Omega^{\Delta}_{\mathbb{S}}$, then $\Omega^h_{\mathcal{P}_n}\subset\Omega^{h+\Delta}_{\mathbb{S}}$, and $d_{\mathbb{S}}\geq d_{\Omega^h_{\mathcal{P}_n}|_{\mathbb{S}}}\geq d_{\Omega^{h+\Delta}_{\mathbb{S}}|_{\mathbb{S}}}$, and finally from Theorem 5, $\|d_{\mathbb{S}}-d_{\Omega^h_{\mathcal{P}_n}}\|_{L^\infty(\mathbb{S})}\leq C_{\mathbb{S}}\sqrt{h+\Delta}$.

Let $\mathbb{S}\subset\Omega^h_{\mathcal{P}_n}$, $f\in\mathcal{F}^{\Delta}_{\mathbb{S}}$, and $1\leq i,j\leq n$. Then, after using the triangle inequality a number of times, we can write the bound

$$\begin{aligned}\left| d_{\mathbb{S}}(f(p_i),f(p_j)) - d_{\Omega^h_{\mathcal{P}_n}}(p_i,p_j)\right| \leq\ & 2\sup_{f\in\mathcal{F}^{\Delta}_{\mathbb{S}}}\max_{p\in\mathcal{P}_n} d_{\mathbb{S}}(f(p),\Pi_{\mathbb{S}}(p))\\ &+\max_{p,q\in\mathbb{S}}\left| d_{\mathbb{S}}(p,q)-d_{\Omega^h_{\mathcal{P}_n}}(p,q)\right|\\ &+\max_{p,q\in\mathcal{P}_n}\left| d_{\Omega^h_{\mathcal{P}_n}}(p,q)-d_{\Omega^h_{\mathcal{P}_n}}(\Pi_{\mathbb{S}}(p),\Pi_{\mathbb{S}}(q))\right|.\end{aligned}$$

The last term can be bounded by 2Δ, the one in the middle has already been discussed, and hence we are left with the first one. Using Corollary 2, we find that since $f(p)\in B(\Pi_{\mathbb{S}}(p),2\Delta)\cap\mathbb{S}$, then in fact $f(p)\in B_{\mathbb{S}}(\Pi_{\mathbb{S}}(p),2\Delta(1+C_{\mathbb{S}}\sqrt{\Delta}))$ and $d_{\mathbb{S}}(f(p),\Pi_{\mathbb{S}}(p))\leq 2\Delta(1+C_{\mathbb{S}}\sqrt{2}\sqrt{\Delta})$. Summing up, under the condition $\mathbb{S}\subset\Omega^h_{\mathcal{P}_n}$, we obtain the desired result,

$$\mathcal{L}_{\mathbb{S}}(\mathcal{P}_n;\Delta,h)\leq C_{\mathbb{S}}\sqrt{h+\Delta}+2\Delta(2+\sqrt{2}C_{\mathbb{S}}\sqrt{\Delta}). \tag{19}$$

6.2. Random setting. Assume that $\{p_1,\ldots,p_n\}$ is a set of i.i.d. random points such that each $p_i\sim\mathbf{U}[\Omega^{\Delta}_{\mathbb{S}}]$. At this time, we want to estimate the probability of having $\mathbb{S}\subseteq\Omega^h_{\mathcal{P}_n}$. It is easy to see that as a first "reality compliant" condition one should have that the noise level not be too big with respect to h. We will impose $h\geq\Delta$ for simplicity's sake, as can be understood from the convergence theorem below. Since the techniques are similar to those used in the noise-free case, we will present its proof in Appendix C.

THEOREM 8. *Let $\mathbb{S}$ be a k-dimensional smooth compact submanifold of $\mathbb{R}^d$. Let $\mathcal{P}_n=\{p_1,\ldots,p_n\}$ be such that $p_i\sim\mathbf{U}[\Omega^{\Delta}_{\mathbb{S}}]$ for $1\leq i\leq n$. Then if $h=h_n$, $\Delta=\Delta_n$*

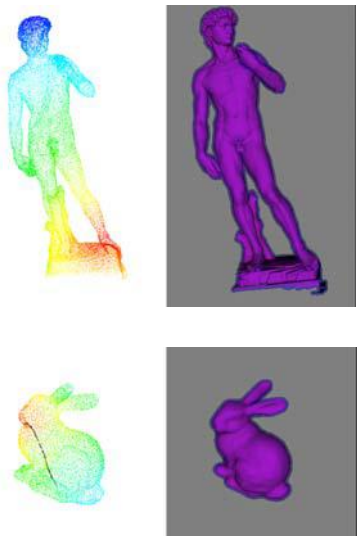


FIG. 1. *Intrinsic distance function for a point cloud. A point is selected in the head of the David, and the intrinsic distance is computed following the framework introduced here. The point cloud is colored according to the intrinsic distance to the selected point, going from bright red (far) to dark blue (close). The offset band, given by the union of balls, is shown next to the distance figure. Bottom: Same as before, with a geodesic curve between two selected points.*

are such that $\Delta_n \leq h_n$ *and* $h_n \downarrow 0$ *and* $\Delta_n^k \gtrsim \frac{\log n}{n}$ *as* $n \uparrow \infty$*, we have that for any* $\varepsilon > 0$*,*

$$\mathbb{P}\left(\mathcal{L}_{\mathbb{S}}(\mathcal{P}_n; \Delta, h) > \varepsilon\right) \xrightarrow{n\uparrow\infty} 0.$$

We have now concluded the analysis of the most general case for noisy sampling of manifolds. Note that, although the results in this and in previous sections were presented for Euclidean balls, they can easily be extended to more general covering shapes (check Corollary 1 above), e.g., following [15, 36], or using minimal spanning trees, or from the local directions of the data [56]. In addition, the recently developed approach reported in [57] can be used for defining the offset band in an adaptive fashion. This will improve the bounds reported here. Similarly, the results can be extended to other sampling or noise models following the same techniques developed here.

7. Implementation details and examples. We now present examples of distance matrices and geodesics for point clouds (Figure 1), use these computations to find intrinsic Voronoi diagrams (Figure 2; see also [42, 43, 71]); and compare the results with those obtained with mesh-based techniques (Figure 3).[21] We also present examples in high dimensions and use, following and extending [24], our results to compare manifolds given by point clouds. All these exercises are to exemplify the importance of computing distance functions and geodesics on point clouds, and are by no means exhaustive. The 3D data sets used come from real point cloud data and have been obtained either from range scanners (*David* model) or via photometric stereo techniques (man and woman).

[21] All the figures in this paper are in color. VRML files corresponding to these examples can be found at http://mountains.ece.umn.edu/∼guille/pc.htm.

FIG. 2. *Voronoi diagram for point clouds. Four points (left) and two points (right) are selected on the cloud, and the point cloud is divided (colored) according to the geodesic distance to these four points. Note that this is a surface Voronoi, based on geodesics computed with our proposed framework, not a Euclidean one.*

The theoretical results presented in the previous sections show that the intrinsic distance and geodesics can be approximated by the Euclidean ones computed in the band defined (for example) by the union of balls centered at the points of the cloud. The problem is then simplified to first computing this band (no need for mesh computation, of course), and then using well-known computationally optimal techniques to compute the distances and geodesics inside this band, exactly as done in [49] for implicit surfaces (where the interested reader can also find explicit computational timings and accuracy comparisons with mesh-based approaches). The band itself can be computed in several ways, and for the examples below we have used constant radii. Locally adaptive radii can be used, based, for example, on diameters obtained from minimal spanning trees or on the recent work reported in [57]. Automatic and local estimation of h defining $\Omega^h_{\mathcal{P}_n}$, which will improve the bounds reported here, was not pursued in this paper and is the subject of current implementation efforts.

The software implementation of the algorithm is based on using the fast Euclidean distance computation algorithms, usually referred to as *fast marching* algorithms [35, 62, 63, 69], twice. We omit the description of this algorithm since it is well known. The starting point is defining a grid over which all the computations are performed. This amounts to choosing Δ_{x_i}, the grid spacing in each direction $i = 1, \ldots, d$, which will determine the accuracy of the numerical implementation (the offset band includes fewer than 10 grid points).[22] In the first round we compute the band $\Omega^h_{\mathcal{P}_n} = \{x \in \mathbb{R}^d : d(\mathcal{P}_n, x) \leq h\}$ by specifying a value of zero for the function $\Psi(x) = d(\mathcal{P}_n, x)$ on the points $x \in \mathcal{P}_n$. Since in general these points will not be on the grid, we use a simple multilinear interpolation procedure to specify the values on neighboring grid points. The second use of the fast distance algorithm is also simply reduced to using Ψ to define $\Omega^h_{\mathcal{P}_n}$ by using the simple modification reported in [49]. The computation of geodesics was done using a simple Runge–Kutta gradient descent procedure, much in the way described in [49], with some obvious modifications.

[22] Adaptive grids inside the fixed or variable width offset band could be used as well; see, for example, [30].

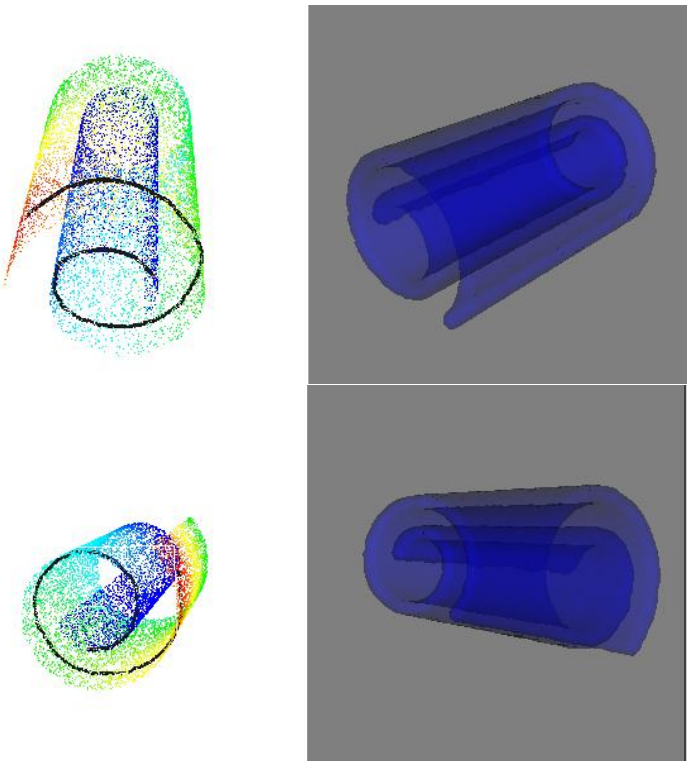


FIG. 3. *Examples of geodesic computations. This data is used to study the algorithm robustness to noise, see Appendix A.*

All the code and 3D visualization was developed in C++ using both Flujos (which is written using Blitz++; see [7]) and VTK (see [70]). For matrix manipulation and visualization of other results we used MATLAB. We are currently working on a more advanced implementation of the proposed framework that permits us to work with high-dimensional data without having the memory allocation problems that result from blind and straightforward allocation of resources to empty and nonused grids.

7.1. High-dimensional data. In this section we present a simple example for high-dimensional data. We embed a circle of radius 15 in $\mathbb{R}^5$, and use a grid of size $34 \times 4 \times 4 \times 4 \times 34$ (with uniform spacing $\Delta x = 1$) such that each of the sample points is of the form $p_i = 15\,(\cos(\frac{2\pi i}{N}), 0, 0, 0, \sin(\frac{2\pi i}{N})) + (17, 2, 2, 2, 17)$, for $1 \leq i \leq N$. We then use our approach to compute the (approximate) distance function d_h in a band in $\mathbb{R}^5$, and then the error $e_{ij} = |d_{\mathcal{S}}(p_i, p_j) - d_h(p_i, p_j)|$ for $i, j \in \{1, \ldots, N\}$. In our experiments we used $h = 2.5 > \Delta x\sqrt{5}$.[23] We randomly sampled 500 points from the $N = 1000$ points used to construct the union of balls to build the 500×500 error matrix $((e_{ij}))$. We found $\max_{ij}\{e_{ij}\} = 2.0275$, that is, a 4.3% L_∞-error. In Figure 4 we show the histogram of all the (500^2) entries of $((e_{ij}))$. We should also note that when following the dimensionality reduction approach in [67], with the geodesic distance computation proposed here, the correct dimensionality of the circle was obtained.

[23] For a discussion on how to make a preliminary estimation of the value of h, see [49].

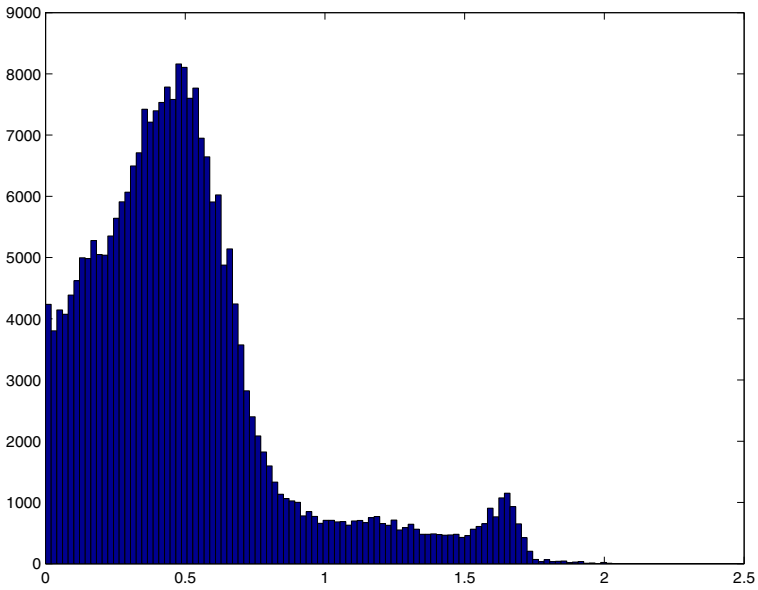


FIG. 4. *Histogram for the error in the case of a circle embedded in* $\mathbb{R}^5$.

In high dimensions, when the grid is too large, our current numerical implementation becomes unusable. The problem stems from the fact that we require too much memory space, most of which is not really used, since the computations are conducted only in a band around $\mathcal{P} \subset \mathbb{R}^d$. To be more precise, the memory requirements of our current direct implementation, which uses a d-dimensional array to make the computations, are $\simeq (\max_i l_i)^d$, whereas we really need a storage capacity of order $\mu_k(\mathcal{S})h^{d-k}$, where l_i is the size of $\mathcal{P}$'s bounding box along the ith direction, $1 \leq i \leq d$, and $\mu_k(\mathcal{S})$ is the measure of the k-dimensional manifold $\mathcal{S}$ (embedded in $\mathbb{R}^d$). This memory problem is to be addressed by a computation that is not based on discretizing the whole band. (Note, of course, that the theoretical foundations presented in this paper are independent of the particular implementation.) We are currently working on addressing this specific issue.

7.2. Object recognition. The goal of this application is to use our framework to compare manifolds given by point clouds. The comparison is done in an intrinsic way, that is, isometrically (bending) invariant. This application is motivated by [24], where they use geodesic distances (computed using a graph-based approach) to compare 3D triangulated surfaces. In contrast with [24], we compare point clouds using our framework (which is not only based in the original raw data, but also, as shown in Appendix A, more robust to noise than mesh approaches such as those of [24] and is valid in any dimensions), and use a different procedure/similarity metric between the manifolds. The authors in [24] basically project into low-dimensional manifolds and use eigenvalues and eigenvectors of a centralized matrix related to the *distance matrices* (matrices which in each entry (i, j) have the value of the intrinsic distance between (projected) points p_i and p_j of the cloud), which are clearly not sufficient to distinguish nonisometric objects. (Nonisometric objects can have distance matrices with the same eigenvalues.) A different study, based on direct comparisons of distance matrices, is used here and detailed in Appendix E.

TABLE 1
Information about the models used in our recognition experiments.

Dataset	Number of points in the cloud (n)	Grid size used
Bunny	15862	$80 \times 80 \times 70$
MAN2	26186	$120 \times 90 \times 200$
MAN3	26186	$120 \times 90 \times 165$
MAN5	26186	$120 \times 85 \times 160$
WOMAN2	29624	$120 \times 105 \times 175$
WOMAN3	29624	$120 \times 100 \times 180$

Our task then is to compare two manifolds in an intrinsic way; i.e., we want to check whether they are isometric or not. We want to check this condition by using point clouds representing each one of the manifolds. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two submanifolds of $\mathbb{R}^d$ and sample on each of them the two point clouds $\mathcal{P}_n^{(1)} \subset \mathcal{S}_1$ and $\mathcal{P}_n^{(2)} \subset \mathcal{S}_2$. Then, following our theory, we compute the corresponding distances in the offset bands for these two sets of points, $d_{\Omega^{h_1}_{\mathcal{P}_n^{(1)}}}$ and $d_{\Omega^{h_2}_{\mathcal{P}_n^{(2)}}}$, and for point subsets $\{q_1^{(1)}, \ldots, q_m^{(1)}\} = \mathcal{Q}_m^{(1)} \subseteq \mathcal{P}_n^{(1)}$, $\{q_1^{(2)}, \ldots, q_m^{(2)}\} = \mathcal{Q}_m^{(2)} \subseteq \mathcal{P}_n^{(2)}$ we compute the corresponding $m \times m$ pairwise distance matrices (as defined above)

$$D_1 = \left(\left(d_{\Omega^{h_1}_{\mathcal{P}_n^{(1)}}} (q_i^{(1)}, q_j^{(1)}) \right) \right) \quad \text{and} \quad D_2 = \left(\left(d_{\Omega^{h_2}_{\mathcal{P}_n^{(2)}}} (q_i^{(2)}, q_j^{(2)}) \right) \right).$$

Let $\mathcal{PM}_m$ be the set of $m \times m$ permutation matrices and $\|\cdot\|$ a unitary transformation invariant norm[24] (fix the Frobenius norm: $\|A\| = \sqrt{\sum_i \sum_j a_{ij}^2}$). Then we define the $\mathcal{J}$-distance between (distance) matrices D_1 and D_2 as

$$d_{\mathcal{J}}(D_1, D_2) \stackrel{\triangle}{=} \min_{P \in \mathcal{PM}_m} \|D_1 - P D_2 P^T\|.$$

Clearly, if $d_{\mathcal{J}}(D_1, D_2) = 0$, then we have an isometry between the discrete metric sets $(\mathcal{Q}_n^{(1)}, d_{\Omega^{h_1}_{\mathcal{P}_n^{(1)}}})$ and $(\mathcal{Q}_n^{(2)}, d_{\Omega^{h_2}_{\mathcal{P}_n^{(2)}}})$. This should allow us to establish a *rough isometry* (see [14, section 4.4]) between $\mathcal{S}_1$ and $\mathcal{S}_2$ with interesting constants.

The exact details on how this metric is approximated and how the subsets of points $\mathcal{Q}$ are selected is presented in Appendix E. For the experiments regarding recognition of shapes we used the datasets listed in Table 1.

In Figure 5 we present the histogram of the error $e(100)/100$ for 20 different 100×100 distance matrices corresponding to the full Bunny model, with the 100 points chosen as in the "packing procedure" described in Appendix E, where the exact definition of $e(\cdot)$ is also given (see (28)). We computed the mean of $e(100)/100$ over the $19 \times 18 \times \cdots \times 1 = 190$ comparison experiments to be 0.4774 with standard deviation 0.0189. This can be interpreted as indicating that when one considers a large enough set of points, the information contained in the packing set is representative of the metric information of the manifold, independently of the particular choice of the packing set. This claim needs some further theoretical justification, which could come if a result of the following fashion were proved:[25]

[24] $\|AU\| = \|A\|$ for any matrix A and any *unitary* matrix U.

[25] Note added in proof: After this paper was submitted for publication, we proved that a properly modified version of the above claim holds *in probability*; see [50] for details.

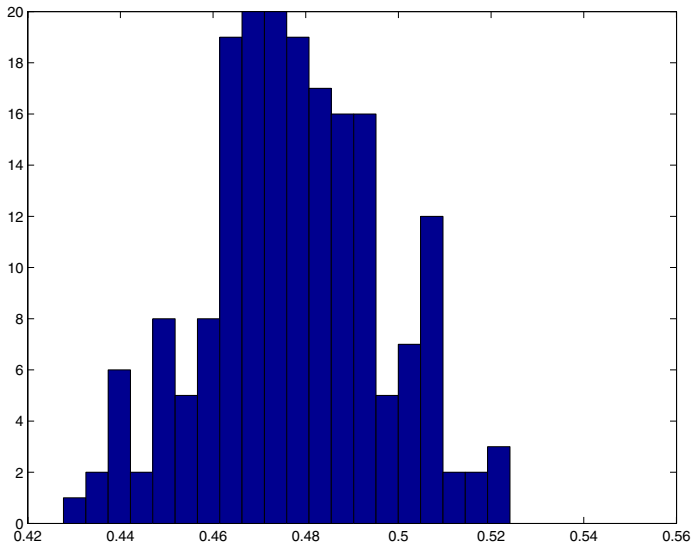


FIG. 5. *Histogram showing the errors for different selections of point clouds on the bunny model.*

Let $\mathcal{S}$ be a smooth compact k-dimensional submanifold of $\mathbb{R}^k$ such that its Ricci curvature is bounded below by $\kappa(n-1)$ with $\kappa \leq 0$. Let $\mathcal{Q}_m^{(r)} \subset \mathcal{S}$, $r = 1, 2$, be such that $d_{\mathcal{S}}(q_i^{(r)}, q_j^{(r)}) \geq \varepsilon$ and $B_{\mathcal{S}}(\mathcal{Q}_m^{(r)}, R)$ covers $\mathcal{S}$ for some $R > \varepsilon > 0$. Then, with D_1 and D_2 defined as before,

$$d_{\mathcal{J}}(D_1, D_2) \leq 2mC_{\mathcal{S}}\sqrt{h} + C(R, \varepsilon, m),$$

where the exact form of $C(R, \varepsilon, m)$ is to be determined, leading to an optimal choice of m (the size of the subset).

Using the same procedure, described in Appendix E, to choose the sets $\mathcal{Q}_m^{(i)}$, we computed the errors (according to $e(D_1, D_2)$) for five artificial human models; three of them are bendings of a man and two are bendings of a woman; see Figures 6 and 7. Details on these models are also given in Table 1. The results of this cross-comparison are presented in Table 2 below.

TABLE 2

Cross-comparisons for the human models using the error measure $e(300)/300$ normalized by the maximum of the errors.

MODEL	Man2	Man3	Man5	Woman2	Woman3
Man2	$*$	0.0514	0.0570	0.4690	0.4853
Man3	$*$	$*$	0.0206	0.4701	0.4859
Man5	$*$	$*$	$*$	0.4702	0.4862
Woman2	$*$	$*$	$*$	$*$	0.2639
Woman3	$*$	$*$	$*$	$*$	$*$

These examples show how our geodesic distance computation technique, when complemented with the matrix metric in Appendix E, can be used to compare manifolds given by point clouds, in a bending-invariant fashion and without explicit manifold reconstruction. More exhaustive experimentation and additional theoretical justification will be reported elsewhere.

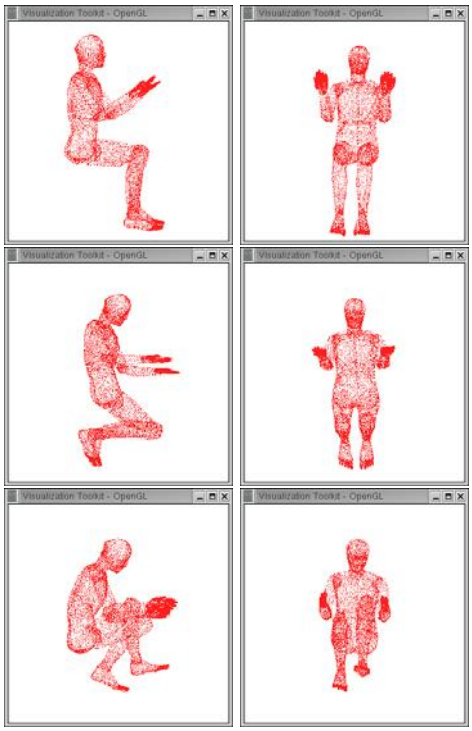


FIG. 6. *MAN models. From top to bottom (two views of each model): MAN*2*, MAN*3*, and MAN*5*.*

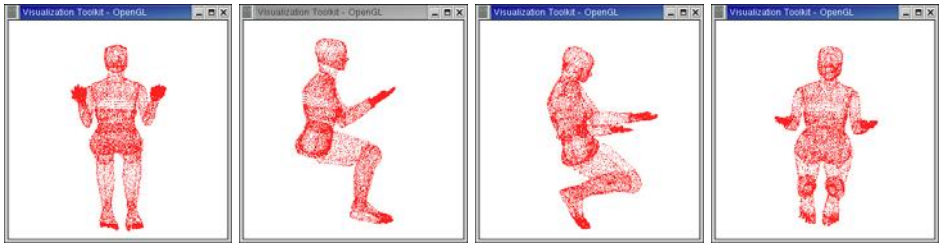


FIG. 7. *WOMAN models. From left to right (two views of each model): WOMAN*2 *and WOMAN*3*.*

Before concluding, we should comment that, as frequently done in the literature, we could normalize the geodesic distances if scale invariance were also required. Moreover, we could also consider in the distance matrix only nonzero entries for local neighborhoods. In addition, the use of techniques for computing eigenvalues and eigenvectors such as those in the work of Coifman and colleagues [17], on high-dimensional geometric multiscale analysis should be explored.

8. Concluding remarks. In this paper, we have extended our previous work [49] to deal with (smooth) submanifolds of $\mathbb{R}^d$ (of any codimension) and possibly with boundary, and using these extensions, we have also shown how to compute intrinsic distance functions on a generic manifold defined by a point cloud, without the intermediate step of manifold reconstruction. The basic idea is to use well-developed computational algorithms for computing Euclidean distances in an offset band sur-

rounding the manifold, to approximate the intrinsic distance. The underlying theoretical results were complemented by experimental illustrations.

As mentioned in the introduction, an alternative technique for computing geodesic distances was introduced in [9, 67] (see also [31]). In contrast with our work, the effects of noise were not addressed in [9, 31]. Moreover, as one can see from considerations in Appendix A, our framework seems to be more robust to noise. We should note that the memory requirements of the current way of implementing our framework are large, and this needs to be addressed for very high dimensions (the framework is, of course, still valid). In particular, we are interested in direct ways of computing distances inside regions defined by union of balls, without the need to use the Hamilton–Jacobi approach. Several classical computer science implementation tricks can be applied to avoid this memory allocation problem, and this is part of our current implementation efforts.

We are currently working on the use of this framework to create multiresolution representations of point clouds (in collaboration with C. Moenning and N. Dyn; see [55] and also [11, 18, 20, 58]), to further perform object recognition for larger libraries, and to compute basic geometric characteristics of the underlying manifold—all this, of course, without reconstructing the manifold. (See [54] for recent results on normal computations for 2D and 3D noisy point clouds.) Some results in these directions are reported in [50, 55]. Further applications of our framework for high-dimensional data are also currently being addressed, beyond the preliminary (toy) results reported in section 7. Of particular interest in this direction is the combination of this work with the one developed by Coifman and colleagues and the recent one in [31].

Appendix A. Comparison with mesh-based strategies for distance calculation in the presence of noise. We now make some very basic comparisons between our approach to geodesic distance computations and those based on graph approximations to the manifold, such as the one in Isomap [67, 31].[26] (Comparisons of the band framework with the one reported in [40] for 3D triangulated surfaces are reported in [49].) The goal is to show that such graph-based techniques are more sensitive to noise in the point cloud sample (and the error can even increase to infinity with the increase in the number of points). This is expected, since the geodesic in such techniques goes through the noisy samples, while in our approach, they just go through the union of balls. We make our argument only for the $1D$ case, while the high-dimensional cases can be similarly studied.

A.1. 1D theoretical case. Let us consider a rectilinear segment of length L and $n+1$ equispaced points $p_1, \ldots, p_{n+1}$ in that segment. Consider the *noisy* points $q_i = p_i + \zeta_i \vec{n}$, where $\vec{n}$ is the normal to the segment and ζ_i $1 \leq i \leq n$ are independent R.V. uniformly distributed in $[-\Delta, \Delta]$. Let $l = L/n$ denote the distance between adjacent p_i's. Let d_g^Δ denote the length of the polygonal path $\overline{q_1 q_2 \ldots q_{n+1}}$ and $d_0 = L$. Then obviously $d_g^\Delta \geq d_0$ for any realization of the R.V.'s ζ_i. Let $d_i = \|p_i - p_{i+1}\|$; then by Pythagoras theorem $d_i = \sqrt{l^2 + z_i^2}$, where $z_i = \zeta_i - \zeta_{i+1}$ are R.V.'s with triangular density in $[-2\Delta, 2\Delta]$.

Next we compute $\mathbb{E}(d_i) = \frac{1}{2\Delta} \int_{-2\Delta}^{2\Delta} \sqrt{l^2 + z^2}(1 - \frac{|z|}{2\Delta})\, dz$. The result is

$$\mathbb{E}(d_i) = \sqrt{l^2 + 4\Delta^2} + \frac{l^2}{2\Delta} \log\left(\frac{2\Delta + \sqrt{l^2 + 4\Delta^2}}{l}\right) - \frac{1}{6\Delta^2}\left((l^2 + 4\Delta^2)^{3/2} - l^3\right).$$

[26] Isomap builds a mesh by locally connecting the (noisy) samples.

TABLE 3
Results of simulations with the Swiss Roll dataset.

Noise power (n_k^2)	$\max_{ij}\lvert D_{ij}^{g,n_k} - D_{ij}^{g,0}\rvert$	k	$\max_{ij}\lvert D_{ij}^{h,n_k} - D_{ij}^{h,0}\rvert$	h
0.0001	2.5222	7	0.5266	1.8
0.01	4.6409	7	0.9430	1.8
0.04	5.1737	7	1.2489	1.8
0.09	5.3292	7	1.4682	1.8
0.16	5.4651	7	1.7965	1.8

Now assuming $\frac{\Delta}{l} \ll 1$, we find that up to first order $\mathbb{E}(d_i) \simeq l + \Delta$ and

$$\mathbb{E}\left(d_g^\Delta - d_0\right) \simeq n\Delta.$$

From this we also get[27]

$$p_g \triangleq \mathbb{P}\left(d_g^\Delta - d_0 > \varepsilon\right) \lesssim \frac{n\Delta}{\varepsilon}.$$

On the other hand, for our approximation d_h^Δ, if the segment is contained in the union of the balls centered at the sampling points, $d_h^\Delta = d_0$. The probability of covering the segment by the band can be made arbitrarily close to 1 by increasing n. More precisely, one can prove that if p stands for the value of the probability of *not covering* the segment, then $p \leq k\frac{L}{\Delta}(1 - k'\frac{\Delta}{L})^n$, for some positive constants k and k'. Then we can write

$$p_h \triangleq \mathbb{P}\left(d_h^\Delta - d_0 > \varepsilon\right) \leq \frac{k''}{\varepsilon}\frac{L}{\Delta}\left(1 - k'\frac{\Delta}{L}\right)^{n+1}.$$

The comparison is now easy. We see that in order to have p_g vanish as $n \uparrow \infty$, Δ must go to zero *faster* than $\frac{1}{n}$. However, we know that by requiring $\Delta \simeq \frac{\log n}{n} \gtrsim \frac{1}{n}$ we have $p_h \downarrow 0$ as $n \uparrow \infty$. This means that the graph approximation of the distance is more sensitive to noise than ours.[28] This gives some evidence about why our approach is more robust than popular mesh-based ones. Next we present results of some simulations carried out in order to further verify our claim.

A.2. Simulations. In Table 3 we present results of simulations carried out for the SwissRoll dataset [67]; see Figure 3. We used 10,000 points to define the manifold. We then generated 10,000 noise vectors, each component being uniform with power one and zero mean. Then we generated noisy datasets from the noiseless SwissRoll dataset by adding the noise vector times a constant n_k to each vector of the noiseless initial dataset. We then chose 1000 corresponding points in each dataset and computed the intrinsic pairwise distance approximation, obtaining the matrices $\{(D_{ij}^{g,n_k})\}$ and $\{(D_{ij}^{h,n_k})\}$ for the graph-based and our approach, respectively, where $k = 1, 2, \ldots, 5$, $i, j \in [1, 1000]$, and n_k denotes the noise level. We then computed the values of $\max_{ij}\lvert D_{ij}^{g,n_k} - D_{ij}^{g,0}\rvert$ and $\max_{ij}\lvert D_{ij}^{h,n_k} - D_{ij}^{h,0}\rvert$ for each k, where $D_{ij}^{g,0}$ and $D_{ij}^{h,0}$ stand for noiseless intrinsic distance approximations. In Table 3, h indicates the radii and k the size of the neighborhood for Isomap. The graph approximation shows

[27] Also, with similar arguments we can prove that $\max_{\zeta_1,\ldots,\zeta_{n+1}}(d_g^\Delta - d_0) \simeq \frac{2n^2\Delta^2}{L}$.

[28] Another way of seeing this is by noting that, for a fixed noise level Δ, by increasing n we actually worsen the graph approximation, whereas we are making our approximation better.

less robustness to noise than our method, as was argued above. This is also true for the sensitivity,[29] where our approach outperforms the graph-based one by at least one order of magnitude. Note that the sensitivity for our approach can be formally studied from Theorem 3.

Appendix B. Properties of Euclidean distance functions. The references for this section are [2, pp. 12–16], and [26].

THEOREM 9 (see [2]). *Let $\Gamma \subset \mathbb{R}^d$ be a compact, smooth manifold without boundary. Then $\eta(x) \stackrel{\triangle}{=} \frac{1}{2}d^2(\Gamma, x)$ is smooth in a tubular neighborhood U of Γ. Also, in U it satisfies $\|D\eta\|^2 = 2\eta$.*

COROLLARY 5. *The projection operator $\Pi : U \to \Gamma$, for a given $x \in U$, can be written as $\Pi(x) = x - D\eta(x)$. Moreover, this operator is smooth.*

Remark 8. Differentiation of the relation $\langle D\eta, D\eta\rangle = 2\eta$ gives us $D^2\eta D\eta = D\eta$. Differentiating once more, we also find $D^3\eta D\eta = D^2\eta$.

THEOREM 10 (see [2]). *Let Γ and U be as in Theorem 9, and let $y \in U$ and $x = y - D\eta(x) \in \Gamma$, $k = \dim(\Gamma)$. Then, denoting by $\lambda_1, \ldots, \lambda_n$ the eigenvalues of $D^2\eta(y)$,*

$$\lambda_i(y) = \begin{cases} \frac{d(\Gamma,y)\kappa_i(x)}{1+d(\Gamma,y)\kappa_i(x)} & \text{if } 1 \le i \le k, \\ 1 & \text{if } k < i \le n, \end{cases}$$

where $\kappa_i(x)$ are the principal curvatures of Γ at x along $Dd(\Gamma, y) \in N_x\Gamma$, where $N_x\Gamma$ is the normal space to Γ at x.

Appendix C. Deferred proofs.

Proof of Corollary 3. We present only a sketch of the proof. Let $\mathcal{M}$ be an extension of $\mathcal{S}$ such that $\mathcal{S}$ is still strongly convex in $\mathcal{M}$, and let $0 < \delta \stackrel{\triangle}{=} \min_{x\in\mathcal{S}} \min_{z\in\mathcal{M}} \|x - z\|$. Then, $\overline{B(x,\alpha)} \cap \overline{B(z,\beta)} = \emptyset$ for all $x \in \mathcal{S}$, $z \in \partial\mathcal{M}$, and $\alpha, \beta < \frac{\delta}{3}$. Hence, $\Omega^\alpha_{\mathcal{S}} \cap \Omega^\beta_{\partial\mathcal{M}} = \emptyset$ for $\alpha, \beta \le \frac{\delta}{3}$.

For any $x, y \in \mathcal{S}$ consider γ_h the $\Omega^h_{\mathcal{M}}$-minimizing geodesic, $\mathbf{L}(\gamma_h) = d_{\Omega^h_{\mathcal{M}}}(x, y)$.

By the convexity of $\mathcal{S}$ there exists a unique $\mathcal{M}$-minimizing geodesic $\gamma_0 \subset \mathcal{S}$ joining x, y, and then, by Theorem 4, γ_h uniformly converges to γ_0. In particular, for any $\epsilon > 0$ there exists $h_\epsilon > 0$ such that $\gamma_h \subset \Omega^\epsilon_{\gamma_0}$ for all $h < h_\epsilon$. Choose $\epsilon \le \frac{\delta}{3}$; then $\gamma_h \subset \Omega^\epsilon_{\gamma_0} \subset \Omega^\epsilon_{\mathcal{S}}$. Furthermore, if $h \le \frac{\delta}{3}$, then $\Omega^\epsilon_{\gamma_0} \cap \Omega^h_{\mathcal{M}} = \emptyset$, and therefore γ_h does not touch $\partial\Omega^h_{\mathcal{M}} \cap \partial\Omega^h_{\partial\mathcal{M}}$. Thus, γ_h is $C^{1,1}$ for $h \le \frac{\delta}{3}$. Note that with this choice of h we have $\Omega^h_{\mathcal{S}} \cap \mathcal{M} \subset \mathrm{int}(\mathcal{M})$, and therefore we also have a smooth orthogonal projection operator $\Pi : \Omega^h_{\mathcal{S}} \to \mathcal{M}$.

Proceeding as in the first steps of the proof of Theorem 5, we have $\mathbf{L}(\gamma_h) = d_{\Omega^h_{\mathcal{M}}}(x, y) \le d_{\mathcal{M}}(x, y) \le \mathbf{L}(\Pi(\gamma_h))$, since $\Pi(\gamma_h) \subset \mathcal{M}$ but may not be a minimizing path. Then, using the convexity of $\mathcal{S}$ in $\mathcal{M}$, $d_{\mathcal{M}}(x, y) = d_{\mathcal{S}}(x, y)$, and therefore $0 \le d_{\mathcal{S}}(x, y) - d_{\Omega^h_{\mathcal{M}}}(x, y) \le |\mathbf{L}(\Pi\gamma_h) - \mathbf{L}(\gamma_h)|$, which can be bounded by a constant times $\sqrt{h}$ just mimicking the proof of Theorem 5. We conclude by noting that $\Omega^h_{\mathcal{S}} \subset \Omega^h_{\mathcal{M}}$, and hence $d_{\mathcal{S}}(x, y) - d_{\Omega^h_{\mathcal{M}}}(x, y) \ge d_{\mathcal{S}}(x, y) - d_{\Omega^h_{\mathcal{S}}}(x, y)$. ∎

Proof of Lemma 3. We now estimate the covering number $\mathcal{N}(\mathcal{S}, \delta)$. The idea is constructive, very simple, and of course standard. We consider the following procedure (adopted from [9]): Let q_1 be any point in $\mathcal{S}$, and choose $q_2 \in \mathcal{S}\backslash B_{\mathcal{S}}(q_1, \delta)$. Then choose $q_3 \in \mathcal{S}\backslash\{B_{\mathcal{S}}(q_1, \delta) \cup B_{\mathcal{S}}(q_2, \delta)\}$. Iterate this procedure until it is no longer possible

[29] Sensitivity is defined as $\left|1 - \frac{\text{distance for noisy points}}{\text{distance for clean points}}\right|$.

to choose any point $q \in \mathcal{S}\backslash\{\cup_{k=1}^{\mathcal{N}(\mathcal{S},\delta)} B_{\mathcal{S}}(q_k,\delta)\}$; in such a case $\mathcal{S} = \cup_{k=1}^{\mathcal{N}(\mathcal{S},\delta)} B_{\mathcal{S}}(q_k,\delta)$. Note that $B_{\mathcal{S}}(q_k,\frac{\delta}{2}) \cap B_{\mathcal{S}}(q_l,\frac{\delta}{2}) = \emptyset$ if $k \neq l$, and therefore we can bound $\mathcal{N}(\mathcal{S},\delta) \leq \frac{\mu(\mathcal{S})}{\min_{x\in\mathcal{S}} \mu(B_{\mathcal{S}}(x,\delta/2))}$. Therefore, using the Bishop–Günther inequalities in the same manner as in Lemma 1, we find (14). ∎

Proof of Corollary 4. Note first that the random variable $d_{\mathcal{H}}(\mathcal{S},\Omega^h_{\mathcal{P}_n})$ is bounded by $\max\{\text{diam }(\mathcal{S}) + h, h\}$. By definition of the Haussdorf distance, $d_{\mathcal{H}}(\mathcal{S},\Omega^h_{\mathcal{P}_n}) = \max\left(\sup_{x\in\mathcal{S}} d(x,\Omega^h_{\mathcal{P}}), \sup_{y\in\Omega^h_{\mathcal{P}_n}} d(y,\mathcal{S})\right)$. Then, $\sup_{x\in\mathcal{S}} d(x,\Omega^h_{\mathcal{P}_n}) \leq \text{diam }(\mathcal{S}) + h$ by the triangle inequality, and $\sup_{y\in\Omega^h_{\mathcal{P}}} d(y,\mathcal{S}) \leq h$, trivially.

Now, we can write $\mathbb{E}(d_{\mathcal{H}}(\mathcal{S},\Omega^h_{\mathcal{P}_n})) = \mathbb{E}(\mathbb{E}(d_{\mathcal{H}}(\mathcal{S},\Omega^h_{\mathcal{P}})\|\mathbb{1}_{\left[\mathcal{S}\subseteq\Omega^h_{\mathcal{P}}\right]}))$, but the inner expected value can be bounded by h when $\mathbb{1}_{\left[\mathcal{S}\subseteq\Omega^h_{\mathcal{P}_n}\right]} = 1$, and by $\max\{\text{diam}\,(\mathcal{S})+h, h\}$ when $\mathbb{1}_{\left[\mathcal{S}\subseteq\Omega^h_{\mathcal{P}}\right]} = 0$. Using Chebyshev's inequality, we find

$$\begin{aligned}\mathbb{P}\left(d_{\mathcal{H}}(\mathcal{S},\Omega^h_{\mathcal{P}_n}) > \delta\right) &\leq \frac{h}{\delta}\mathbb{P}\left(\{\mathcal{S}\subseteq\Omega^h_{\mathcal{P}}\}\right) + \frac{\max\{\text{diam}\,(\mathcal{S})+h, h\}}{\delta}\left(1-\mathbb{P}\left(\{\mathcal{S}\subseteq\Omega^h_{\mathcal{P}}\}\right)\right)\\ &\leq \frac{h}{\delta} + \frac{\text{diam}\,(S)+h}{\delta}\left(1-\mathbb{P}\left(\{\mathcal{S}\subseteq\Omega^h_{\mathcal{P}}\}\right)\right),\end{aligned}$$

a quantity that goes to zero for any fixed $\delta > 0$ as $h \downarrow 0$ and $n \uparrow \infty$, provided that (16) holds. ∎

Proof of Theorem 8. Since the proof is almost identical to that of Theorem 7, many steps will be skipped. Note that since $\mathcal{S}$ is compact, there exists an upper bound K for all its sectional curvatures. This will allow us to use the volume comparison theorems as before.

We can start from the adequate version of (7). We must bound both $\mathbb{P}(\{\mathcal{S}\subseteq\Omega^h_{\mathcal{P}_n}\}^c)$ and $\mathbb{P}(\mathcal{L}_{\mathcal{S}}(\mathcal{P}_n;\Delta,h) > \varepsilon \,|\, \{\mathcal{S}\subseteq\Omega^h_{\mathcal{P}_n}\})$. The second term can be bounded in an identical way as its $\Delta = 0$ counterpart was, obtaining

$$\mathbb{P}\left(\mathcal{L}_{\mathcal{S}}(\mathcal{P}_n;\Delta,h) > \varepsilon \,|\, \{\mathcal{S}\subseteq\Omega^h_{\mathcal{P}_n}\}\right) \leq \frac{C_{\mathcal{S}}\sqrt{h+\Delta} + 2\Delta(2+\sqrt{2}C_{\mathcal{S}}\sqrt{\Delta})}{\varepsilon}, \tag{20}$$

which vanishes as $n \uparrow \infty$.

Now we upper bound $\mathbb{P}(\{\mathcal{S}\subseteq\Omega^h_{\mathcal{P}_n}\}^c)$. Everything carries over in the same fashion as in the proof of Lemma 1, except that now we must take into consideration that the p_i's are not necessarily on $\mathcal{S}$ but inside $\Omega^{\Delta}_{\mathcal{S}}$. Following the described steps, we obtain

$$\mathbb{P}\left(\left\{x \notin \Omega^h_{\mathcal{P}_n}\cap\mathcal{S}\right\}\right) \leq \left(1 - \frac{\mu\left(B(x,h)\cap\Omega^{\Delta}_{\mathcal{S}}\right)}{\mu\left(\Omega^{\Delta}_{\mathcal{S}}\right)}\right)^n. \tag{21}$$

Notice that, since we are working with $h \geq \Delta$, we have $B(x,\Delta) \subset B(x,h)\cap\Omega^{\Delta}_{\mathcal{S}}$ (see Figure 8), and we can rewrite the bound in (21) as

$$\mathbb{P}\left(\left\{x \notin \Omega^h_{\mathcal{P}_n}\cap\mathcal{S}\right\}\right) \leq \left(1 - \frac{\mu\left(B(x,\Delta)\right)}{\mu\left(\Omega^{\Delta}_{\mathcal{S}}\right)}\right)^n \tag{22}$$

$$= \left(1 - \frac{\mu\left(B(\cdot,\Delta)\right)}{\mu\left(\Omega^{\Delta}_{\mathcal{S}}\right)}\right)^n. \tag{23}$$

We can bound this quantity using formulas akin to Weyl's tube theorem. More precisely, as explained in Appendix D, we can write

$$\mu\left(\Omega^{\Delta}_{S}\right) = \mu\left(\mathcal{S}\right) v(d-k,\Delta) + \varphi_{\mathcal{S}}(\Delta),$$

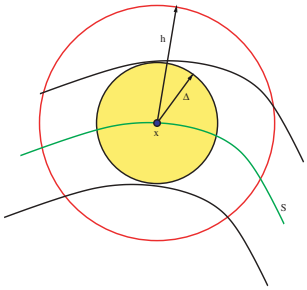


FIG. 8. $B(x,\Delta) \subset B(x,h) \cap \Omega_S^{\Delta}$.

where $\frac{\varphi_{\mathcal{S}}(\Delta)}{\Delta^{d-k+1}} \to 0$ as $\Delta \to 0$ and $v(D,R)$ is the volume of the ball of radius R in D-dimensional Euclidean space.

Now, for $x \in \mathcal{S}$ we must find a bound for $\mathbb{P}(B_{\mathcal{S}}(x,\delta) \not\subseteq \Omega^h_{\mathcal{P}_n})$, but as in the proof of Lemma 2, $\mathbb{P}(B_{\mathcal{S}}(x,\delta) \not\subseteq \Omega^h_{\mathcal{P}_n}) \leq \mathbb{P}(x \notin \Omega^{h-\delta}_{\mathcal{P}_n} \cap \mathcal{S})$, which can be bounded by (22). Also the bound (14) for the covering number still works in this case, and thus we can write $\mathbb{P}(\mathcal{S} \not\subseteq \Omega^h_{\mathcal{P}_n} \cap \mathcal{S}) \leq \frac{(1-y_\Delta)^n}{x_h}$, where $y_\Delta \stackrel{\triangle}{=} \frac{\mu(B(\cdot,\Delta))}{\mu(\Omega_S^{\Delta})}$. Also since $h \geq \Delta$, $x_h \geq x_\Delta$, then

$$\mathbb{P}\left(\mathcal{S} \not\subseteq \Omega^h_{\mathcal{P}_n}\right) \leq \frac{(1-y_\Delta)^n}{x_\Delta}.$$

But with Δ small enough, $y_\Delta \simeq \alpha\Delta^k$ and $x_\Delta \simeq \beta\Delta^k$, and then Lemma 4 and the hypotheses guarantee that $\mathbb{P}(\{\mathcal{S} \subset \Omega^h_{\mathcal{P}_n}\}^c) \to 0$ as $n \uparrow \infty$. □

Appendix D. Basic differential geometry facts. In this section we collect some facts that were used throughout the article, following [32].

D.1. Measure of a *d*-dimensional ball. Recall the definition of the Γ function:

$$\Gamma(\alpha) = \int_0^{+\infty} e^{-t} t^{\alpha-1} dt.$$

THEOREM 11. *The volume of d-dimensional ball of radius r is given by*

$$v(d,r) \stackrel{\triangle}{=} \mu\left(B(\cdot,r)\right) = \omega_d r^d,$$

where $\omega_d = \frac{2\pi^{d/2}}{d\Gamma(d/2)}$.

D.2. Bishop–Günther inequalities for the measure of a geodesic ball.

THEOREM 12. *Let $\mathcal{S}$ be a complete k-dimensional Riemannian manifold, assume r to be smaller than the distance between $m \in \mathcal{S}$ and $Cut(m,\mathcal{S})$ (cut locus of the points m in $\mathcal{S}$). Let $K^{\mathcal{S}}$ be the sectional curvatures of $\mathcal{S}$ and γ a constant. Then if* $\widehat{V}_\gamma(r) \stackrel{\triangle}{=} \frac{2\pi^{k/2}}{\Gamma(k/2)} \int_0^r \left(\frac{\sin(t\sqrt{\gamma})}{\sqrt{\gamma}}\right)^{k-1} dt$, *then*

$$K^{\mathcal{S}} \geq \gamma \ \textit{implies} \ \mu\left(B_{\mathcal{S}}(m,r)\right) \leq \widehat{V}_\gamma(r), \tag{24}$$

$$K^{\mathcal{S}} \leq \gamma \ \textit{implies} \ \mu\left(B_{\mathcal{S}}(m,r)\right) \geq \widehat{V}_\gamma(r). \tag{25}$$

PROPOSITION 3. *We have the following Taylor expansion for $\widehat{V}_\gamma(r)$, the volume of a geodesic ball in a space of constant sectional curvature γ:*

$$\widehat{V}_\gamma(r) = \omega_k r^k \left(1 - r^2 \frac{\gamma}{6} \frac{k(k-1)}{k+2}\right) + \phi(r),$$

where $\frac{\phi(r)}{r^{k+2}} \longrightarrow 0$ as $r \downarrow 0$.

D.3. Weyl's tube theorem.

THEOREM 13. *Let $\mathcal{S}$ be a k-dimensional manifold topologically embedded in $\mathbb{R}^d$. Assume that $\mathcal{S}$ is compact closure, and that every point in the tube $T(\mathcal{S}, r) = \{x \in \mathbb{R}^d$ such that, $d(\mathcal{S}, x) \leq r\}$ has a unique shortest geodesic connecting it with $\mathcal{S}$; then the volume $\mu(T(\mathcal{S}, r))$ of the tube is given by*

$$\mu(T(\mathcal{S}, r)) = r \frac{\sqrt{\pi}}{\Gamma(3/2)} \sum_{i=0}^{\left[\frac{d-1}{2}\right]} \frac{k_{2i}(\mathcal{S}) r^{2i}}{I(i)}, \tag{26}$$

where $I(i) = 1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2i+1)$ and the numbers k_{2i} depend on the curvature structure of $\mathcal{S}$. For our purposes we need know only that $k_0 = \mu(\mathcal{S})$.

COROLLARY 6. *The volume of the tube $T(\mathcal{S}, r)$ can be expanded as*

$$\mu(T(\mathcal{S}, r)) = \mu(\mathcal{S})\ v(d-k; r)\ +\ \phi_{\mathcal{S}}(r),$$

where $\frac{\phi_{\mathcal{S}}(r)}{r^{d-k}} \xrightarrow{r \downarrow 0} 0$

Appendix E. Details on object recognition. The ideal objective is to actually compute the $\mathfrak{I}$-distance between D_1 and D_2 as described in section 7.2; however, this is a very hard problem since there are $m!$ $m \times m$ permutation matrices. The choice of m is subject to compromise: on one hand, we want it to be big enough so as to capture the metric structure of $\mathcal{S}_i$ with the information given by $(\Omega_n^{(i)}, d_{\Omega_{\mathcal{P}_n^{(i)}}^{h_i}})$; on the other hand, we want to be able to actually make the computations involved without too much processing cost. Therefore we should attempt to circumvent this $m!$ search space by exploiting some other information we might have.

One possibility for bypassing this difficulty is to try to upper bound the $\mathfrak{I}$-distance by some difference between eigenvalues of the matrices. However, it turns out that one can easily find two distance matrices which have positive $\mathfrak{I}$-distance (they are not *cogredient*) but have the same spectra. Then an upper bound should take into account also another term that measures our inability to really differentiate distance functions by looking only at their eigenvalues. Of course this information must then be contained in the eigenvectors.[30]

A way of dealing with this particular issue is working with the spectral factorization of each of the matrices. Let $D_1 = QDQ^T$ and $D_2 = \widehat{Q}\widehat{D}\widehat{Q}^T$, where Q and $\widehat{Q}$ are unitary matrices and D and $\widehat{D}$ are diagonal matrices whose entries are the eigenvalues of D_1 and D_2, respectively. Note that we are not saying anything about the order in which those eigenvalues are presented; for convenience, let

[30] Another idea, for example, is the following: We know that the searched-for isometry (if it exists) must be a Lipschitz continuous map, and therefore it makes no sense to consider the huge set of transformations spanned by $\mathcal{PM}_m$. We leave the exploitation of this idea for future work.

$D_{11} = |D_{11}| > |D_{22}| > \cdots > |D_{mm}|$ and $\widehat{D}_{11} = |\widehat{D}_{11}| > |\widehat{D}_{22}| > \cdots > |\widehat{D}_{mm}|$.[31] Then with little effort we can write

$$\min_{P \in \mathcal{PM}_m} \|D_1 - P D_2 P^T\| \leq \|(Q - P\widehat{Q})D\| + \|(Q - P\widehat{Q})\widehat{D}\| + \|D - \widehat{D}\|. \tag{27}$$

Note that if W is any matrix and T is diagonal, then $\|WT\|^2 = \sum_k \|W_{(:,k)} t_{kk}\|^2 = \sum_k \|W_{(:,k)}\|^2 t_{kk}^2$ where $W_{(:,k)}$ is the kth column vector of W. Using this observation, we note that the first two terms in (27) can be bounded as follows (let $Q = (q_1|\ldots|q_m)$, and $\widehat{Q} = (\widehat{q}_1|\ldots|\widehat{q}_m)$):

$$\|(Q - P\widehat{Q})D\| + \|(Q - P\widehat{Q})\widehat{D}\| = \sqrt{\sum_k D_{kk}^2 \|q_k - P\widehat{q}_k\|^2} + \sqrt{\sum_k \widehat{D}_{kk}^2 \|q_k - P\widehat{q}_k\|^2}.$$

Now, using the trivial inequality $\frac{\sqrt{a}+\sqrt{b}}{2} \leq \sqrt{\frac{a+b}{2}}$ for all $a, b \geq 0$, we finally arrive at the expression

$$\min_{P \in \mathcal{PM}_m} \|D_1 - P D_2 P^T\| \leq \sqrt{\sum_{k=1}^m (D_{kk} - \widehat{D}_{kk})^2} + \sqrt{2}\sqrt{\sum_{k=1}^m (D_{kk}^2 + \widehat{D}_{kk}^2)\|q_k - P\widehat{q}_k\|^2}. \tag{28}$$

This inequality holds for any $P \in \mathcal{PM}_m$. It is important to note that in case D_1 and D_2 are cogredient, all their eigenvectors will also be related through that same permutation; therefore this inequality is sharp.[32]

Note that in the second term of (28), the values of $\|q_i - P\widehat{q}_i\|$ are weighted by $(D_{ii}^2 + \widehat{D}_{ii}^2)$, so one can think that since $\|q_i - P\widehat{q}_i\| \leq 2$, the most important terms of the sum will be those for which $(D_{ii}^2 + \widehat{D}_{ii}^2)$ is large. This is not a rigorous consideration, but gives some guidelines on how to compute an approximate bound when the sizes of the distance matrices are prohibitively large.

In some situations, the choice of the subsampled set size m that guarantees a good metric approximation in the sense discussed above might be too large, making the computation of the full bound (28) onerous. But still a measure of similarity must be provided which does not require the computation of all of the eigenvalues and eigenvectors of each distance matrix. Therefore, in order to estimate $d_{\mathcal{J}}(D_1, D_2)$, we use the following idea: Instead of computing all the eigenvalues and eigenvectors of the matrices D_1 and D_2, compute the $N \ll m$ more important ones, where important means, in the light of the expression for the bound, those with the largest moduli, at least for the part of the bound involving eigenvectors. Then, for a (computationally) reasonable N we define the approximate error bound (still letting P be any convenient choice of a permutation matrix)

$$e(N) \triangleq \sqrt{\sum_{k=1}^N (D_{kk} - \widehat{D}_{kk})^2} + \sqrt{2}\sqrt{\sum_{k=1}^N (D_{kk}^2 + \widehat{D}_{kk}^2)\|q_k - P\widehat{q}_k\|^2}. \tag{29}$$

[31] We have used Frobenius theorem [51], which asserts that nonnegative matrices have a positive largest absolute value eigenvalue. Note that we have also assumed that there are no repeated eigenvalues.

[32] Note that from (27) one can obtain $d_{\mathcal{J}}(D_1, D_2) \leq \|D - \widehat{D}\| + (\|D\| + \|\widehat{D}\|)\|Q\widehat{Q}^T - P\|$; then one further idea to be explored is how to best approximate a given unitary matrix by a permutation matrix. This would not only allow us to obtain an explicit bound for the $\mathcal{J}$-distance, but would also provide us with a low metric distortion way of mapping S_1 ($\mathcal{P}_n^{(1)}$) into S_2 ($\mathcal{P}_n^{(2)}$), with applications like texture mapping, brain warping, etc.

Now, we fix the permutation P as follows: Let S be the permutation matrix such that Sq_1 is a column vector whose components are sorted from largest to smallest. Do the same with $\widehat{q}_1$ to obtain $\widehat{S}$; then compare Sq_1 with $\widehat{S}\widehat{q}_1$, which amounts to comparing q_1 with $S^T\widehat{S}$; hence we let $P = S^T\widehat{S}$. We could again use a more sophisticated way of choosing P, but this one suffices for demonstration purposes and, of course, achieves equality in (28) when both matrices are cogredient.

Another possibility is to directly compare the distance matrices according to the expression $\|D_1 - PD_2P^T\|$, using a certain sensible choice for P. We first put both matrices in a "canonical" order. Let (i_1, j_1) be one position on the matrix D_1 with the maximum value. We then order the rest of the points in the set according to their distances to either $q_{i_1}^{(1)}$ or $q_{i_2}^{(1)}$ from smallest to largest.[33] This induces an ordering for the matrix D_1, letting P_1 be the underlying permutation matrix. We do the same with D_2 and obtain P_2. Finally we let

$$e_G(D_1, D_2) \stackrel{\triangle}{=} \|D_1 - P_1^T P_2 D_2 P_2^T P_1\|,$$

and note that obviously $d_{\mathcal{J}}(D_1, D_2) \leq e_G(D_1, D_2)$ and that the inequality is sharp.

E.1. Choice of the point cloud subset $\mathcal{Q}^{(i)}$. In general, the number of points in the cloud is too big. This means that the actual computation of the distance matrices, if done using all the points in the cloud, and subsequent eigenvalue and eigenvector computations (if needed) become onerous. Therefore we need a procedure which allows us to select a small cardinality subset $\mathcal{Q}_m$ of $\mathcal{P}_n$ for which we will actually compute the approximate distance matrix, but still using $\mathcal{P}_n$ to define the offset $\Omega^h_{\mathcal{P}_n}$ inside which the computations are performed. This subset $\mathcal{C}_r \subset \mathcal{P}_n$ must be "representative" of the geometry of the underlying manifold. One way of selecting those points is by not allowing them to cluster inside any region of the manifold. This can be accomplished in practice by using the "packing idea" in [24]: Given $m < n$, choose the first point $c_1 \in \mathcal{C}_m$ randomly, then proceed by always choosing a point as far as possible from the set of points that have already been chosen. End the process when m points have been chosen. This is the procedure used in the experiments.

Acknowledgments. We acknowledge useful conversations on the topic of this paper with L. Aspirot, P. Bermolén, R. Coifman, D. Donoho, N. Dyn, J. Giesen, O. Gil, R. Gulliver, R. Kimmel, A. Pardo, and O. Zeitouni. We thank M. Levoy and the Digital Michelangelo Project for data provided for this project. We thank the anonymous reviewers for their comments, which helped improve the presentation of the paper.

REFERENCES

[1] R. Alexander and S. Alexander, *Geodesics in Riemannian manifolds with boundary*, Indiana Univ. Math. J., 30 (1981), pp. 481–488.

[2] L. Ambrosio and H. M. Soner, *Level set approach to mean curvature flow in arbitrary codimension*, J. Differential Geom., 43 (1996), pp. 693–737.

[3] N. Amenta, S. Choi, and R. Kolluri, *The power crust, unions of balls, and the medial axis transform*, Comput. Geom., 19 (2001), pp. 127–153.

[4] N. Amenta, S. Choi, and R. Kolluri, *The power crust*, in Proceedings of the 6th Annual ACM Symposium on Solid Modeling, Ann Arbor, MI, 2001, ACM, New York, pp. 249–260.

[33] In our current implementation we don't worry about repeated distance values, since this can be easily handled. We will present more details and further refinements elsewhere.

[5] N. Amenta and R. Kolluri, *Accurate and efficient unions of balls*, in Proceedings of the ACM Symposium on Computational Geometry, Hong Kong, ACM, New York, 2000, pp. 119–128.
[6] T. Apostol, *Mathematical Analysis*, Addison–Wesley Ser. Math., Addison–Wesley, Reading, MA, 1974.
[7] A. Bartesaghi and F. Mémoli, *Flujos software*, http://iie.fing.edu.uy/investigacion/grupos/gti/flujos/flujos.html.
[8] A. Bartesaghi and G. Sapiro, *A system for the generation of curves on 3D brain images*, Human Brain Mapping, 14 (2001), pp. 1–15.
[9] M. Bernstein, V. de Silva, J. Langford, and J. Tenenbaum, *Graph approximations to geodesics on embedded manifolds*, http://isomap.stanford.edu/BdSLT.ps.
[10] *Blitz++ website*, http://www.oonumerics.org/blitz.
[11] J-D. Boissonnat and F. Cazals, *Coarse-to-fine surface simplification with geometric guarantees*, in Proceedings of EUROGRAPHICS 2001, A. Chalmers and T-M. Rhyne, eds., Manchester, UK, 2001.
[12] M. Botsch, A. Wiratanaya, and L. Kobbelt, *Efficient high quality rendering of point sampled geometry*, in Proceedings of the 13th EUROGRAPHICS Workshop on Rendering, Pisa, Italy, 2002, pp. 53–64.
[13] E. Calabi, P. J. Olver, and A. Tannenbaum, *Affine geometry, curve flows, and invariant numerical approximations*, Adv. Math., 124 (1996), pp. 154–196.
[14] I. Chavel, *Riemannian Geometry : A Modern Introduction*, Cambridge University Press, Cambridge, UK, 1993.
[15] R. Coifman, *personal communication*, Yale University, New Haven, CT, 2002.
[16] R. Coifman, *personal communication*, Yale University, New Haven, CT, 2003 (talk presented at IPAM-UCLA).
[17] R. Coifman, *personal communication*, Yale University, New Haven, CT, 2003 (talk presented at University of Minnesota).
[18] T. K. Dey, J. Giesen, and J. Hudson, *Decimating samples for mesh simplification*, in Proceedings of the 13th Annual Canadian Conference on Computational Geometry, Waterloo, ON, 2001, pp. 85–88.
[19] D. L. Donoho and C. Grimes, *When Does ISOMAP Recover the Natural Parametrization of Families of Articulated Images?*, Technical Report 2002-27, Department of Statistics, Stanford University, Stanford, CA, 2002.
[20] N. Dyn, M. S. Floater, and A. Iske, *Adaptive thinning for bivariate scattered data*, J. Comput. Appl. Math., 145 (2002), pp. 505–517.
[21] M. doCarmo, *Riemannian Geometry*, Birkhäuser Boston, Cambridge, MA, 1992.
[22] A. Dvoretsky, *On covering a circle by randomly placed arcs*, Proc. Natl. Acad. Sci. USA, 42, (1956), pp. 199–203.
[23] H. Edelsbrunner, *The union of balls and its dual shape*, Discrete Comput. Geom., 13 (1995), pp. 415–440.
[24] A. Elad (Elbaz) and R. Kimmel, *Bending invariant representations for surfaces*, in Proceedings of the Computer Vision and Pattern Recognition (CVPR'01), Kauai, HI, 2001, pp. I-168–I-174.
[25] R. Ellis, X. Jia, and C. Yan, *On Random Points in the Unit Disk*, submitted; available online at http://www.math.tamu.edu/∼rellis/papers/5random.pdf.
[26] H. Federer, *Curvature measures*, Trans. Amer. Math. Soc., 93 (1959), pp. 418–491.
[27] L. Flatto and D. J. Newman, *Random coverings*, Acta Math., 138 (1977), pp. 241–64.
[28] L. Flatto, *A limit theorem for random coverings of a circle*, Israel J. Math., 15 (1973), pp. 167–184.
[29] M. S. Floater and A. Iske, *Thinning algorithms for scattered data interpolation*, BIT, 38 (1998), pp. 705–720.
[30] S. F. Frisken, R. N. Perry, A. P. Rockwood, and T. R. Jones, *Adaptively sampled distance fields: A general representation of shape for computer graphics*, in Proceedings of SIGGRAPH 2000, New Orleans, LA, ACM, New York, 2000, pp. 249–254.
[31] J. Giesen and U. Wagner, *Shape dimension and intrinsic metric from samples of manifolds with high co-dimension*, in Proceedings of the 19th ACM Symposium on Computational Geometry, San Diego, CA, 2003, pp. 329–337.
[32] A. Gray, *Tubes*, Addison–Wesley, Reading, MA, 1990.
[33] M. Alexa, M. Gross, M. Pauly, H. Pfister, M. Stamminger, and M. Zwicker, *Point Based Computer Graphics*, EUROGRAPHICS Lecture Notes, 2002, available online at http://graphics.stanford.edu/∼niloy/research/papers/ETH/PointBasedComputerGraphics_TutorialNotes.pdf.
[34] P. Hall, *Introduction to the Theory of Coverage Processes*, Wiley Series in Probability and

Mathematical Statistics, John Wiley & Sons, New York, 1988.
[35] J. Helmsen, E. G. Puckett, P. Collela, and M. Dorr, *Two new methods for simulating photolithography development in* 3*D*, in Proceedings of the SPIE Microlithography, IX (1996), pp. 253.
[36] P. W. Jones, *Rectifable sets and the traveling salesman problem*, Invent. Math., 102 (1990), pp. 1–15.
[37] S. Janson, *Random coverings in several dimensions*, Acta Math., 156 (1986), pp. 83–118.
[38] J. Hoffmann-Jørgensen, *Coverings of metric spaces with randomly placed balls*, Math. Scand., 32 (1973), pp. 169–186.
[39] M. G. Kendall and P. A. P. Morgan, *Geometrical Probability*, Griffin, London, 1963.
[40] R. Kimmel and J. A. Sethian, *Computing geodesic paths on manifolds*, Proc. Natl. Acad. Sci. USA, 95 (1998), pp. 8431–8435.
[41] A. N. Kolmogorov and S. V. Fomin, *Elements of the Theory of Functions and Functional Analysis*, Dover, New York, 1999.
[42] R. Kunze, F. E. Wolter, and T. Rausch, *Geodesic Voronoi diagrams on parametric surfaces*, in Proceedings of the Computer Graphics International (CGI) '97, IEEE, Computer Society Press, Kinepolis, Belgium, 1997, Piscataway, NJ, 1997, pp. 230–237.
[43] G. Leibon and D. Letscher, *Delaunay triangulations and Voronoi diagrams for Riemannian manifolds*, ACM Symposium on Computational Geometry 2000, Hong Kong, ACM, New York, 2000, pp. 341–349.
[44] G. Lerman, *How to Partition a Low-Dimensional Data Set into Disjoint Clusters of Different Geometric Structure*, preprint, 2000.
[45] L. Linsen, *Point Cloud Representation*, CS Technical Report, University of Karlsruhe, Karlsruhe, Germany, 2001.
[46] L. Linsen and H. Prautzsch, *Local versus global triangulations*, in Proceedings of EUROGRAPHICS, 2001, A. Chalmers and T-M. Rhyne, eds., Manchester, UK.
[47] C. Mantegazza and A.C. Menucci, *Hamilton-Jacobi equations and distance functions on Riemannian manifolds*, Appl. Math. Optim., 47 (2003), pp. 1–25.
[48] A. Marino and D. Scolozzi, *Geodetiche con ostacolo*, Boll. Un. Mat. Ital., 6 (1983), pp. 1–31.
[49] F. Mémoli and G. Sapiro, *Fast computation of weighted distance functions and geodesics on implicit hyper-surfaces*, J. Comput. Phys., 173 (2001), pp. 730–764.
[50] F. Mémoli and G. Sapiro, *Comparing point clouds*, in Proceedings of the 2nd Annual EUROGRAPHICS Symposium on Geometry Processing, Nice, France, 2004.
[51] H. Minc, *Nonnegative Matrices*, Wiley Ser. Discrete Math. Optim., Wiley Interscience, New York, 1988.
[52] J. S. B. Mitchell, *An algorithmic approach to some problems in terrain navigation*, Artificial Intelligence, 37 (1988), pp. 171–201.
[53] J. S. B. Mitchell, D. Payton, and D. Keirsey, *Planning and reasoning for autonomous vehicle control*, Internat. J. Intelligent Systems, 2 (1987), pp. 129–198.
[54] N. J. Mitra and A. Nguyen, *Estimating surface normals in noisy point cloud data*, in Proceedings of the ACM Symposium on Computational Geometry, San Diego, 2003, ACM, New York, 2003, pp. 322–328.
[55] C. Moenning, F. Mémoli, G. Sapiro, N. Dyn, and N. A. Dodgson, *Meshless Geometric Subdivision*, Technical report IMA TR 1977, 2004, available online from http://www.ima.umn.edu/preprints/apr2004/apr2004.html.
[56] M. Pauly and M. Gross, *Spectral processing of point-sampled geometry*, in Proceedings of SIGGRAPH 2001, Los Angeles, 2001, ACM, New York, pp. 379–386.
[57] M. Pauly, N. J. Mitra, and L. Guibas, *Uncertainty and variability in point cloud surface data*, in Proceedings of the Symposium on Point-Based Graphics, Zurich, 2004.
[58] M. Pauly, M. Gross, and L. Kobbelt, *Efficient simplification of point-sampled surfaces*, in Proceedings of the IEEE Visualization meeting, 2002, pp. 163–170.
[59] S. Rusinkiewicz and M. Levoy, *QSplat: A multiresolution point rendering system for large meshes*, in Proceedings of SIGGRAPH 2000, New Orleans, LA, ACM, New York, 2000, pp. 343–352.
[60] T. Sakai, *Riemannian Geometry*, AMS Translations of Mathematical Monographs 149, AMS, New York, 1996.
[61] E. Schwartz, A. Shaw, and E. Wolfson, *A numerical solution to the generalized mapmaker's problem: Flattening nonconvex polyhedral surfaces*, IEEE Trans. Pattern Anal. Machine Intelligence, 11 (1989), pp. 1005–1008.
[62] J. Sethian, *Fast marching level set methods for three-dimensional photolithography development*, in Proceedings of the SPIE International Symposium on Microlithography, Santa Clara, CA, SPIE, Bellingham, WA, 1996.
[63] J. A. Sethian, *A fast marching level-set method for monotonically advancing fronts*, Proc.

Natl. Acad. Sci. USA, 93 (1996), pp. 1591–1595.
[64] C. E. Shannon, *Coding theorems for a discrete source with a fidelity criterion*, in Information and Decision Processes, McGraw–Hill, New York, 1960.
[65] L. A. Shepp, *Covering the circle with random arcs*, Israel J. Math., 11 (1972), pp. 328–345.
[66] H. Solomon, *Geometric Probability*, in CBMS-NSF Reg. Conf. Ser. Appl. Math., 28, SIAM, Philadelphia, 1978.
[67] J. B. Tenenbaum, V. de Silva, and J. C. Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science, 290 (2002), pp. 2319–2323.
[68] Y.-H. R. Tsai, L.-T. Cheng, S. Osher, and H.-K. Zhao, *Fast sweeping algorithms for a class of Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 41 (2003), pp. 673–694.
[69] J. N. Tsitsiklis, *Efficient algorithms for globally optimal trajectories*, IEEE Trans. Automat. Control, 40 (1995), pp. 1528–1538.
[70] *The Visualization Toolkit*, available online at http://www.vtk.org.
[71] F. E. Wolter, *Cut Loci in Bordered and Unbordered Riemannian Manifolds*, Doctoral Dissertation, Technische Universität Berlin, Berlin, 1985.
[72] M. Zwicker, M. Pauly, O. Knoll, and M. Gross, *PointShop 3D: An interactive system for point-based surface editing*, Proceedings of SIGGRAPH 2002, San Antonio, TX, 2002, pp. 322–329.

SIAM J. APPL. MATH.
Vol. 65, No. 4, pp. 1261–1284

A MIXTURE THEORY FOR THE GENESIS OF RESIDUAL STRESSES IN GROWING TISSUES I: A GENERAL FORMULATION*

ROBYN P. ARAUJO† AND D. L. SEAN MCELWAIN†

Abstract. In this paper a theoretical framework for the study of residual stresses in growing tissues is presented using the theory of mixtures. Such a formulation must necessarily be a solid-multiphase model, comprising at least one phase with solid characteristics, owing to the fundamental role played by the incompatibility of strains in generating residual stresses. Since biological growth involves mass exchange between cellular and extracellular phases, field equations are presented for individual phases and for the mixture as a whole which incorporate this phenomenon. Appropriate constitutive equations are then deduced from first principles, appealing to the second law of thermodynamics.

The analysis shows that the distinguishing feature of multiphase models involving mass exchange is the necessity to propose an additional constitutive postulate between the variables in the mass-balance equation in order to close the model. In particular, the defining characteristic of a solid-multiphase model which describes biological growth is a constitutive postulate which relates the process of interphase mass exchange (cell proliferation/cell death) with the expansion or contraction of the solid phase. Thus, the framework presented here represents a new class of mathematical models which extends the concepts of poroelasticity to accommodate continuous volumetric growth. A set of modelling equations is then proposed for the simplest case of a solid-multiphase model, being a biphasic mixture of a linear-elastic solid and an inviscid fluid.

Key words. tissue growth, mixture theory, residual stresses, continuum mechanics, constitutive equations, porous media

AMS subject classifications. 70S99, 74A, 74F, 74L15, 76S05

DOI. 10.1137/040607113

1. Introduction.

The evolution and spatial distribution of tissue stresses is of fundamental importance in a number of physiological phenomena. The experimentally observed phenomenon of vascular collapse in tumors, for example, which has been attributed to the elevated tissue stresses resulting from confined proliferation of tumor cells [5, 9], represents a significant barrier to the delivery of blood-borne therapeutic drugs. Such stresses are *residual* in nature, arising in the tissue when it is free of external loads, and result from the incompatibility of growth strains [21, 38, 40].

Fung [20] further notes the existence of residual stresses in living organs and highlights the importance of such stresses to physiological functions, asserting that "in a living organism, the function of its organs depends on the levels of their internal stress and strain."

Hence continuum models of growing tissues would provide a theoretical framework for a wide range of studies in biology, ranging from tumor biology and anticancer therapies [23, 38] to studies in embryology [7, 34], developmental biology, and plant physiology [18], in addition to providing tools for prediction and analysis for a wide range of projects in the rapidly growing field of tissue engineering [31].

Nevertheless, the underlying phenomenological determinants of residual stresses, as well as their purpose and implications in both normal tissue development and

*Received by the editors January 5, 2004; accepted for publication (in revised form) October 20, 2004; published electronically April 14, 2005.
http://www.siam.org/journals/siap/65-4/60711.html
†School of Mathematical Sciences, Queensland University of Technology, PO Box 2434 Brisbane, QLD 4001 Australia (s.mcelwain@qut.edu.au).

various pathological conditions, are poorly understood since there is a paucity of mathematical models to elucidate these phenomena.

Gatenby [22] explains that "recent research in tumour biology, particularly that using new techniques from molecular biology, has produced information at an explosive pace. Yet a conceptual framework within which all these new (and old) data can be fitted is lacking." Gatenby and Maini [23] add that "clinical oncologists and tumour biologists possess virtually no comprehensive theoretical model to serve as a framework for understanding, organizing and applying these data," noting the necessity to "(develop) mechanistic models that provide real insights into critical parameters that control system dynamics." Murray [32] concurs, arguing that "the goal is to develop models which capture the essence of various interactions allowing their outcome to be more fully understood."

Indeed, while experimental approaches may attest to the *existence* of residual stresses and provide information about their distribution in tissues, the underlying mechanisms governing their genesis cannot be fully elucidated in the absence of mathematical modeling owing to the fundamental role played by the incompatibility of growth strains in their formation [40]. Mathematical analysis provides the key to identifying incompatible growth and represents a tool for investigating the roles of a variety of phenomenological aspects of growing tissues—distribution of nutrients, growth-related density changes, stress modulated cell-proliferation and apoptosis, geometric effects—in promoting incompatibilities and the associated residual stresses.

An important consideration in the mathematical modelling of tissue growth is the choice between single-phase mechanics and mixture theory [2]. The former, which appeals to an analogy with thermal expansion, incorporates a *source term* in the balance of mass, with the phase or phases responsible for the mass source remaining implicit to the model. While Skalak [39] claims that volumetric growth is analogous to thermal expansion—an analogy which forms the basis of the tissue growth models by Shannon and Rubinsky [38], Jones et al. [26], and Araujo and McElwain [4]—it does not consider all the processes which determine the stresses induced during biological tissue growth. Indeed, Araujo and McElwain [4] note that the single-constituent framework does not take into account the net fluid movement associated with the growth process and the Darcy-like drag terms in the equilibrium of forces—a consideration which may be significant when the elastic (residual) stresses are small.

Multiphase models, on the other hand, which are based on mixture theory, clarify the nature of any mass sources, and consider the role of interstitial fluid in the growth process. While several *fluid* multiphase models of growing tissues have been proposed recently [15, 16, 28], it is essential to recognize that these models provide no basis for examining the genesis of *residual* stresses in tissues, which requires a consideration of the tissue's *solid* characteristics.

Hence, a theoretical framework which enables the growth process and the associated development of tissue stresses to be modeled naturally, without recourse to an analogy with thermal expansion, is lacking.

This paper is the first in a series of papers which elaborate a mixture theory for the genesis of residual stresses in growing soft tissues, based on field equations which incorporate interphase mass exchange. This paper presents a general formulation for deducing thermodynamically appropriate constitutive equations relevant to the study of biological growth. In section 2, the field equations are presented, being adapted from the classical field theories developed by Truesdell and Toupin [42] and the theory of mixtures developed by Bowen [12] and manipulated into the forms most useful to further exploration of the problem at hand. In sections 3 and 4, these classical

theories are used as a guide to developing a particular form of the second axiom of thermodynamics from which the relevant constitutive equations may be deduced most readily, neglecting the influence of possible density changes associated with a change of phase. Constitutive assumptions for a general mixture of n phases are then outlined in section 5, following the pioneering work on constitutive modeling by Coleman and Noll [17], Ehlers [19], and Bowen [12].

In sections 6 through 8, a simple two-phase mixture of an elastic solid and an inviscid fluid is considered in detail. Since this solid-biphasic mixture must be able to exhibit continuous volumetric expansion to model the process of biological growth and an associated evolution of residual stresses, this paper stands alone in the mathematical literature pertaining to solid tumor growth, enucleating the very essence of the closure problem relevant to continuous growth of a tissue with solid characteristics. The solutions to these biphasic equations will be presented in the next paper in the series.

1.1. Differentiation conventions and index of symbols. The following conventions will be adopted throughout this paper.

If $\hat{\alpha_i}$ and $\boldsymbol{\alpha}_i$ are scalar and vector/tensor properties of the ith constituent, respectively, then $\boldsymbol{\nabla}\hat{\alpha_i}$ and $\boldsymbol{\nabla}.\boldsymbol{\alpha}_i$ denote the gradient and the divergence, respectively, with respect to *spatial* coordinates. (Note that in many references and texts in continuum mechanics these symbols are used to denote partial differentiation with respect to the *reference configuration*.) The symbols $\mathbf{Grad}\hat{\alpha_i}$ and $\mathbf{Div}\boldsymbol{\alpha}_i$ will denote the gradient and divergence, respectively, with respect to the *reference configuration*.

In addition, the symbol $\frac{D^i}{Dt}$ denotes the material derivative following the motion defined by $\mathbf{v}_i$, the velocity of the ith constituent. The symbol $\frac{D}{Dt}$, on the other hand, represents the material derivative following the motion defined by $\mathbf{v}_m$, the velocity of the mixture as a whole.

Table 1.1 gives a summary of the nomenclature adopted in this paper, along with the equation in which each symbol first appears. Each quantity is given a more complete description when it is first introduced in the text.

2. Constituent field equations. The balances of mass, linear momentum, angular momentum, and energy for the ith constituent of an n-phase mixture are summarized below. The equations incorporate a mass exchange term, so that the mass of the ith constituent may increase (or decrease) at the expense of other constituents. All constituents are equipresent at each spatial point.

2.1. Balance of mass. The balance of mass for the ith constituent, or *phase*, of an n-phase mixture is given by

$$\frac{D^i(\phi_i\rho_i)}{Dt} + (\phi_i\rho_i)\boldsymbol{\nabla}\cdot\mathbf{v}_i = \Gamma_i, \tag{2.1}$$

or, equivalently,

$$\frac{\partial(\phi_i\rho_i)}{\partial t} + \boldsymbol{\nabla}.(\phi_i\rho_i\mathbf{v}_i) = \Gamma_i, \tag{2.2}$$

where ϕ_i and ρ_i are the volume fraction and density, respectively, of the ith phase and Γ_i is the mass supplied to the ith phase per unit time per unit mixture volume. Truesdell and Toupin's [42] rule for differentiating a determinant gives the identity

$$\frac{D^i}{Dt}(\det\mathbf{F}_i) = (\det\mathbf{F}_i)\boldsymbol{\nabla}\cdot\mathbf{v}_i,$$

TABLE 1.1
Symbols.

Symbol	Description	Equation of first occurrence
ϕ	Volume fraction	(2.1)
ρ	True density	(2.1)
Γ	Mass supply	(2.1)
$\mathbf{F}$	Deformation Gradient	(2.3)
$\mathbf{v}$	Velocity	(2.1)
$\boldsymbol{\sigma}$	Partial Cauchy stress tensor	(2.7)
$\boldsymbol{\sigma}_I$	*Inner* mixture stress tensor	(2.16)
$\mathbf{g}$	Acceleration due to gravity	(2.7)
$\boldsymbol{\pi}$	Phase interaction force	(2.7)
$\mathbf{w}$	Diffusion velocity	(2.11)
$\mathbf{m}$	Angular momentum supply	(2.13)
$\mathbf{L}$	Velocity gradient	(2.18)
u	Internal energy	(2.18)
$\mathbf{q}$	Heat flux	(2.18)
r	Heat production rate	(2.18)
ε	Phase interaction energy supply	(2.18)
η	Entropy	(3.1)
θ	Absolute temperature	(3.1)
ψ	Helmholtz free energy	(3.4)
$\mathbf{K}$	Chemical potential tensor	(3.5)
ζ	Lagrangian multiplier	(4.4)
$\mathbf{X}$	Reference coordinates	(8.2)
$\mathbf{x}$	Spatial coordinates	(8.9)
μ, λ	Lamé constants	(8.22)

which enables (2.1) to be expressed by

$$\frac{D^i}{Dt}\left(\phi_i \rho_i \det\mathbf{F}_i\right) = \Gamma_i \det\mathbf{F}_i, \tag{2.3}$$

where $\mathbf{F}_i$ is the deformation gradient of the ith phase with respect to the reference configuration. The volume fractions, ϕ_i, are subject to the constraint

$$\sum_{i=1}^{n} \phi_i = 1, \tag{2.4}$$

which implies that the mixture is *saturated*.

The balance of mass for the mixture is expressed by

$$\frac{D\rho_m}{Dt} + \rho_m \boldsymbol{\nabla} \cdot \mathbf{v}_m = 0, \tag{2.5}$$

where ρ_m and $\mathbf{v}_m$ are the density and velocity, respectively, of the mixture as a whole. A comparison of (2.5) with the summation of (2.2) over all n phases allows the mixture density, ρ_m, to be defined by

$$\rho_m = \sum_{i=1}^{n} \phi_i \rho_i,$$

and the mixture velocity, $\mathbf{v}_m$, to be defined by

$$\mathbf{v}_m = \frac{1}{\rho_m} \sum_{i=1}^{n} (\phi_i \rho_i \mathbf{v}_i),$$

while yielding the following expression for the conservation of mass:

$$\sum_{i=1}^{n} \Gamma_i = 0. \tag{2.6}$$

2.2. Balance of linear momentum. The balance of linear momentum for the ith phase of an n-phase mixture is given by

$$\phi_i \rho_i \frac{D^i \mathbf{v}_i}{Dt} = \boldsymbol{\nabla} \cdot \boldsymbol{\sigma}_i + \phi_i \rho_i \mathbf{g} + \boldsymbol{\pi}_i, \tag{2.7}$$

or, equivalently,

$$\frac{\partial}{\partial t}(\phi_i \rho_i \mathbf{v}_i) + \boldsymbol{\nabla}.(\phi_i \rho_i \mathbf{v}_i \otimes \mathbf{v}_i) = \boldsymbol{\nabla} \cdot \boldsymbol{\sigma}_i + \phi_i \rho_i \mathbf{g} + \boldsymbol{\pi}_i + \Gamma_i \mathbf{v}_i, \tag{2.8}$$

where $\boldsymbol{\sigma}_i$ is the partial Cauchy stress tensor for the ith phase, $\mathbf{g}$ is the acceleration due to gravity, and $\boldsymbol{\pi}_i$ is the locally produced force per unit volume on the ith phase due to its interactions with the other phases. The symbol $\otimes$ denotes the dyadic vector product.

The balance of linear momentum for the mixture is expressed by

$$\rho_m \frac{D\mathbf{v}_m}{Dt} = \boldsymbol{\nabla} \cdot \boldsymbol{\sigma}_m + \rho_m \mathbf{g}, \tag{2.9}$$

where $\boldsymbol{\sigma}_m$ is the Cauchy stress tensor of the mixture as a whole. A comparison of (2.9) with the summation of (2.8) enables the mixture stress tensor to be defined by

$$\boldsymbol{\sigma}_m = \sum_i (\boldsymbol{\sigma}_i - \phi_i \rho_i \mathbf{w}_i \otimes \mathbf{w}_i), \tag{2.10}$$

where $\mathbf{w}_i$ denotes the diffusion velocity defined by

$$\mathbf{w}_i = \mathbf{v}_i - \mathbf{v}_m \tag{2.11}$$

and gives rise to the following expression for the conservation of linear momentum:

$$\sum_{i=1}^{n} (\boldsymbol{\pi}_i + \Gamma_i \mathbf{v}_i) = 0. \tag{2.12}$$

2.3. Balance of angular momentum. The balance of angular momentum for the ith phase of an n-phase mixture is given by

$$\frac{\partial}{\partial t}\Big(\phi_i \rho_i \mathbf{x} \times \mathbf{v}_i\Big) + \boldsymbol{\nabla} \cdot \Big(\phi_i \rho_i (\mathbf{x} \times \mathbf{v}_i) \otimes \mathbf{v}_i\Big) = \boldsymbol{\nabla} \cdot (\mathbf{x} \times \boldsymbol{\sigma}_i) + \mathbf{x} \times (\phi_i \rho_i \mathbf{g} + \boldsymbol{\pi}_i + \Gamma_i \mathbf{v}_i) + \mathbf{m}_i, \tag{2.13}$$

where $\mathbf{m}_i$ is a vector representing the supply of angular momentum to the ith phase. The symbol $\times$ denotes a cross product, in which the quantity

$$(\mathbf{x} \times \boldsymbol{\sigma}_i)\mathbf{e} = \mathbf{x} \times (\boldsymbol{\sigma}_i \mathbf{e})$$

for all vectors $\mathbf{e}$ (see Bowen [12]). Appealing to the balance of linear momentum produces

$$\mathbf{M}_i = \boldsymbol{\sigma}_i - \boldsymbol{\sigma}_i^T \tag{2.14}$$

from (2.13), where $\mathbf{M}_i$ is a skew-symmetric tensor arising from the angular momentum supply vector, $\mathbf{m}_i$. Since the sum of the momentum supplies over all phases must vanish, then

$$\sum_{i=1}^{n} \mathbf{M}_i = \mathbf{0}. \tag{2.15}$$

Thus, the summation of (2.14) over all phases implies that the *inner* part of the mixture stress tensor, which is defined by Truesdell and Toupin [42] as

$$\boldsymbol{\sigma}_I = \sum_{i=1}^{n} \boldsymbol{\sigma}_i, \tag{2.16}$$

is symmetric. Noting that the quantity

$$\sum_{i=1}^{n} \phi_i \rho_i \mathbf{w}_i \otimes \mathbf{w}_i$$

must also be symmetric implies that the mixture stress tensor is symmetric. Note, however, that the partial Cauchy stress tensors are symmetric if and only if $\mathbf{m}_i = \mathbf{0}$ (and hence $\mathbf{M}_i = \mathbf{0}$), that is, for nonpolar materials. In this particular study, it is assumed that the components of the growing tissue do behave as nonpolar materials, so that $\mathbf{m}_i = \mathbf{0}$ and that

$$\boldsymbol{\sigma}_i = \boldsymbol{\sigma}_i^T. \tag{2.17}$$

A more general theory would have to be developed to consider tissues comprising micropolar fluids.

2.4. Balance of energy. The energy balance for the ith phase of an n-phase mixture is given by

$$\phi_i \rho_i \frac{D^i u_i}{Dt} = tr(\mathbf{L}_i \boldsymbol{\sigma}_i) - \boldsymbol{\nabla} \cdot \mathbf{q}_i + \phi_i \rho_i r_i + \varepsilon_i, \tag{2.18}$$

or, equivalently,

$$\frac{\partial}{\partial t}(\phi_i \rho_i u_i) + \boldsymbol{\nabla} \cdot (\phi_i \rho_i u_i \mathbf{v}_i) = tr(\mathbf{L}_i \boldsymbol{\sigma}_i) - \boldsymbol{\nabla} \cdot \mathbf{q}_i + \phi_i \rho_i r_i + \varepsilon_i + \Gamma_i u_i, \tag{2.19}$$

where $\mathbf{q}_i$ is a measure of the rate of heat flow across a unit area from the ith constituent, r_i is the rate of heat production per unit mass within the ith constituent, ε_i is the energy supply per unit mass per unit time to the ith constituent due to energy exchange between the constituents, u_i is the internal energy per unit mass of the ith constituent, and $\mathbf{L}_i$ is the velocity gradient of the ith constituent with respect to spatial coordinates.

Truesdell and Toupin [42] argue that for the overall conservation of energy in the mixture, "the energy supplied by an excess internal energy rate, plus the energy supplied by the work of the excess inertial forces against diffusion, plus the energy supplied by the creation of mass, must add up to zero for the mixture." This implies the following expression for the conservation of energy:

$$\sum_{i=1}^{n} \left[\varepsilon_i + \Gamma_i \left(u_i + \frac{1}{2} \mathbf{w}_i \cdot \mathbf{w}_i \right) + \mathbf{w}_i \cdot \pi_i \right] = 0,$$

or, equivalently, by appealing to (2.12),

$$\sum_{i=1}^{n}\left[\varepsilon_i+\Gamma_i\left(u_i+\frac{1}{2}\mathbf{v}_i\cdot\mathbf{v}_i\right)+\mathbf{v}_i\cdot\pi_i\right]=0. \tag{2.20}$$

3. The second law of thermodynamics. The second law of thermodynamics, which may be expressed in the form of the Clausius–Duhem inequality, places limitations on the admissible paths of thermodynamic processes, thereby placing restrictions on constitutive equations. The Clausius–Duhem inequality states that the rate of entropy increase is greater than or equal to the entropy input rate.

Following Rajagopal and Tao [37], it is assumed that the second law of thermodynamics holds for the mixture as a whole. In addition, a single, spatially uniform temperature is assumed for all phases since growth involves exchanges of mass among the phases and because the growth process itself is slow in comparison with the time it would take for any possible temperature gradients to equilibrate. Indeed, it is unlikely that stresses arising from a gradient of thermal expansion would be significant in biological tissues. Therefore if η_i denotes the entropy per unit mass of the ith constituent and θ denotes the absolute temperature of the mixture, then the inequality may be expressed by

$$\sum_i\left[\frac{D^i}{Dt}(\phi_i\rho_i\eta_i)+\phi_i\rho_i\eta_i\boldsymbol{\nabla}.\mathbf{v}_i+\boldsymbol{\nabla}.\left(\frac{\mathbf{q}_i}{\theta}\right)-\frac{\phi_i\rho_i r_i}{\theta}\right]\geq 0. \tag{3.1}$$

The entropy inequality for the mixture is given by

$$\rho_m\frac{D\eta_m}{Dt}+\sum_i\boldsymbol{\nabla}.\left(\frac{\mathbf{h}_i}{\theta}\right)-\sum_i\left(\frac{\phi_i\rho_i r_i}{\theta}\right)\geq 0, \tag{3.2}$$

where $\mathbf{h}_i$ is an influx vector for the ith constituent—as yet unrelated to $\mathbf{q}_i$—and η_m is the entropy density for the mixture defined by

$$\eta_m=\frac{1}{\rho_m}\sum_i\phi_i\rho_i\eta_i.$$

Hence, reconciling (3.1) and (3.2) requires the constitutive postulate,

$$\mathbf{h}_i=\mathbf{q}_i+\phi_i\rho_i\theta\eta_i\mathbf{w}_i. \tag{3.3}$$

The second axiom of thermodynamics, as expressed by (3.1), may now be manipulated further to obtain a form from which constitutive equations may be deduced readily. To this end, the internal energies, u_i, will be eliminated in favor of the Helmholtz free energy densities, ψ_i, where

$$\psi_i=u_i-\theta\eta_i. \tag{3.4}$$

Employing the Helmholtz free energy is particularly expedient when deducing constitutive equations since it is the portion of the internal energy available for doing mechanical work at constant temperature [30]. Further, the process of deducing constitutive equations is facilitated by the introduction of the chemical potential for each phase, which, for *general* mixtures, is given by the linear transformation

$$\mathbf{K}_i=\psi_i\mathbf{I}-\frac{\boldsymbol{\sigma}_i}{\phi_i\rho_i}.$$

The change of variables brought about by this transformation is common in the established literature relating to thermodynamic theories of constitutive equations. (See Bowen [11, 12] and Bowen and Wiese [14] for further details on the use of the chemical potential tensor in the study of general mixtures.)

Now, incorporating the balance of mass (2.1) and the energy equation (2.18) and introducing the variables ψ_i and $\mathbf{K}_i$ enable the second law of thermodynamics to be expressed by the dissipation inequality

$$-tr\sum_{i=1}^{n}\phi_i\rho_i\mathbf{K}_i.\mathbf{L}_i-\rho_m\eta_m\frac{D\theta}{Dt}-\sum_{i=1}^{n}\frac{D^i\Psi_i}{Dt}-\sum_{i=1}^{n}\mathbf{v}_i.\left(\boldsymbol{\pi}_i+\frac{\Gamma_i}{2}\mathbf{v}_i\right)\geq 0, \tag{3.5}$$

where

$$\Psi_i=\phi_i\rho_i\psi_i \tag{3.6}$$

represents the Helmholtz free energy of the ith constituent per unit mixture volume. A full derivation of this inequality is given in the appendix.

4. The assumption of incompressibility. In the present paper it will be assumed that each of the n phases is intrinsically incompressible, thereby placing an added constraint on their motion and giving rise to an indeterminacy in the second law of thermodynamics. The assumption of incompressibility is a common one in mathematical models of biological tissues on account of the high water content of the cells and interstitial fluid and the very low compressibility of other extracellular constituents, such as the large macromolecules comprising the extracellular matrix [1].

The balance of mass for the ith phase may now be expressed by

$$\frac{\partial\phi_i}{\partial t}+\boldsymbol{\nabla}.(\phi_i\mathbf{v}_i)=\frac{\Gamma_i}{\rho_i},$$

the summation of which over the n phases gives

$$\sum_{i=1}^{n}(\boldsymbol{\nabla}\phi_i.\mathbf{v}_i+\phi_i\boldsymbol{\nabla}.\mathbf{v}_i)=\sum_{i=1}^{n}\frac{\Gamma_i}{\rho_i}\triangleq\hat{\gamma}, \tag{4.1}$$

employing the saturation constraint (2.4). At this point, the principle of material frame-indifference (or *objectivity*) is considered, which requires that the response of the material and its individual constituents (and hence its constitutive equations, to be developed later from the present analysis) be independent of the observer [6]. Since relative velocities are objective, while individual velocities are not, (4.1) may be expressed in terms of relative velocities by noting that

$$\sum_{i=1}^{n}\mathbf{v}_i.\boldsymbol{\nabla}\phi_i=\sum_{i=1}^{n}(\mathbf{v}_i-\mathbf{v}_1).\boldsymbol{\nabla}\phi_i,$$

where one phase is nominated as the reference phase with the subscript 1. Assuming that the densities of all phases are equal (so that $\rho_i=\rho_m=\rho$ and $\hat{\gamma}=0$) enables (4.1) to reduce to

$$\sum_{i=1}^{n}\Big[(\mathbf{v}_i-\mathbf{v}_1).\boldsymbol{\nabla}\phi_i+\phi_i tr\mathbf{L}_i\Big]=0. \tag{4.2}$$

From a mathematical standpoint, this assumption of equal phase densities allows the model to isolate the growth-induced stresses arising from spatially nonuniform (incompatible) growth, without the potentially confounding effects of additional stresses associated with density changes. In addition, the assumption of equal densities simplifies the ensuing analysis considerably. Further, the argument may be justified from a phenomenological point of view by noting that in a growing tissue, the growth process itself arises from exchanges of mass among individual tissue constituents. In particular, cells grow and proliferate by taking in interstitial fluid—water and proteins (and other molecules contained in the interstitial fluid)—and relinquish these substances on cell death. Thus, while different phases may exhibit fundamentally different mechanical behavior, they are composed of similar substances.

Now, recognizing that

$$\mathbf{L}_i = \grave{\mathbf{F}}_i \mathbf{F}_i^{-1}$$

by the chain rule, where

$$\grave{\mathbf{F}}_i = \frac{D^i \mathbf{F}_i}{Dt},$$

and that

$$\sum_{i=1}^{n} \mathbf{v}_i . \left(\boldsymbol{\pi}_i + \frac{\Gamma_i}{2} \mathbf{v}_i \right) = \sum_{i=1}^{n} \left[\boldsymbol{\pi}_i + \frac{1}{2} \Gamma_i (\mathbf{v}_i - \mathbf{v}_1) \right] . (\mathbf{v}_i - \mathbf{v}_1) \tag{4.3}$$

now enables the second axiom of thermodynamics to be expressed in the form

$$\begin{aligned} &-tr \sum_{i=1}^{n} \mathbf{F}_i^{-1} (\phi_i \rho \mathbf{K}_i - \phi_i \zeta \mathbf{I}) \grave{\mathbf{F}}_i - \rho \eta_m \frac{D\theta}{Dt} - \sum_{i=1}^{n} \frac{D^i \Psi_i}{Dt} \\ &- \sum_{i=1}^{n} \left[\boldsymbol{\pi}_i + \frac{1}{2} \Gamma_i (\mathbf{v}_i - \mathbf{v}_1) - \zeta \boldsymbol{\nabla} \phi_i \right] . (\mathbf{v}_i - \mathbf{v}_1) \geq 0, \end{aligned} \tag{4.4}$$

where ζ is a Lagrangian multiplier.

5. Constitutive assumptions for a general n-phase mixture. As expressed by Coleman and Noll in [17], "a material is defined by a constitutive assumption, which is a restriction on the processes that are admissible in a body consisting of the material."

In discussing the various principles governing constitutive equations, Passman and Nunziato [33] describe the principle of equipresence as "too general," claiming that it is "difficult to accept as a universal axiom appropriate to all mixture theories." (According to this principle, "all dependent variables depend on all independent variables, unless the entropy inequality requires otherwise" [25, 41].) They proceed to explain that "in multiphase mixtures (where) the individual constituents are clearly separated physically, . . . it is plausible to think of the mixture as being ideal, or phase separated. For such mixtures the Principle of Equipresence can reasonably be replaced by the Principle of Phase Separation." By this principle, the material-specific dependent variables of a given phase (such as the stress and the Helmholtz free energy density) depend only on the independent variables of that phase. The interaction variables (such as the momentum transfer term, $\boldsymbol{\pi}_i$) depend on all the independent variables. (See Passman and Nunziato [33] for a more detailed discussion of these principles.)

Much of the classical work in this field has relied on the former, more general principle. Thus, in predicating the current study on these classical, well-established approaches, this paper appeals to this general principle in deducing thermodynamically appropriate constitutive equations in spite of the fact that the individual phases of biological tissues are clearly separated and distinct. Nevertheless, the simpler principle of phase separation is used to advantage in subsequent analysis, allowing the constitutive equations to be manipulated into useable forms.

Furthermore, Ehlers [19] emphasises the fact that "the general constitutive framework must be based on the assumption of second-grade materials . . . , thus making use of the most natural framework in constitutive modelling for multiphase media, additionally avoiding so-called 'simple' results." Thus, following Bowen [12], Bowen and Weise [14], and Ehlers [19], and noting from (4.2) that the constitutive assumptions for $\mathbf{K}_i$ and $\boldsymbol{\pi}_i$ must reflect an indeterminacy consistent with the entropy inequality, the following general constitutive postulate is proposed:

$$\begin{aligned}\left(\Psi_i, \eta_i, \left(\boldsymbol{\pi}_i + \frac{1}{2}\Gamma_i(\mathbf{v}_i - \mathbf{v}_1) - \zeta\boldsymbol{\nabla}\phi_i\right), (\phi_i\rho\mathbf{K}_i - \phi_i\zeta\mathbf{I}), \mathbf{q}\right)\\ = f(\theta, \mathbf{F}_j, \dot{\mathbf{F}}_j, \mathbf{G}_j, \phi_j, \mathbf{n}_j, (\mathbf{v}_j - \mathbf{v}_1)),\end{aligned} \tag{5.1}$$

where f is a smooth function, with the following quantities being defined for clarity:

$$\mathbf{G}_j = \mathbf{Grad}\mathbf{F}_j$$

and

$$\mathbf{n}_j = \boldsymbol{\nabla}\phi_j.$$

As discussed by Bowen in [10] and [12], (5.1) describes a mixture which allows for the combined effects of elasticity, heat conduction, diffusion, viscosity, buoyancy, immiscibility, and variable volume fractions. As noted by Bowen in [13], "an *immiscible* mixture is one where locally one can distinguish between mixture volumes and constituent volumes (and therefore) a model of an immiscible mixture would necessarily allow the volume fractions to effect the mixture response." Having established a general framework, then, it remains for a *particular* constitutive postulate to be chosen to carry the analysis through to completion, to arrive at a full set of modeling equations.

6. A biphasic mechanical model of tissue growth. In this section, the general constitutive assumption (5.1) is applied to a two-phase model comprising an elastic solid (indicated by the subscript s) and an inviscid fluid (indicated by the subscript f), being the simplest case of a solid-multiphase model. In this case, the constitutive equations reduce to

$$\left(\Psi_i, \eta_i, \left(\boldsymbol{\pi}_f + \frac{1}{2}\Gamma_f(\mathbf{v}_f - \mathbf{v}_s) - \zeta\boldsymbol{\nabla}\phi_f\right), (\phi_i\rho\mathbf{K}_i - \phi_i\zeta\mathbf{I}), \mathbf{q}\right) = f(\theta, \mathbf{F}_s, \mathbf{G}_s, (\mathbf{v}_f - \mathbf{v}_s)) \tag{6.1}$$

with $i = f, s$. Since the effect of viscosity is not being considered in this simplified model, the derivatives of the deformation gradients do not appear among the independent variables in (6.1). The volume fractions and their gradients are also omitted from the set of independent variables since the *specific* Helmholtz free energy, ψ_i, is to

be considered independent of volume fraction, with the volume-averaged Helmholtz free energy, Ψ_i, being related to volume fraction via (3.6). Note that in a two-phase model, only one of the mass exchange terms, Γ_i, or volume fraction terms, ϕ_i, need be considered since the constraints (2.4) and (2.6) give the corresponding terms for the other phase. Further, since one of the phases is a solid and the other a fluid, the volume fractions ϕ_s and $\phi_f = 1 - \phi_s$ will henceforth be referred to as the *solidity* and the *porosity*, respectively.

Using (6.1) the total derivative of the Helmholtz free energy for the solid is given by

$$\begin{aligned}\frac{D^s\Psi_s}{Dt} &= \left(\frac{\partial\Psi_s}{\partial\theta}\right)\left(\frac{D\theta}{Dt}\right) + tr\left(\frac{\partial\Psi_s}{\partial\mathbf{F}_s}\right)^T \grave{\mathbf{F}}_s + C\left(\frac{\partial\Psi_s}{\partial\mathbf{G}_s}\right)\otimes\grave{\mathbf{G}}_s \\ &\quad + \left(\frac{\partial\Psi_s}{\partial(\mathbf{v}_f-\mathbf{v}_s)}\right).\left[\frac{D^f(\mathbf{v}_f-\mathbf{v}_s)}{Dt} + \grave{\mathbf{F}}_f\mathbf{F}_f^{-1}(\mathbf{v}_s-\mathbf{v}_f) - \grave{\mathbf{F}}_s\mathbf{F}_s^{-1}(\mathbf{v}_s-\mathbf{v}_f)\right],\end{aligned}$$

while the total derivative of the Helmholtz free energy for the fluid is given by

$$\begin{aligned}\frac{D^f\Psi_f}{Dt} &= \left(\frac{\partial\Psi_f}{\partial\theta}\right)\left(\frac{D\theta}{Dt}\right) + tr\left(\frac{\partial\Psi_f}{\partial\mathbf{F}_s}\right)^T \left[\grave{\mathbf{F}}_s + \mathbf{G}_s\mathbf{F}_s^{-1}(\mathbf{v}_f-\mathbf{v}_s)\right] \\ &\quad + C\left(\frac{\partial\Psi_f}{\partial\mathbf{G}_s}\right)\otimes\left[\grave{\mathbf{G}}_s + (\mathbf{GradG}_s)\mathbf{F}_s^{-1}(\mathbf{v}_f-\mathbf{v}_s)\right] \\ &\quad + \left(\frac{\partial\Psi_f}{\partial(\mathbf{v}_f-\mathbf{v}_s)}\right).\left[\frac{D^f(\mathbf{v}_f-\mathbf{v}_s)}{Dt}\right].\end{aligned}$$

Therefore the entropy inequality becomes

$$\begin{aligned}&-tr\sum_{i=f,s}\mathbf{F}_i^{-1}\left(\phi_i\rho\mathbf{K}_i - \phi_i\zeta\mathbf{I}\right)\grave{\mathbf{F}}_i - \rho\eta_m\frac{d\theta}{dt} \\ &-\left(\boldsymbol{\pi}_f + \frac{1}{2}\Gamma_f(\mathbf{v}_f-\mathbf{v}_s) - \zeta\boldsymbol{\nabla}\phi_f\right).(\mathbf{v}_f-\mathbf{v}_s) - \left(\frac{\partial\Psi_I}{\partial\theta}\right)\frac{D\theta}{Dt} \\ &-tr\mathbf{F}_s^{-1}\left(\mathbf{F}_s\left(\frac{\partial\Psi_I}{\partial\mathbf{F}_s}\right)^T - (\mathbf{v}_s-\mathbf{v}_f)\otimes\left(\frac{\partial\Psi_s}{\partial(\mathbf{v}_f-\mathbf{v}_s)}\right)\right)\grave{\mathbf{F}}_s \\ &-(\mathbf{v}_f-\mathbf{v}_s).\left(\mathbf{F}_s^{-1T}\left(\frac{\partial\Psi_f}{\partial\mathbf{F}_s}[\mathbf{G}_s]\right)\right) \\ &-C\left(\frac{\partial\Psi_I}{\partial\mathbf{G}_s}\right)\otimes\grave{\mathbf{G}}_s - (\mathbf{v}_f-\mathbf{v}_s).\left(\mathbf{F}_s^{-1T}\left(\frac{\partial\Psi_f}{\partial\mathbf{G}_s}[\mathbf{GradG}_s]\right)\right) \\ &-\left(\frac{\partial\Psi_I}{\partial(\mathbf{v}_f-\mathbf{v}_s)}\right).\left[\frac{D^f(\mathbf{v}_f-\mathbf{v}_s)}{Dt}\right] \\ &-tr\mathbf{F}_f^{-1}\left((\mathbf{v}_s-\mathbf{v}_f)\otimes\left(\frac{\partial\Psi_s}{\partial(\mathbf{v}_f-\mathbf{v}_s)}\right)\right)\grave{\mathbf{F}}_f \geq 0,\end{aligned}$$

where

$$\Psi_I = \sum_{i=f,s}\Psi_i = \sum_{i=f,s}\phi_i\rho_i\psi_i$$

denotes the inner part of the mixture Helmholtz free energy. Here, the notation $\mathbf{X}[\mathbf{Y}]$, where $\mathbf{X}$ is a tensor of rank p and $\mathbf{Y}$ is a tensor of rank $p+1$, denotes a vector defined

in component form by

$$\mathbf{X}[\mathbf{Y}] = X_{k_1 k_2 \ldots k_p} Y^{k_1 k_2 \ldots k_p q} \mathbf{e}_q.$$

where $\mathbf{e}_q$ are basis vectors. (See, for example, (1.10) in Bowen and Weise [14] or (1.1.58) in Bowen [12].) Now rearranging the inequality produces

$$\begin{aligned}
&-tr\mathbf{F}_s^{-1}\left(\phi\rho_s\mathbf{K}_s - \phi_s\zeta\mathbf{I} + \mathbf{F}_s\left(\frac{\partial\Psi_I}{\partial\mathbf{F}_s}\right)^T - (\mathbf{v}_s - \mathbf{v}_f)\otimes\left(\frac{\partial\Psi_s}{\partial(\mathbf{v}_f - \mathbf{v}_s)}\right)\right)\dot{\mathbf{F}}_s \\
&- tr\mathbf{F}_f^{-1}\left(\phi_f\rho\mathbf{K}_f - \phi_f\zeta\mathbf{I} + (\mathbf{v}_s - \mathbf{v}_f)\otimes\left(\frac{\partial\Psi_s}{\partial(\mathbf{v}_f - \mathbf{v}_s)}\right)\right)\dot{\mathbf{F}}_f \\
&- \left(\rho\eta_m + \frac{\partial\Psi_I}{\partial\theta}\right)\frac{D\theta}{Dt} - \left(\frac{\partial\Psi_I}{\partial\mathbf{g}}\right).\frac{D\mathbf{g}}{Dt} - C\left(\frac{\partial\Psi_I}{\partial\mathbf{G}_s}\right)\otimes\dot{\mathbf{G}}_s \\
&- (\mathbf{v}_f - \mathbf{v}_s).\left(\mathbf{F}_s^{-1T}\left(\frac{\partial\Psi_f}{\partial\mathbf{G}_s}[\mathbf{GradG}_s]\right)\right) - \left(\frac{\partial\Psi_I}{\partial(\mathbf{v}_f - \mathbf{v}_s)}\right).\left[\frac{D^f(\mathbf{v}_f - \mathbf{v}_s)}{Dt}\right] \\
&- \left[\boldsymbol{\pi}_f + \frac{1}{2}\Gamma_f(\mathbf{v}_f - \mathbf{v}_s) - \zeta\boldsymbol{\nabla}\phi_f + \mathbf{F}_s^{-1T}\left(\frac{\partial\Psi_f}{\partial\mathbf{F}_s}[\mathbf{G}_s]\right)\right].(\mathbf{v}_f - \mathbf{v}_s) \geq 0.
\end{aligned}$$

Following Coleman and Noll's argument [17],

$$\theta,\ \mathbf{F}_s,\ \mathbf{G}_s \text{ and } (\mathbf{v}_f - \mathbf{v}_s)$$

are held *fixed* while *varying* the quantities

$$\frac{D\theta}{Dt},\ \dot{\mathbf{F}}_s,\ \dot{\mathbf{G}}_s,\ \mathbf{GradG}_s \text{ and } \frac{D^f(\mathbf{v}_f - \mathbf{v}_s)}{Dt}.$$

This yields the following necessary and sufficient conditions:

$$\phi_s\rho\mathbf{K}_s - \phi_s\zeta\mathbf{I} + \mathbf{F}_s\left(\frac{\partial\Psi_I}{\partial\mathbf{F}_s}\right)^T - (\mathbf{v}_s - \mathbf{v}_f)\otimes\left(\frac{\partial\Psi_s}{\partial(\mathbf{v}_f - \mathbf{v}_s)}\right) = \mathbf{0}, \tag{6.2}$$

$$\phi_f\rho\mathbf{K}_f - \phi_f\zeta\mathbf{I} + (\mathbf{v}_s - \mathbf{v}_f)\otimes\left(\frac{\partial\Psi_s}{\partial(\mathbf{v}_f - \mathbf{v}_s)}\right) = \mathbf{0}, \tag{6.3}$$

$$\rho\eta_m = -\frac{\partial\Psi_I}{\partial\theta},$$

$$\frac{\partial\Psi_I}{\partial\mathbf{G}_s} = \mathbf{0},$$

$$\sum_{i=f,s}(\mathbf{v}_i - \mathbf{v}_s).\left(\mathbf{F}_s^{-1T}\left(\frac{\partial\Psi_f}{\partial\mathbf{G}_s}[\mathbf{GradG}_s]\right)\right) = \mathbf{0},$$

$$\frac{\partial\Psi_I}{\partial(\mathbf{v}_f - \mathbf{v}_s)} = \mathbf{0},$$

and

$$-\mathbf{f}.(\mathbf{v}_f - \mathbf{v}_s) \geq 0, \tag{6.4}$$

where

$$\mathbf{f} = \boldsymbol{\pi}_f + \frac{1}{2}\Gamma_f(\mathbf{v}_f - \mathbf{v}_s) - \zeta\boldsymbol{\nabla}\phi_f + \mathbf{F}_s^{-1T}\left(\frac{\partial\Psi_f}{\partial\mathbf{F}_s}[\mathbf{G}_s]\right).$$

7. Development of linearized constitutive equations. In this section, the constitutive equations (6.2), (6.3), and (6.4) deduced in the previous section will be linearized about the thermodynamic equilibrium using the method discussed by Bowen [12], in order to express the equations in useable forms.

Let $\hat{\xi}$ denote the complete set of independent variables

$$\hat{\xi} = (\mathbf{F}_s, \mathbf{G}_s, (\mathbf{v}_f - \mathbf{v}_s)),$$

noting that there is no longer any dependence on temperature or temperature gradients. Let $\hat{\xi}^{0E}$ denote the subset of these variables,

$$\hat{\xi}^{0E} = (\mathbf{F}_s, \mathbf{G}_s, \mathbf{0}).$$

Let $\hat{\xi}^{0R}$ denote the reference state about which the constitutive equations for $\mathbf{f}$ and $\boldsymbol{\sigma}_s$ will be linearized,

$$\hat{\xi}^{0R} = (\mathbf{I}, \mathbf{0}, \mathbf{0}).$$

At the state $\hat{\xi}^{0E}$, where $\mathbf{v}_s = \mathbf{v}_f$, the quantity

$$-\mathbf{f}.(\mathbf{v}_f - \mathbf{v}_s) \geq 0$$

is a minimum, such that $\hat{\xi}^{0E}$ defines the thermodynamic equilibrium. Now in the vicinity of the thermodynamic equilibrium,

$$\phi_s \rho \mathbf{K}_s = \phi_s \zeta \mathbf{I} - \mathbf{F}_s \left(\frac{\partial \Psi_I}{\partial \mathbf{F}_s} \right)^T,$$

$$\phi_f \rho \mathbf{K}_f = \phi_f \zeta \mathbf{I},$$

and

$$\mathbf{f} = -\kappa(\mathbf{v}_s - \mathbf{v}_f),$$

where κ is a constant, sometimes referred to as the diffusive drag coefficient. Therefore,

$$\boldsymbol{\sigma}_s = \phi_s \rho \psi_s \mathbf{I} - \phi_s \zeta \mathbf{I} + \mathbf{F}_s \left(\frac{\partial \Psi_I}{\partial \mathbf{F}_s} \right)^T, \tag{7.1}$$

$$\boldsymbol{\sigma}_f = \phi_f \rho \psi_f \mathbf{I} - \phi_f \zeta \mathbf{I}, \tag{7.2}$$

and

$$\boldsymbol{\pi}_f = -\kappa(\mathbf{v}_s - \mathbf{v}_f) - \frac{1}{2}\Gamma_f(\mathbf{v}_s - \mathbf{v}_f) + \zeta \boldsymbol{\nabla} \phi_f - \mathbf{F}_s^{-1T} \left(\frac{\partial \Psi_f}{\partial \mathbf{F}_s}[\mathbf{G}_s] \right). \tag{7.3}$$

Now, appealing to the principle of phase separation,

$$\begin{aligned} \boldsymbol{\sigma}_s &= (\phi_s \rho \psi_s - \phi_s \zeta)\mathbf{I} + \rho \mathbf{F}_s \left(\frac{\partial (\phi_s \psi_s + \phi_f \psi_f)}{\partial \mathbf{F}_s} \right)^T \\ &= -\phi_s(\zeta - \rho \psi_s)\mathbf{I} + \rho \mathbf{F}_s \left(\frac{\partial \phi_s}{\partial \mathbf{F}_s}\psi_s + \phi_s \frac{\partial \psi_s}{\partial \mathbf{F}_s} - \frac{\partial \phi_s}{\partial \mathbf{F}_s}\psi_f \right)^T. \end{aligned}$$

Hence,

$$\boldsymbol{\sigma}_s = -\phi_s(\zeta - \rho\psi_s)\mathbf{I} + \phi_s\rho\mathbf{F}_s\left(\frac{\partial\psi_s}{\partial\mathbf{F}_s}\right)^T + \rho(\psi_s - \psi_f)\mathbf{F}_s\left(\frac{\partial\phi_s}{\partial\mathbf{F}_s}\right)^T. \tag{7.4}$$

Similarly, for the momentum transfer terms, the principle of phase separation gives

$$\boldsymbol{\pi}_f = -\kappa(\mathbf{v}_f - \mathbf{v}_s) - \frac{1}{2}\Gamma_f(\mathbf{v}_f - \mathbf{v}_s) + \zeta\boldsymbol{\nabla}\phi_f + \mathbf{F}_s^{-1T}\left(\rho\psi_f\left(\frac{\partial\phi_s}{\partial\mathbf{F}_s}\right)[\mathbf{G}_s]\right). \tag{7.5}$$

8. Modeling biological growth: Mass exchanges, solid deformation, and fluid flow. To reduce (7.4) to a usable form, an expression for $\frac{\partial\phi_s}{\partial\mathbf{F}_s}$ must be deduced from the balance of mass for the solid phase, which is given by

$$\frac{D^s}{Dt}(\rho\phi_s\mathrm{det}\mathbf{F}_s) = \Gamma_s\mathrm{det}\mathbf{F}_s. \tag{8.1}$$

Thus, in general

$$\rho\phi_s\mathrm{det}\mathbf{F}_s = \int_0^t \Gamma_s(\mathbf{X}_s,\tau)\mathrm{det}\mathbf{F}_s(\mathbf{X}_s,\tau)d\tau \triangleq \hat{\Theta}_s, \tag{8.2}$$

where

$$\hat{\Theta}_s = \hat{\Theta}_s(\mathbf{X}_s, t),$$

and $\mathbf{X}_s$ denotes the reference coordinates. Hence, using Jacobi's identity [42],

$$\frac{\partial(\mathrm{det}\mathbf{F}_s)}{\partial\mathbf{F}_s} = (\mathrm{det}\mathbf{F}_s)\mathbf{F}_s^{-1T},$$

the derivative of the solidity with respect to the solid deformation gradient is given by

$$\frac{\partial\phi_s}{\partial\mathbf{F}_s} = -\phi_s\mathbf{F}_s^{-1T} + \frac{1}{\rho\mathrm{det}\mathbf{F}_s}\frac{\partial\hat{\Theta}_s}{\partial\mathbf{F}_s}. \tag{8.3}$$

Therefore, (7.4) becomes

$$\boldsymbol{\sigma}_s = -\phi_s(\zeta - \rho\psi_s)\mathbf{I} + \phi_s\rho\mathbf{F}_s\left(\frac{\partial\psi_s}{\partial\mathbf{F}_s}\right)^T - \phi_s\rho(\psi_s - \psi_f) + \frac{(\psi_s - \psi_f)}{\mathrm{det}\mathbf{F}_s}\mathbf{F}_s\left(\frac{\partial\hat{\Theta}_s}{\partial\mathbf{F}_s}\right)^T. \tag{8.4}$$

Thus the constitutive equations are

$$\boldsymbol{\sigma}_s = -\phi_s P\mathbf{I} + \phi_s\rho\mathbf{F}_s\left(\frac{\partial\psi_s}{\partial\mathbf{F}_s}\right)^T + \frac{(\psi_s - \psi_f)}{\mathrm{det}\mathbf{F}_s}\mathbf{F}_s\left(\frac{\partial\hat{\Theta}_s}{\partial\mathbf{F}_s}\right)^T \tag{8.5}$$

and

$$\boldsymbol{\sigma}_f = -\phi_f P\mathbf{I}, \tag{8.6}$$

where

$$P = \zeta - \rho\psi_f. \tag{8.7}$$

In addition,

$$\frac{\partial\phi_s}{\partial\mathbf{X}_s} = \frac{\partial\phi_s}{\partial\mathbf{F}_s}\left[\frac{\partial\mathbf{F}_s}{\partial\mathbf{X}_s}\right] = \left(-\phi_s\mathbf{F}_s^{-1T} + \frac{1}{\rho\mathrm{det}\mathbf{F}_s}\frac{\partial\hat{\Theta}_s}{\partial\mathbf{F}_s}\right)[\mathbf{G}_s]. \tag{8.8}$$

Now

$$\boldsymbol{\nabla}\phi_s = \frac{\partial \phi_s}{\partial \mathbf{x}_s} = \mathbf{F}_s^{-1T}\frac{\partial \phi_s}{\partial \mathbf{X}_s}, \tag{8.9}$$

where $\mathbf{x}_s$ denotes spatial coordinates. Hence

$$\boldsymbol{\nabla}\phi_s = \mathbf{F}_s^{-1T}\left(\frac{\partial \phi_s}{\partial \mathbf{F}_s}\right)[\mathbf{G}_s] = \mathbf{F}_s^{-1T}\left(-\phi_s\mathbf{F}_s^{-1T} + \frac{1}{\rho \det\mathbf{F}_s}\frac{\partial \hat{\Theta}_s}{\partial \mathbf{F}_s}\right)[\mathbf{G}_s].$$

Thus, (7.5) for the momentum transfer term for the fluid phase, $\boldsymbol{\pi}_f$, now becomes

$$\begin{aligned}\boldsymbol{\pi}_f &= -\kappa(\mathbf{v}_f - \mathbf{v}_s) - \frac{1}{2}\Gamma_f(\mathbf{v}_f - \mathbf{v}_s) + \zeta\boldsymbol{\nabla}\phi_f + \rho\psi_f\boldsymbol{\nabla}\phi_s \\ &= P\boldsymbol{\nabla}\phi_f - \kappa(\mathbf{v}_f - \mathbf{v}_s) - \frac{1}{2}\Gamma_f(\mathbf{v}_f - \mathbf{v}_s).\end{aligned} \tag{8.10}$$

Noting the conservation of linear momentum,

$$\boldsymbol{\pi}_f + \Gamma_f\mathbf{v}_f + \boldsymbol{\pi}_s + \Gamma_s\mathbf{v}_s = \mathbf{0}$$

then gives the momentum transfer term for the solid phase, $\boldsymbol{\pi}_s$, being

$$\boldsymbol{\pi}_s = P\boldsymbol{\nabla}\phi_s + \kappa(\mathbf{v}_f - \mathbf{v}_s) - \frac{1}{2}\Gamma_f(\mathbf{v}_f - \mathbf{v}_s). \tag{8.11}$$

In a detailed analysis of Darcy's law for growing porous media, Preziosi and Farina [36] have shown that the process of interphase mass exchange plays a negligible role in momentum transfer. Thus, the final term in each of (8.10) and (8.11) may be neglected in practice.

Returning to the constitutive equation for the stress in the solid (8.5), note that the quantity $\hat{\Theta}_s$ depends on both the mass exchange term, Γ_s, and the solid phase deformation, $\mathbf{F}_s$. Clearly, then, manipulation of this equation into a useable form requires a phenomenological assumption about the functional form of the mass exchange term. Furthermore, the distinguishing feature of a mass-exchange model which describes biological growth, as opposed to, say, models which describe solidification/melting or some other phase change, is the fact that the mass-exchange term, Γ_s, and the expansion/contraction of the solid matrix, $\det\mathbf{F}_s$, *are not independent*. Indeed, such is the very essence of the unique closure problem peculiar to the study of biological growth, and it is a novel feature of the present work that such phenomenological aspects of growing tissues may be incorporated into the modeling framework.

Consider that the balance of mass for the solid phase may also be expressed in the form

$$\rho\frac{D^s\phi_s}{Dt} + \phi_s\rho\frac{1}{\det\mathbf{F}_s}\frac{D^s(\det\mathbf{F}_s)}{Dt} = \Gamma_s. \tag{8.12}$$

Note that

$$\rho\frac{D^s\phi_s}{Dt} = \frac{D^s}{Dt}\left(\frac{m_s}{V_m}\right),$$

where m_s is the mass of the solid phase in V_m, the volume of the mixture. Now

$$\frac{D^s}{Dt}\left(\frac{m_s}{V_m}\right) = \frac{D^s m_s}{Dt}\left(\frac{1}{V_m}\right) - \frac{\phi_s\rho}{V_m}\frac{D^s V_m}{Dt}.$$

Since

$$\frac{D^s m_s}{Dt}\left(\frac{1}{V_m}\right) = \Gamma_s,$$

the quantity

$$-\frac{\phi_s \rho}{V_m}\frac{D^s V_m}{Dt}$$

reflects the time rate of change of the solidity which results from the flow of fluid into, or out of, the deformed solid matrix. Now since

$$\Gamma_s = \underbrace{\frac{D^s}{Dt}\left(\frac{m_s}{V_m}\right)}_{total} - \underbrace{\left(-\frac{\phi_s \rho}{V_m}\frac{D^s V_m}{Dt}\right)}_{flow/deformation},$$

then

$$\underbrace{\frac{D^s(\phi_s \rho)}{Dt}}_{total} + \underbrace{\phi_s \rho \frac{1}{\det \mathbf{F}_s}\frac{D^s(\det \mathbf{F}_s)}{Dt}}_{flow/deformation} = \underbrace{\frac{D^s}{Dt}\left(\frac{m_s}{V_m}\right)}_{total} + \underbrace{\left(\frac{\phi_s \rho}{V_m}\frac{D^s V_m}{Dt}\right)}_{flow/deformation} = \Gamma_s$$

describes the mass balance for the solid phase.

Therefore, since the solidity is regulated by two separate processes—mass exchange and solid matrix deformation/fluid flow—a further constitutive postulate is required to relate any two of the three quantities ϕ_s, Γ_s, and $\det\mathbf{F}_s$ to decompose the balance of mass into two independent equations.

Suppose, for example, that the mass exchange and solid matrix deformation are related in such a way as to keep the volume fractions constant, reflecting a tissue which tends to exhibit a "natural" ratio of cells to extracellular fluid. (Note that this particular choice of constitutive postulate would be insufficient to model a growing tumor tissue which contains regions of coagulative necrosis, since these regions consist predominantly of fluid and cellular debris and are therefore characterized by significantly higher proportions of fluid than the rest of the tissue.) In this case, the balance of mass would be represented by

$$\rho \frac{D^s \phi_s}{Dt} = 0 \tag{8.13}$$

and

$$\rho\phi_s \frac{1}{\det\mathbf{F}_s}\frac{D^s(\det\mathbf{F}_s)}{Dt} = \rho\phi_s \boldsymbol{\nabla}\cdot\mathbf{v}_s = \Gamma_s. \tag{8.14}$$

Now that the constitutive postulate (8.14) has been proposed, the phenomenological assumption for Γ_s is no longer required to be a function of $\det\mathbf{F}_s$.

Assuming, for example, that Γ_s is proportional to the effective cell density, $\rho\phi_s$, and to some regulating factor such as nutrient concentration, c, implies that

$$\Gamma_s = \alpha\rho\phi_s c, \tag{8.15}$$

where α is a constant of proportionality and where the functional form for c will also be determined by a phenomenological assumption (appealing to a diffusion model, say).

Now (8.1) becomes

$$\frac{1}{\rho\phi_s \det\mathbf{F}_s}\frac{D^s}{Dt}(\rho\phi_s \det\mathbf{F}_s) = \alpha c,$$

so that

$$\rho\phi_s \det\mathbf{F}_s = e^{\int_0^t \alpha c(\mathbf{X}_s,\tau)d\tau},$$

which gives

$$\frac{\partial\phi_s}{\partial\mathbf{F}_s} = -\phi_s\mathbf{F}_s^{-1T}. \tag{8.16}$$

Now the constitutive equations reduce to

$$\boldsymbol{\sigma}_s = -\phi_s P\mathbf{I} + \phi_s\rho\mathbf{F}_s\left(\frac{\partial\psi_s}{\partial\mathbf{F}_s}\right)^T \tag{8.17}$$

and

$$\boldsymbol{\sigma}_f = -\phi_f P\mathbf{I}, \tag{8.18}$$

with P being given by (8.7). Substitution of the new definition for $\frac{\partial\phi_s}{\partial\mathbf{F}_s}$ into (8.8) and (8.9) then produces (8.10) and (8.11), illustrating that the momentum equations are unaffected by the simplified definition for $\frac{\partial\phi_s}{\partial\mathbf{F}_s}$.

8.1. Linear elasticity. If the solid phase is assumed to be elastically isotropic, the Helmholtz free energy density is a function of the solid deformation gradient, $\mathbf{F}_s$, through the left Cauchy–Green strain tensor defined by $\mathbf{B}_s = \mathbf{F}_s\mathbf{F}_s^T$, so that

$$\mathbf{F}_s\left(\frac{\partial\psi_s}{\partial\mathbf{F}_s}\right)^T = 2\mathbf{B}_s\frac{\partial\psi_s}{\partial\mathbf{B}_s}. \tag{8.19}$$

To formulate a linearized constitutive equation, an approximate expression is required for the right-hand side of (8.19) which is valid in the vicinity of the reference state $\hat{\xi}^{0R}$. Departures from $\hat{\xi}^{0R}$ may be measured by the quantity ϵ defined by

$$\epsilon^2 = tr\mathbf{H}_s\mathbf{H}_s^T + C\left(\mathbf{Grad}\mathbf{H_s}\otimes\mathbf{Grad}\mathbf{H_s}\right) + (\mathbf{v}_f - \mathbf{v}_s)\cdot(\mathbf{v}_f - \mathbf{v}_s),$$

where $\mathbf{H}_s = \mathbf{F}_s - \mathbf{I}$ is the displacement gradient of the solid phase. Thus, departure from the reference state, $\hat{\xi}^{0R}$, is *small* when $\epsilon < 1$.

Moreover,

$$\begin{aligned}\mathbf{B}_s &= \mathbf{I} + 2\mathbf{E}_s + \mathbf{H}_s\mathbf{H}_s^T\\ &= \mathbf{I} + 2\mathbf{E}_s + O(\epsilon^2),\end{aligned}$$

where $\mathbf{E}_s$ is the classical infinitesimal strain tensor defined by

$$\mathbf{E}_s = \frac{1}{2}\left(\mathbf{H}_s + \mathbf{H}_s^T\right). \tag{8.20}$$

Thus,

$$\mathbf{F}_s\left(\frac{\partial\psi_s}{\partial\mathbf{F}_s}\right)^T = \frac{\partial\psi_s}{\partial\mathbf{E}_s} + 2\mathbf{E}_s\left(\frac{\partial\psi_s}{\partial\mathbf{E}_s}(\hat{\xi}^{0R})\right) \tag{8.21}$$

in the linear theory. The Helmholtz free energy density for the solid phase, ψ_s, is now to be expanded into a polynomial about $\hat{\xi}^{0R}$, including terms up to second order since ψ_s must be differentiated to obtain the stress. Furthermore, since the Helmholtz free energy is, by definition, the energy available to do mechanical work, it is a function only of the component of the strain tensor associated with a stress response. Indeed, the strain tensor $\mathbf{E}_s$ may be decomposed into the contribution due to growth $\mathbf{E}_s^G$ and the contribution due to stress $\mathbf{E}_s^S$, i.e.,

$$\mathbf{E}_s = \mathbf{E}_s^G + \mathbf{E}_s^S.$$

Now,

$$\psi_s(\mathbf{B}_s) = \psi_s(\mathbf{I}) + \sigma_0\left(tr\mathbf{E}_s^S\right) + \frac{1}{2}\lambda_0\left(tr\mathbf{E}_s^S\right)^2 + \mu_0 tr\left(\mathbf{E}_s^S\mathbf{E}_s^S\right) + O(\epsilon^3). \tag{8.22}$$

Thus, substituting (8.22) into (8.21) yields

$$\boldsymbol{\sigma}_s = -\phi_s P\mathbf{I} + \lambda\left(tr\mathbf{E}_s^S\right)\mathbf{I} + 2\mu\mathbf{E}_s^S,$$

where

$$\lambda = \lambda(\rho_s, \phi_s) = \lambda_0$$

and

$$\mu = \mu(\rho_s, \phi_s) = \mu_0 + \sigma_0,$$

where σ_0, the so-called prestress, will be assumed zero. Now, the portion of the strain tensor due to growth may be expressed by

$$\mathbf{E}_s^G = g\boldsymbol{\Omega},$$

where g is the increase in volume per unit volume of the solid matrix due to growth (as yet unrelated to Γ_s), and

$$\boldsymbol{\Omega} \triangleq \begin{bmatrix} \gamma_1 & 0 & 0 \\ 0 & \gamma_2 & 0 \\ 0 & 0 & \gamma_3 \end{bmatrix} \tag{8.23}$$

defines the anisotropy tensor, where γ_1, γ_2, and γ_3 are the anisotropic growth multipliers defined by Araujo and McElwain [4, 3] with $\gamma_1 + \gamma_2 + \gamma_3 = 1$. Hence, isotropic growth corresponds to $\gamma_1 = \gamma_2 = \gamma_3 = \frac{1}{3}$. By allowing the tissue to grow anisotropically in response to the prevailing stress field, so that the expansion occurs preferentially in directions of least stress, the constitutive law is able to exhibit stress-relaxation in the absence of viscous dissipation. The mathematical theory of anisotropic growth has been developed by Araujo and McElwain [4, 3], while the phenomenon has been demonstrated experimentally by Helmlinger et al. [24]. Thus, in regularizing the elasticity by incorporating stress-relaxation into the growth component of the constitutive equation rather than in the stress-response component, anisotropic growth may be said to impart a *pseudo-viscoelasticity* to growing tissues. Appropriate functional forms for the anisotropic growth multipliers will be considered in the next paper in this series.

Now the constitutive equation for the solid phase becomes

$$\begin{aligned}\boldsymbol{\sigma}_s &= -\phi_s P\mathbf{I} + \lambda\left(tr\mathbf{E}_s - g\right)\mathbf{I} + 2\mu\left(\mathbf{E}_s - g\boldsymbol{\Omega}\right)\\ &= -\phi_s P\mathbf{I} + \lambda tr\mathbf{E}_s\mathbf{I} + 2\mu\mathbf{E}_s - g\left(3\lambda + 2\mu\right)\boldsymbol{\Omega}.\end{aligned} \tag{8.24}$$

Note that (8.24) was derived based on the two key assumptions of the intrinsic incompressibility of the phases and constant volume fractions, which together imply that $\lambda \to \infty$. While this may generally be introduced to the model by initially permitting compressibility and then allowing λ to tend to infinity in the *solution* of the boundary value problem, a more convenient approach in this case is to make the strain tensor the subject of the equation by noting that the trace of (8.24) is

$$tr\mathbf{E}_s = \frac{tr\boldsymbol{\sigma}_s + 3\phi_s P}{3\lambda + 2\mu} + g,$$

which gives

$$\mathbf{E}_s = \frac{1}{2\mu}\boldsymbol{\sigma}_s - \frac{\lambda}{2\mu(3\lambda + 2\mu)}(tr\boldsymbol{\sigma}_s + 3\phi_s P)\mathbf{I} + g\boldsymbol{\Omega}. \tag{8.25}$$

Now, in the limit as $\lambda \to \infty$, (8.25) becomes

$$\mathbf{E}_s = \frac{1}{2\mu}\boldsymbol{\sigma}_s - \left(\frac{tr\boldsymbol{\sigma}_s + 3\phi_s P}{6\mu}\right)\mathbf{I} + g\boldsymbol{\Omega}. \tag{8.26}$$

Note that while (8.26) represents a correct statement of the relationship between stress and strain when growth occurs, it must be able to reflect the fact that growth is a continuous process which creates movement. Thus, to accommodate the continuous expansion of the solid matrix due to the growth process, (8.25) must be differentiated with respect to time using an objective convected tensorial derivative such as the corotational (Jaumann) derivative (see, for example, [8] or [27]). Thus, (8.26) becomes

$$\frac{\mathcal{D}\mathbf{E}_s}{\mathcal{D}t} = \frac{1}{2\mu}\frac{\mathcal{D}\boldsymbol{\sigma}_s}{\mathcal{D}t} - \frac{1}{6\mu}\frac{\mathcal{D}}{\mathcal{D}t}\left(tr\boldsymbol{\sigma}_s + 3\phi_s P\right)\mathbf{I} + \frac{\mathcal{D}g}{\mathcal{D}t}\boldsymbol{\Omega},$$

where the notation $\frac{\mathcal{D}}{\mathcal{D}t}$ denotes an appropriate convected derivative. Taking the trace of this new equation now gives[1]

$$\frac{\mathcal{D}g}{\mathcal{D}t} = \boldsymbol{\nabla}\cdot\mathbf{v}_s + \frac{3\phi_s}{2\mu}\frac{\mathcal{D}P}{\mathcal{D}t},$$

which identifies the relationship between g and Γ_s via (8.14), being

$$\Gamma_s = \phi_s\rho\frac{\mathcal{D}g}{\mathcal{D}t} - \frac{3\phi_s^2\rho}{2\mu}\frac{\mathcal{D}P}{\mathcal{D}t}. \tag{8.27}$$

This enables the constitutive equation to be expressed in the form

$$\frac{\mathcal{D}\mathbf{E}_s}{\mathcal{D}t} = \frac{1}{2\mu}\frac{\mathcal{D}\boldsymbol{\sigma}_s}{\mathcal{D}t} - \frac{1}{6\mu}\frac{\mathcal{D}}{\mathcal{D}t}\left(tr\boldsymbol{\sigma}_s + 3\phi_s P\right)\mathbf{I} + \left(\boldsymbol{\nabla}\cdot\mathbf{v}_s + \frac{3\phi_s}{2\mu}\frac{\mathcal{D}P}{\mathcal{D}t}\right)\boldsymbol{\Omega}$$

[1]Strictly, the trace of the convected derivative of the infinitesimal strain tensor should be the divergence of the velocity vector with respect to the *reference* coordinates, i.e., $\mathbf{Div}(\mathbf{v_s})$ rather than $\boldsymbol{\nabla}\cdot\mathbf{v}_s$, based on the definition of the infinitesimal strain tensor given in (8.20). Nevertheless, if the strains required to ensure compatibility are small, as appropriate to the use of linearized constitutive equations, then $\mathbf{Div}(\mathbf{v_s})$ and $\boldsymbol{\nabla}\cdot\mathbf{v}_s$ may be used interchangeably here.

or

$$\frac{\mathcal{D}\mathbf{E}_s}{\mathcal{D}t} = \boldsymbol{\nabla}\cdot\mathbf{v}_s\boldsymbol{\Omega} + \frac{1}{2\mu}\frac{\mathcal{D}}{\mathcal{D}t}\left(\boldsymbol{\sigma}_s + \frac{1}{3}tr\boldsymbol{\sigma}_s\right) + \frac{\phi_s}{2\mu}\frac{\mathcal{D}P}{\mathcal{D}t}\left(3\boldsymbol{\Omega} - \mathbf{I}\right), \tag{8.28}$$

which reduces to

$$\frac{\mathcal{D}\mathbf{E}_s}{\mathcal{D}t} = \frac{1}{3}\boldsymbol{\nabla}\cdot\mathbf{v}_s\mathbf{I} + \frac{1}{2\mu}\frac{\mathcal{D}}{\mathcal{D}t}\left(\boldsymbol{\sigma}_s + \frac{1}{3}tr\boldsymbol{\sigma}_s\right) \tag{8.29}$$

in the special case of isotropic growth.

9. Summary of biphasic equations and comparison with single phase equations. Table 9.1 gives a summary of the suite of equations for the biphasic model of a growing tissue developed in this paper. Intriguingly, only the special case of isotropic growth gives rise to a solid phase constitutive equation identical to that used in single phase models, which consider growth as an analogy to thermal expansion. Nevertheless, it is essential to recognize that this combination of elasticity and isotropic growth does not incorporate the crucial aspect of stress-relaxation into the constitutive law. Indeed Lubkin and Jackson [29] explain that "the fatal mathematical combination of multiple phases, elasticity, and contractility renders the contractile-poroelastic model ill-posed.... The elasticity must then be regularized by a viscous term in order for solutions to exist." Araujo and McElwain [4] have shown that the elasticity may be regularized by considering *anisotropic* growth, thereby obviating the necessity to appeal to more complicated viscoelastic principles in many situations.

TABLE 9.1
Comparison of biphasic equations with single phase equivalents.

Equation type	Biphasic equations	Single phase equation
Balance of mass	$\phi_s\rho\boldsymbol{\nabla}\cdot\mathbf{v}_s = \Gamma_s$ $\rho\frac{D^s\phi_s}{Dt} = 0$ $\boldsymbol{\nabla}\cdot(\phi_s\mathbf{v}_s + \phi_f\mathbf{v}_f) = 0$	$\rho\boldsymbol{\nabla}\cdot\mathbf{v} = \Gamma$
Constitutive equations: Isotropic growth	$\frac{\mathcal{D}\mathbf{E}_s}{\mathcal{D}t} = \frac{1}{3}\boldsymbol{\nabla}\cdot\mathbf{v}_s\mathbf{I} + \frac{1}{2\mu}\frac{\mathcal{D}}{\mathcal{D}t}\left(\boldsymbol{\sigma}_s + \frac{1}{3}tr\boldsymbol{\sigma}_s\right)$ $\boldsymbol{\sigma}_f = -\phi_f P\mathbf{I}$	$\frac{\mathcal{D}\mathbf{E}_s}{\mathcal{D}t} = \frac{1}{3}\boldsymbol{\nabla}\cdot\mathbf{v}_s\mathbf{I} + \frac{1}{2\mu}\frac{\mathcal{D}}{\mathcal{D}t}\left(\boldsymbol{\sigma}_s + \frac{1}{3}tr\boldsymbol{\sigma}_s\right)$
Anisotropic growth	$\frac{\mathcal{D}\mathbf{E}_s}{\mathcal{D}t} = \boldsymbol{\nabla}\cdot\mathbf{v}_s\boldsymbol{\Omega} + \frac{1}{2\mu}\frac{\mathcal{D}}{\mathcal{D}t}\left(\boldsymbol{\sigma}_s + \frac{1}{3}tr\boldsymbol{\sigma}_s\right) + \frac{\phi_s}{2\mu}\frac{\mathcal{D}P}{\mathcal{D}t}\left(3\boldsymbol{\Omega} - \mathbf{I}\right)$ $\boldsymbol{\sigma}_f = -\phi_f P\mathbf{I}$	$\frac{\mathcal{D}\mathbf{E}_s}{\mathcal{D}t} = \boldsymbol{\nabla}\cdot\mathbf{v}_s\boldsymbol{\Omega} + \frac{1}{2\mu}\frac{\mathcal{D}}{\mathcal{D}t}\left(\boldsymbol{\sigma}_s + \frac{1}{3}tr\boldsymbol{\sigma}_s\right)$
Momentum equations[2]	$\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_s + \kappa(\mathbf{v}_f - \mathbf{v}_s) = 0$ $\phi_f\boldsymbol{\nabla}P = -\kappa(\mathbf{v}_f - \mathbf{v}_s)$	$\boldsymbol{\nabla}\cdot\boldsymbol{\sigma} = 0$

[2]Note that, in keeping with other published models of growing tissues [26, 36], inertial and body forces, as well as mass-exchange effects, are neglected in the momentum equations presented here.

10. Concluding remarks. In this paper, a theoretical framework for a solid-multiphase model of a growing tissue has been presented which extends the concepts of poroelasticity to accommodate continuous volumetric growth. Moreover, in incorporating a solid phase, the model provides a basis for the study of residual stresses, which is of fundamental importance in a wide range of studies in biology, physiology, and tissue engineering.

The general equations developed in sections 2 through 6 have been applied to a two-phase model of an elastic solid and an inviscid fluid in sections 7 through 9. The analysis points to a crucial phenomenological aspect of tissue growth, illustrating that such a process must consist of a coordinated combination of the "swelling" of the solid (cellular) phase due to the influx of extracellular fluid—which is, in essence, the inverse of the consolidation concept of poroelasticity—and the exchange of mass whereby extracellular fluid is incorporated into the cellular phase. This combination of processes necessitates the inclusion of an additional constitutive postulate—in which the mass-exchange term is related to the solid phase expansion—among the modeling equations to close the model.

In the present paper, a particular constitutive postulate has been chosen which reflects a tissue whose ratio of cells to extracellular fluid is constant throughout its volume. The assumption of linear-elasticity and *mechanical* isotropy (cf. isotropic *growth*) for the solid phase then enables simple constitutive equations between stress and strain to be specified for both the solid and fluid phases. Solutions to these biphasic equations will be presented in the next paper in this series.

This work may be extended in a number of ways. More complicated relationships between interphase mass exchange and solid phase expansion may be proposed, enabling the model to consider the formation of necrotic regions. Additionally, the equations could be rederived by incorporating a dependence of the Helmholtz free energy of the solid phase, Ψ_s, on both the solid deformation gradient, $\mathbf{F}_s$, and its convected derivative, $\overset{\nabla}{\mathbf{F}}_s$ (see sections 6 and 7 of the present paper) to produce a viscoelastic constitutive law (see, for example, Pioletti et al. [35]). This would enable the elasticity of the solid phase to be regularized in situations where anisotropic growth provides insufficient stress-relaxation [3].

Appendix. Development of the dissipation inequality. In this section, the second axiom of thermodynamics as expressed by (3.1) will be manipulated further to obtain a form from which constitutive equations may be deduced readily. Incorporating the balance of mass as expressed by (2.1) enables (3.1) to be expressed in the form

$$\sum_{i=1}^{n} \frac{1}{\theta}\left[\Gamma_i \eta_i \theta + \phi_i \rho_i \theta \frac{d^i \eta_i}{dt} + \theta \boldsymbol{\nabla}.\left(\frac{\mathbf{q}_i}{\theta}\right) - \phi_i \rho_i r_i\right] \geq 0. \tag{A.1}$$

Further, incorporating the energy equation (2.18) enables (A.1) to be expressed in a form which does not include the rate of heat production per unit mass within the ith constituent, r_i, explicitly, being

$$\sum_{i=1}^{n} \frac{1}{\theta}\left[\Gamma_i \eta_i \theta + \phi_i \rho_i \theta \frac{d^i \eta_i}{dt} + \theta \boldsymbol{\nabla}.\left(\frac{\mathbf{q}_i}{\theta}\right) - \phi_i \rho_i \frac{d^i u_i}{dt} + tr(\mathbf{L}_i \boldsymbol{\sigma}_i) - \boldsymbol{\nabla}.\mathbf{q}_i + \varepsilon_i\right] \geq 0.$$

Introducing the relation

$$\hat{\epsilon_i} \triangleq \varepsilon_i + \Gamma_i \left(u_i + \frac{1}{2}\mathbf{v}_i.\mathbf{v}_i\right) + \mathbf{v}_i.\pi_i$$

now enables the inequality to be expressed in the form

$$\sum_{i=1}^{n}\left[\phi_i\rho_i\left(\theta\frac{d^i\eta_i}{dt}-\frac{d^iu_i}{dt}\right)+tr(\mathbf{L}_i\boldsymbol{\sigma}_i)+\hat{\epsilon_i}-\Gamma_i\left(u_i+\frac{1}{2}\mathbf{v}_i.\mathbf{v}_i-\eta_i\theta\right)-\mathbf{v}_i.\pi_i\right]\geq 0.$$

The internal energies, u_i, will be eliminated at this point in favor of the Helmholtz free energy densities, ψ_i, where

$$\psi_i = u_i - \theta\eta_i. \tag{A.2}$$

Hence, the inequality becomes

$$-\sum_{i=1}^{n}\phi_i\rho_i\frac{d^i\psi_i}{dt}-\rho_m\eta_m\frac{d\theta}{dt}+tr\sum_{i=1}^{n}(\mathbf{L}_i\boldsymbol{\sigma}_i)-\sum_{i=1}^{n}\Gamma_i\left(\psi_i+\frac{1}{2}\mathbf{v}_i.\mathbf{v}_i\right)-\sum_{i=1}^{n}\mathbf{v}_i.\pi_i\geq 0.$$

At this stage, the chemical potential is introduced, being the linear transformation defined by

$$\mathbf{K}_i=\psi_i\mathbf{I}-\frac{\boldsymbol{\sigma}_i}{\phi_i\rho_i};$$

see Bowen and Wiese [14], noting that in the present paper it is assumed that $\boldsymbol{\sigma}_i=\boldsymbol{\sigma}_i^T$ (see section 2.3). Therefore,

$$\begin{aligned}-\sum_{i=1}^{n}\phi_i\rho_i\frac{d^i\psi_i}{dt}-\rho_m\eta_m\frac{d\theta}{dt}-tr\sum_{i=1}^{n}\phi_i\rho_i\mathbf{K}_i.\mathbf{L}_i+tr\sum_{i=1}^{n}\phi_i\rho_i\psi_i\mathbf{L}_i\\ -\sum_{i=1}^{n}\Gamma_i\psi_i-\sum_{i=1}^{n}\mathbf{v}_i.\left(\pi_i+\frac{\Gamma_i}{2}\mathbf{v}_i\right)\geq 0,\end{aligned}$$

which further reduces to

$$-tr\sum_{i=1}^{n}\phi_i\rho_i\mathbf{K}_i.\mathbf{L}_i-\rho_m\eta_m\frac{d\theta}{dt}-\sum_{i=1}^{n}\frac{d^i}{dt}(\phi_i\rho_i\psi_i)-\sum_{i=1}^{n}\mathbf{v}_i.\left(\pi_i+\frac{\Gamma_i}{2}\mathbf{v}_i\right)\geq 0 \tag{A.3}$$

by appealing to the balance of mass. Some authors (see, for example, Bowen [12]) define the quantity

$$\Psi_i=\phi_i\rho_i\psi_i$$

which represents the Helmholtz free energy of the ith constituent per unit mixture volume. Rewriting (A.3) in terms of Ψ_i gives

$$-tr\sum_{i=1}^{n}\phi_i\rho_i\mathbf{K}_i.\mathbf{L}_i-\rho_m\eta_m\frac{d\theta}{dt}-\sum_{i=1}^{n}\frac{d^i\Psi_i}{dt}-\sum_{i=1}^{n}\mathbf{v}_i.\left(\boldsymbol{\pi}_i+\frac{\Gamma_i}{2}\mathbf{v}_i\right)\geq 0. \tag{A.4}$$

Acknowledgments. The authors thank the referees for their helpful comments and constructive criticisms.

REFERENCES

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, 4th ed., Garland Science, New York, 2002.

[2] D. Ambrosi and F. Mollica, *On the mechanics of a growing tumor*, Internat. J. Engrg. Sci., 40 (2002), pp. 1297–1316.

[3] R. P. Araujo and D. L. S. McElwain, *The nature of the stresses induced during tissue growth*, Appl. Math. Lett., in press.

[4] R. P. Araujo and D. L. S. McElwain, *A linear-elastic model of anisotropic tumour growth*, Euro. J. Appl. Math., 15 (2004), pp. 365–384.

[5] R. P. Araujo and D. L. S. McElwain, *New insights into vascular collapse and growth dynamics in solid tumours*, J. Theoret. Biol., 228 (2004), pp. 335–346.

[6] A. Bertram and B. Svendsen, *On material objectivity and reduced constitutive equations*, Arch. Mech., 53 (2001), pp. 653–675.

[7] D. A. Beysens, G. Forgacs, and J. A. Glazier, *Embryonic tissues are viscoelastic materials*, Canad. J. Phys., 78 (2000), pp. 243–251.

[8] R. B. Bird, R. C. Armstrong, and O. Hassager, *Dynamics of Polymeric Liquids*, John Wiley & Sons, New York, 1977.

[9] Y. Boucher and R. K. Jain, *Microvascular pressure is the principal driving force for interstitial hypertension in solid tumours: Implications for vascular collapse*, Cancer Res., 52 (1992), pp. 5110–5114.

[10] R. M. Bowen, *Compressible porous media models by use of the theory of mixtures*, Internat. J. Engrg. Sci., 20, pp. 697–735.

[11] R. M. Bowen, *Toward a thermodynamics and mechanics of mixtures*, Arch. Ration. Mech. Anal., 24 (1967), pp. 370–403.

[12] R. M. Bowen, *Theory of mixtures*, in Continuum Physics, Vol. 3, A. C. Eringen, ed., Academic Press, New York, 1976.

[13] R. M. Bowen, *Incompressible porous media models by use of the theory of mixtures*, Internat. J. Engrg. Sci., 18 (1980), pp. 1129–1148.

[14] R. M. Bowen and J. C. Wiese, *Diffusion in mixtures of elastic materials*, Internat. J. Engrg. Sci., 7 (1969), pp. 689–722.

[15] H. M. Byrne, J. R. King, D. L. S. McElwain, and L. Preziosi, *A two-phase model of solid tumour growth*, Appl. Math. Lett., 16 (2003), pp. 567–573.

[16] C. Y. Chen, H. M. Byrne, and J. R. King, *The influence of growth-induced stress from the surrounding medium on the development of multicell spheroids*, J. Math. Biol., 43 (2001), pp. 191–220.

[17] B. D. Coleman and W. Noll, *Thermodynamics of elastic materials with conduction and viscosity*, Arch. Ration. Mech. Anal., 13 (1963), pp. 167–178.

[18] J. Dumais and C. R. Steele, *New evidence for the role of mechanical forces in the shoot apical meristem*, J. Plant Growth Regul., 19 (2000), pp. 7–18.

[19] W. Ehlers, *Constitutive equations for granular materials in geomechanical context*, in Environmental Sciences and Geophysics, K. Hutter, ed., CISM Courses and Lectures 337, Springer, Berlin, 1993.

[20] Y. C. Fung, *What are the residual stresses doing in our blood vessels?*, Ann. Biomed. Engrg., 19 (1991), pp. 237–249.

[21] Y. C. Fung, *Stress, strain, growth, and remodeling of living organisms*, Math. Phys., 46 (1995), pp. S469–S482.

[22] R. A. Gatenby, *Mathematical models of tumor-host interactions*, Cancer J., 11 (1998), pp. 289–293.

[23] R. A. Gatenby and P. K. Maini, *Cancer summed up*, Nature, 421 (2003), p. 321.

[24] G. Helmlinger, P. A. Netti, H. D. Lichtenbeld, R. J. Melder, and R. K. Jain, *Solid stress inhibits the growth of multicellular tumour spheroids*, Nature Biotech., 15 (1997), pp. 778–783.

[25] J. M. Huyghe and J. D. Janssen, *Quadriphasic mechanics of swelling incompressible porous media*, Internat. J. Engrg. Sci., 35, pp. 793–802.

[26] A. F. Jones, H. M. Byrne, J. S. Gibson, and J. W. Dold, *A mathematical model of the stress induced during avascular tumour growth*, J. Math. Biol., 40 (2000), pp. 473–499.

[27] D. D. Joseph, *Fluid Dynamics of Viscoelastic Liquids*, Springer-Verlag, New York, 1990.

[28] K. A. Landman and C. P. Please, *Tumour dynamics and necrosis: Surface tension and stability*, IMA J. Math. Appl. Medicine Biol., 18 (2001), pp. 131–158.

[29] S. R. Lubkin and T. Jackson, *Multiphase mechanics of capsule formation in tumors*, J. Biomech. Engrg., 124 (2002), pp. 237–243.

[30] L. E. MALVERN, *Introduction to the Mechanics of a Continuous Medium*, Prentice-Hall, Englewood Cliffs, NJ, 1969.
[31] V. MIRONOV, T. BOLAND, T. TRUSK, G. FORGACS, AND R. R. MARKWALD, *Organ printing: Computer-aided jet-based 3D tissue engineering*, Trends Biotech., 21 (2003), pp. 157–161.
[32] J. D. MURRAY, *Mathematical Biology,* I: *An Introduction*, Springer-Verlag, Berlin, 2002.
[33] S. L. PASSMAN AND J. W. NUNZIATO, *A theory of multiphase mixtures*, in Rational Thermodynamics, C. Truesdell, ed., Springer-Verlag, New York, 1984.
[34] H. M. PHILLIPS AND M. S. STEINBERG, *Embryonic tissues as elasticoviscous liquids*, J. Cell Sci., 30 (1978), pp. 1–20.
[35] D. P. PIOLETTI, L. R. RAKOTOMANANA, J. F. BENVENUTI, AND P. F. LEYVRAZ, *Viscoelastic constitutive law in large deformations: Application to human knee ligaments and tendons*, J. Biomech., 31 (1998), pp. 753–757.
[36] L. PREZIOSI AND A. FARINA, *On Darcy's law for growing porous media*, Internat. J. Nonlinear Mech., 37 (2002), pp. 485–491.
[37] K. R. RAJAGOPAL AND L. TAO, *Mechanics of Mixtures*, World Scientific, Singapore, 1995.
[38] M. A. SHANNON AND B. RUBINSKY, *The effect of tumour growth on the stress distribution in tissue*, Adv. Biol. Heat Mass Transfer, 231 (1992), pp. 35–38.
[39] R. SKALAK, *Growth as a finite displacement field*, in Proceedings of the IUTAM Symposium on Finite Elasticity, The Hague, 1981, D. E. Carlson and R. T. Shield, eds., Martinus Nijhoff, Dordrecht, Netherlands, pp. 347–355.
[40] R. SKALAK, S. ZARGARYAN, R. K. JAIN, P. A. NETTI, AND A. HOGER, *Compatiblity and the genesis of residual stress by volumetric growth*, J. Math. Biol., 34 (1996), pp. 889–914.
[41] H. SNIJDERS, J. HUYGHE, P. WILLEMS, M. DROST, J. JANSSEN, AND A. HUSON, *A mixture approach to the mechanics of the human intervertebral disc*, in Mechanics of Swelling, T. K. Karalis, ed., Springer-Verlag, Berlin, Heidelberg, 1992.
[42] C. TRUESDELL AND R. TOUPIN, *The classical field theories*, in Handbuch der Physik, Vol. III/I, S. Flugge, ed., Springer-Verlag, Berlin, 1960.

SIAM J. APPL. MATH.
Vol. 65, No. 4, pp. 1285–1304

MODELING, DESIGN, AND OPTIMIZATION OF A SOLID STATE ELECTRON SPIN QUBIT*

R. E. CAFLISCH†, MARK F. GYURE‡, HANS D. ROBINSON§, AND ELI YABLONOVITCH†

Abstract. This paper describes a solid state system in which a qubit is realized as the spin of a single trapped electron in a quantum dot and read functionality is via an adjacent quantum wire with a single or a small number of conductive states. Because of the limited design window for this system, simulation is an important guide to an experimental search for successful designs. We use a semianalytic approximation that is accurate enough to provide meaningful results and computationally simple enough to allow high throughput, as needed for design and optimization. In particular, we find designs that achieve double pinchoff (i.e., a single trapped electron in the dot and a single conductive state in the wire). After relaxing the design requirements to allow for a small number of conductive states in the wire, we find successful designs that are optimally robust, in the sense that their success is unlikely to be affected by fabrication errors.

Key words. qubit, quantum dot, quantum computing, design, optimization

AMS subject classifications. 81P68, 35P20

DOI. 10.1137/040606181

1. Introduction.

Quantum logic, based on manipulation and interaction of binary quantum states or "qubits," has great potential for communication and computation. Quantum communication could offer absolute security [14] and transmission rates beyond the Shannon limit [5]. Quantum computation could greatly accelerate the solution of certain important problems, such as prime factorization [15], database searching [6], and simulation of quantum systems [16]. This potential has motivated a large effort to develop and implement quantum logic. Currently, the foremost problem for quantum communication and computation is the implementation of qubits in a robust and scalable system, which will allow for error correction and control of decoherence. Solid state implementations of a qubit, based on an electron or nuclear spin confined to a quantum dot, have been proposed in [1, 8, 9, 10, 11, 12, 17, 18].

This paper is concerned with the design of a single qubit system in a solid state implementation, as proposed in [3, 17], in which a qubit is represented as the spin of a single electron confined in a quantum dot. A quantum wire is placed below the quantum dot, so that the conductivity of the wire will depend sensitively on the charge present in the dot. The wire can then be used to verify the presence of a single electron in the quantum dot and, in the presence of a spin polarized electron reservoir, to read out the spin of that electron. The latter can be accomplished in several ways, for instance, by measuring tunneling times from the reservoir into the qubit. As the spin singlet state has lower energy than the triplet states [4], we can arrange to make tunneling into a triplet state energetically forbidden, and since the singlet state must

*Received by the editors April 2, 2004; accepted for publication (in revised form) October 6, 2004; published electronically April 14, 2005. This research was supported in part by a DARPA grant as part of the Quantum Information Science and Technology (QuIST) Initiative and by the Center for Scalable and Integrated Nanomanufacturing, NSF grant DMI-0327077.

http://www.siam.org/journals/siap/65-4/60618.html

†Mathematics Department, UCLA, Los Angeles, CA 90095-1555 (caflisch@math.ucla.edu).

‡HRL Laboratories, 3011 Malibu Canyon Road, Malibu, CA 90265 (gyure@hrl.com).

§Electrical Engineering Department, UCLA, Los Angeles, CA 90095 (hansr@ee.ucla.edu, eliy@ee.ucla.edu).

be formed using two opposite spins, the tunneling time will then depend strongly on the relative alignment of the qubit spin to the reservoir spin.

In our design, the quantum dot and wire are formed in two vertically stratified, parallel semiconductor quantum wells and are defined electrostatically using lithographically patterned gates on the surface of the semiconductor. Further description of the geometry and electrostatics of this system are provided in section 2. An alternative design using a horizontal placement of quantum dots has been carried out in [18].

A successful qubit design requires a single electron in the quantum dot and a small number of conduction states in the quantum wire. If there is only a single state in the quantum wire, it can be used both as spin reservoir and charge sensor. Moreover, the design should be robust with respect to fluctuations or errors in modeling and fabrication. These are very stringent requirements that are difficult to satisfy, and numerical simulation can serve as an important guide in the experimental search for successful designs.

The principal goal of the present study is development of a semianalytic model and its application to design and optimization for this quantum system. This reduced order model is based on a number of approximations that restrict its validity. Comparison to full scale numerical simulations, however, indicate that its accuracy is sufficient to provide meaningful results. Computational speed is the model's great virtue, enabling the high throughput that is required for design and optimization of the quantum system.

There are three distinct aspects to simulation of this quantum system: construction of a mathematical model embodying the correct physics, development of an effective numerical method for solving the model, and use of the numerical method to search for a successful design. The semianalytic model and its use for design and optimization, as presented in this paper, address only the last of these. Related efforts, which are beyond the scope of this paper, include a full-scale numerical method for the Schrödinger–Poisson model [2] and simulations using nextnano^3 [13] that include more detailed physics.

Furthermore, design of this qubit system is an intermediate, but important step toward the much more challenging goal of constructing a quantum device. A functioning quantum device using this qubit system must satisfy additional requirements, such as preparation of initial data, coupling of qubits, measurement of the qubit state, control of decoherence, and error correction, that are not included in the present design problem.

In section 2 we develop a Schrödinger–Poisson model for simulation of the electrostatic potential and the single electron wavefunction, and we formulate design goals for performance of this system. A reduced order, semianalytic model is derived in section 3 using square well or parabolic approximations for the electrostatic potential. In section 4 the accuracy and validity of this semianalytic model is assessed by comparison to full scale numerical solution of the Schrödinger–Poisson model from [2]. Successful designs with double pinchoff are found in section 5 through a random search in parameter space. A measure of design robustness, in terms of the sensitivity of the design to fabrication errors, is formulated in section 6. In section 7, an analysis is presented that greatly simplifies the computation of design robustness. Using this simplified analysis, a search for designs that are optimally robust is described in section 8. Finally, conclusions are presented in section 9.

2. Qubit design problem. This sections describes the solid state system and the design goals for a qubit. This description includes one-, two-, and three-dimensional versions of the system.

2.1. Qubit system description. The layered semiconductor system consists of a series of material layers, with layer i consisting of material m_i and having thickness dz_i, in which z is a measure of the distance from the top planar surface, increasing in the downward direction. In addition there are δ-doped layers of zero thickness at the boundaries of some of the material layers, with a density σ_k of ions per area in the kth δ-doped layer. Note that the charge density σ is the doping density times an activation factor, so that it is less than the actual doping density. Volumetric doping, including intrinsic doping, is neglected.

As an example that will be used in this study and is pictured in Figure 1, consider a system that consists of the following layers, in order starting at $z = 0$:

- layer of material A of thickness dz_1;
- δ-doped layer with charge density σ_1;
- layer of material A of thickness dz_2;
- layer of material B of thickness dz_3, the layer containing the quantum dot;
- layer of material A of thickness dz_4;
- layer of material B of thickness dz_5, the layer containing the quantum wire;
- layer of material A of thickness dz_6;
- δ-doped layer with charge density σ_2;
- layer of material A of infinite thickness.

The geometry of these layers is one-dimensional; higher dimensionality is determined by the geometry of the gates. At the top of the material system, i.e., $z = 0$, there are a series of gates G_m on which the electron potential energy ϕ_m is specified. Away from the gates, the energy is taken to be equal to a constant free surface potential ϕ_0.

In this study the following gate geometries and potentials are considered:

- gate G_g consisting of a circle $r < R_g/2$ for the three-dimensional geometry or an interval $|x| < R_g/2$ for the two-dimensional geometry, on which the potential energy is ϕ_g;
- two gates $G_{b\pm}$, in which G_{b+} consists of points with $x > R_b/2$ and G_{b-} consists of points with $x < -R_b/2$ in both the two-dimensional and three-dimensional geometries, with potential ϕ_b on both gates;
- no gates for the one-dimensional geometry;
- potential energies $\phi_g = -V_g + \phi_{schottky}$ and $\phi_b = -V_b + \phi_{schottky}$, where $\phi_{schottky}$ is the Schottky barrier, and V_g and V_b are the voltages applied to the gates.

In this description all distances are measured in nm, the doping densities are measured in units of electrons cm^{-2}, and the energy ϕ is in units of eV.

A drawing of the device structure, with parameters from an optimally robust design as in (8.1), is shown in Figure 1. A schematic drawing of the gates, the potentials in the quantum wells, and the electron densities is shown in Figure 2. Positive potential energy (ϕ_b) on the planar side gates raises the potential on the sides of both quantum wells. Negative potential energy (ϕ_g) on the circular, central gates lowers the potential energy, primarily in the upper well. This leads to a localized electron density (i.e., a quantum dot) in the upper well and an electron density along a line (i.e., a quantum wire) in the lower quantum well.

The layer widths dz_i and the charge densities σ_i are determined during the material growth, and the gate sizes R_i are determined during the device fabrication. These parameters cannot be changed after fabrication. Thus the parameters can be divided into two sets: the vector of operation parameters $v_o = (\phi_g, \phi_b)$, which can be varied during operation of the device, and the vector of design parameters

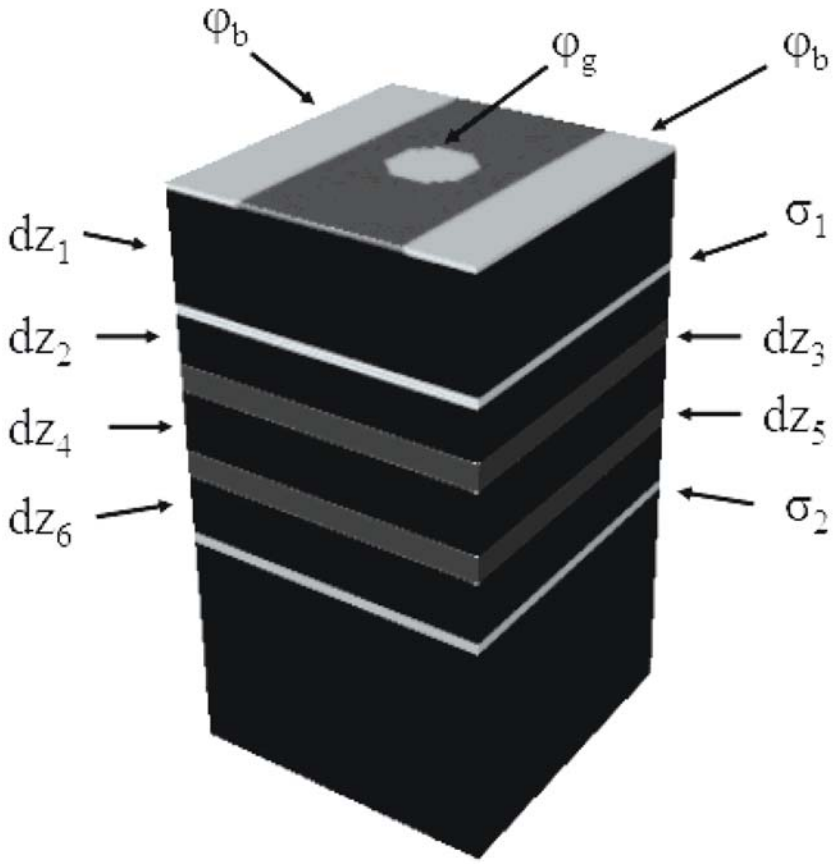


FIG. 1. *Schematic drawing of the device geometry. The depths of the material layers are dz_i, in which dz_3 is the depth of the upper quantum well, which contains the quantum dot, and dz_5 is the depth of the lower quantum well, which contains the quantum wire. The charge densities in the delta-doped layers are σ_i. The electrostatic potential of the central gate (a circle of diameter R_g in a three-dimensional geometry) is ϕ_g, and the electrostatic potential of the two side gates (separated by a distance R_b) is ϕ_b. In a two-dimensional geometry the central gate would be an infinite strip parallel to the side gates.*

$v_d = (dz_1, dz_2, dz_3, dz_4, dz_5, dz_6, R_g, R_b, \sigma_1, \sigma_2)$, which cannot be changed during operation. A device design can be identified with a choice of the design vector v_d. These are chosen from a subset C of R^{10} that has been determined from some external consideration, such as additional constraints or previous experience. These have the form

$$\underline{dz}_i < dz_i < \overline{dz}_i \quad \text{for} \quad 1 \le i \le 6, \tag{2.1}$$

$$\underline{R}_g < R_g < \overline{R}_g, \tag{2.2}$$

$$\underline{R}_b < R_b < \overline{R}_b, \tag{2.3}$$

$$\underline{\sigma}_i < \sigma_i < \overline{\sigma}_i \quad \text{for} \quad 1 \le i \le 2, \tag{2.4}$$

$$\underline{\phi}_g < -\phi_g < \overline{\phi}_g, \tag{2.5}$$

$$\underline{\phi}_b < \phi_b < \overline{\phi}_b. \tag{2.6}$$

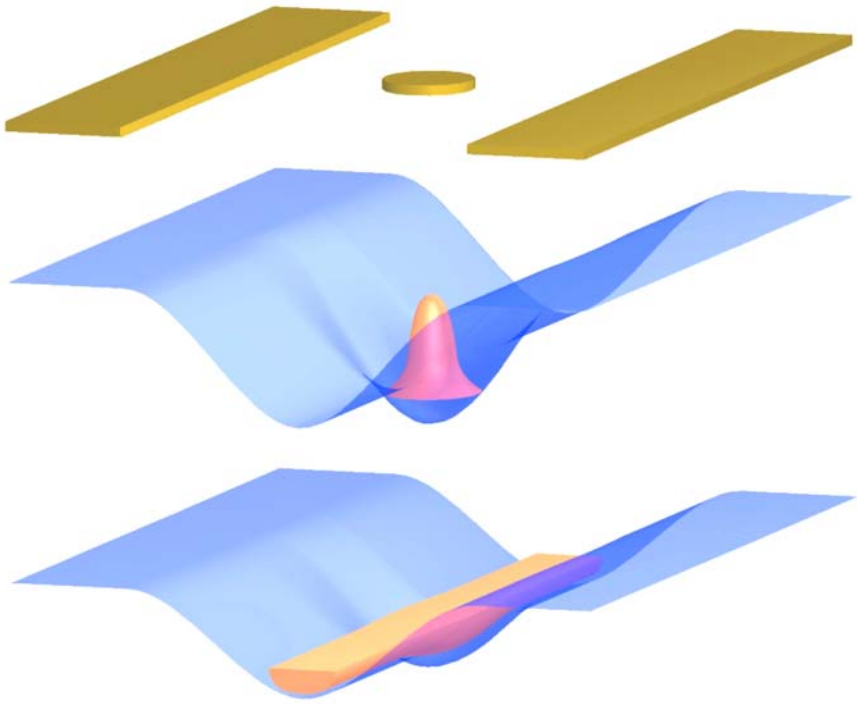


FIG. 2. *Schematic drawing of the gates, the potential energies in the quantum wells, and the electron densities in the dot and the wire.*

In the design searches conducted below, a typical set of constraints was the following:

$$
\begin{aligned}
20 &< dz_1 < 40(nm),\\
20 &< dz_2 < 40(nm),\\
5 &< dz_3 < 10(nm),\\
dz_4 &= 16(nm),\\
5 &< dz_5 < 15(nm),\\
dz_6 &= 50(nm),\\
50 &< R_g < 100(nm),\\
50 &< (R_b - R_g)/2 < 400(nm),\\
0 &< \sigma_1 < 4 \times 10^{11}(cm^{-2}),\\
0 &< \sigma_2 < 4 \times 10^{11}(cm^{-2}),\\
0.1 &< -\phi_g < 0.3(eV),\\
0 &< \phi_b < 2.0(eV).
\end{aligned}
\tag{2.7}
$$

The constraints on dz_4 were chosen to allow electrons to tunnel between dot and wire on a ms time scale; 16 nm is appropriate for the InP/InGaAs system. The lower bounds on R_g and R_b are representative of what can be easily accomplished with e-beam lithography. The upper bounds on σ_i are set to avoid hopping conduction through the doping layers. The upper bound on ϕ_b and the lower bound on ϕ_g are set

to avoid passing any current through the gates into the sample, which would disrupt the qubit. The lower bound on ϕ_g is set to ensure formation of a quantum dot with a bound electron state; at smaller voltages, the electron may not be bound, which the central gate, the parabolic approximation in section 3 may fail to predict. The remaining constraints are reasonable limits that were imposed to speed the design search.

2.2. Schrödinger–Poisson model. The electrostatic potential Φ is assumed to satisfy the Poisson equation

$$\nabla \cdot \epsilon \nabla \Phi = \sigma_1 \delta_1 + \sigma_2 \delta_2 - \rho_\psi \tag{2.8}$$

in which $\epsilon = k\epsilon_0/e^2$ is the scaled dielectric constant, δ_i is a δ-function on ith δ-doped interface, and ρ_ψ is the number density for electrons in the wire, which must be determined self-consistently as described below. The function Φ is the potential energy for an electron, measured in eV. For this equation the boundary conditions are taken to be Dirichlet conditions (i.e., Φ prescribed) at the top, Neumann conditions (i.e., $\partial\Phi/\partial n = 0$) at the bottom, and periodic conditions on the sides.

The electronic wave function for the unbound electron of lowest energy is assumed to satisfy a single particle Schrödinger equation

$$-\nabla \cdot \left(\frac{\hbar^2}{2m} \nabla \Psi \right) = -(\Phi + U)\Psi + \lambda \Psi \tag{2.9}$$

in which e is the electron charge, m is the electron effective mass, U is the conduction band offset relative to material A (InP in the examples below), λ is the energy level (eigenvalue) in units of eV, and Ψ is wave function (eigenfunction).

As an example, for InP, $In_xGa_{1-x}As$, and $Al_yIn_{1-y}As$, with alloy fractions $x = 0.53$ and $y = 0.48$, the material parameters are given in Table 1 and the relevant physical constants are given in Table 2.

TABLE 1
Material parameters.

Parameter	InP	$In_{0.53}Ga_{0.48}As$	$Al_{0.48}In_{0.52}As$	Units
k	12.61	13.9	12.7	1
$\epsilon = k\epsilon_0/e^2$	.697	.769	0.702	1/(eV nm)
m	.079	.041	0.0733	m_0
$\hbar^2/2m$	.484	.94	0.522	eV/nm^2
U	0	.224	0.25	eV

TABLE 2
Physical constants.

Constant	Value	Units
$\hbar^2/2m_0$	.0382	eV/nm^2
ϵ_0	8.854×10^{-12}	$C^2N^{-1}m^{-2}$
ϵ_0/e^2	.0553	1/(eV nm)

For the solutions of interest in this study, the electrons are localized either in a dot in the upper layer or along a wire in the lower layer. This implies that the eigenfunctions for the Schrödinger equation (2.9) are each localized in either the dot or the wire. The eigenvalues and eigenfunctions in the quantum dot are labeled

λ_k^d, Ψ_k^d; those in the wire are labeled λ_k^w, Ψ_k^w. Also denote $d\lambda = \lambda_2 - \lambda_1$ as the difference between the first two eigenvalues. The eigenfunctions are normalized so that $\int |\Psi|^2 dx = 1$. The self-consistent charge density is $\rho_\psi = \sum |\Psi|^2$, summed over all $\lambda < E_F$, in which the Fermi energy E_F is set to 0.

To emphasize the dependence of the eigenvalues λ on the gate voltages ϕ_g and ϕ_b and the design vector v, we shall sometimes write $\lambda = \lambda(\phi_g, \phi_b, v)$.

Note that the eigenfunctions for the quantum dot are quite distinct from those for the quantum wire. So computation of these eigenfunctions is equivalent to a computation using two separate wave functions for the dot and wire, and so it correctly represents the charge density in the wire and its affect on the dot. Interaction terms between the dot and wire are omitted, because they are small. On the other hand, tunneling effects between the dot and wire are important for detection of an electron in the dot using the wire. These tunneling effects are beyond the scope of the current model.

2.3. Design goals. The design goals for the quantum dot are to have a single confined electron under the gate and no confined electron states away from the gate. The design goals for the quantum wire are to have a small number k of conduction states in the wire, with no additional states in the wire under the gate. Denote E_k^d and E_k^w to be the energy for k electrons in the dot and for k conduction states in the wire, respectively. Also denote $E_k^d(0)$ and $E_k^w(0)$ to be the same energies but with no voltage on the central gate; i.e., $\phi_g = 0$. An unbound electron will be localized if its energy is less than the Fermi energy $E_F = 0$. The design goals can thus be stated as

$$E_1^d < 0 < E_2^d, \tag{2.10}$$

$$0 < E_1^d(0), \tag{2.11}$$

$$E_k^w < 0 < E_{k+1}^w, \tag{2.12}$$

$$E_k^w(0) < 0 < E_{k+1}^w(0). \tag{2.13}$$

Since $E_k^w < E_k^w(0)$, (2.12), and (2.13) can be recombined as

$$E_k^w(0) < 0 < E_{k+1}^w. \tag{2.14}$$

As shown below, the energy level E_2^d is smaller than $E_1^d(0)$ in the regime of interest, so that (2.11) is redundant.

In the quantum dot, the energy for a single electron is the lowest eigenvalue, so that $E_1^d = \lambda_1^d$ and $E_1^d(0) = \lambda_1^d(0)$. For the energy of two electrons, there is an interaction (Coulomb) correction $E_2^d = \lambda_2^d + \tilde{E}_2^d$. In the wire, we identify a conduction state, as an eigenfunction for the cross-section of the wire and neglect the interaction among different conduction states. Thus $E_k^w = \lambda_k^{w2D}$ and $E_k^w(0) = \lambda_k^{w2D}(0)$, in which λ_k^{w2D} and $\lambda_k^{w2D}(0)$ are the two-dimensional eigenvalues and $\phi_g = 0$ for $\lambda_k^{w2D}(0)$. Therefore the operation goals can be rewritten as

$$\lambda_1^d < 0 < \lambda_2^d + \tilde{E}_2^d, \tag{2.15}$$

$$\lambda_1^{w2D}(0) < 0 < \lambda_2^{w2D}. \tag{2.16}$$

The design goal is to find a device design v_d, for which there is a choice of operation parameters v_o such that the operation goals (2.15) and (2.16) are satisfied. A second design goal, that the operation goals are still met in the presence of growth and fabrication uncertainties, is formulated in section 6.

3. Semianalytic model. In this section we formulate a simplified semianalytic model that represents an approximate solution of the Schrödinger–Poisson equation. As described below, the potential Φ in each of the upper and lower quantum wells is approximated as a parabola in the lateral directions x and y and a square well in the depth direction z. Because the layered geometry is independent of x and y and the gates have a reflection symmetry with respect to both x and y, the first derivatives Φ_x and Φ_y are 0 on the centerline $x = y = 0$. Thus the lateral variation of the potential near the either quantum dot and quantum wire is approximately given by $\frac{1}{2}(x^2\Phi_{xx} + y^2\Phi_{yy})$.

3.1. Approximations for electrostatics. For the potential $\Phi = \Phi^{1D}$ due to modulationally doped layers but not including the effect of the gates, put the bottom boundary condition at ∞, omit any self-consistent terms, and neglect the variation in dielectric constant by using the value for material A throughout to obtain

$$\Phi^{1D} = \begin{cases} \phi_{top} - \epsilon_A^{-1}(\sigma_1 + \sigma_2)z, & 0 < z < z_1, \\ \phi_{top} - \epsilon_A^{-1}(\sigma_1 z - \sigma_2 z_1), & z_1 < z < z_2, \\ \phi_{top} - \epsilon_A^{-1}(\sigma_1 z_2 - \sigma_2 z_1), & z_2 < z, \end{cases} \tag{3.1}$$

in which $z_1 = dz_1$ and $z_2 = dz_1 + dz_2 + dz_3 + dz_4 + dz_5 + dz_6$ are the positions of the δ-doped layers.

The potential $\Phi = \Phi_L^{2D}$, due to a gate that is a strip (in three dimensions) (i.e., $|x| < L/2, z = 0$) with potential $\Phi = 1$ on the gate and $\Phi = 0$ away from gate, is

$$\Phi_L^{2D}(x,z) = \pi^{-1}\left(\arctan\left(\frac{x + L/2}{z}\right) - \arctan\left(\frac{x - L/2}{z}\right)\right). \tag{3.2}$$

On the central axis $x = 0$, the values of Φ and its second derivative are

$$\Phi_L^{2D}(x = 0, z) = 2\pi^{-1}\arctan(L/2z), \tag{3.3}$$

$$\Phi_{Lxx}^{2D}(x = 0, z) = -\pi^{-1}z^{-2}\frac{2L/z}{(1 + L^2/4z^2)^2}. \tag{3.4}$$

The potential $\Phi = \Phi_d^{3D}$, due to a gate that is a circle (i.e., $r = |(x,y)| < d/2$) with potential $\Phi = 1$ on the gate and $\Phi = 0$ away from gate, is

$$\begin{aligned} \Phi_d^{3D}(\mathbf{x}) &= \Phi(r,z) \\ &= \frac{|z|}{2\pi}\int_0^{2\pi}\int_0^{d/2} |\mathbf{x} - \mathbf{x}'|^{-3} r'\,dr'\,d\theta' \\ &= \frac{|z|}{2\pi}\int_0^{2\pi}\int_0^{d/2} (z^2 + (r - r'\cos\theta')^2 + r'^2\sin^2\theta')^{-3/2} r'\,dr'\,d\theta'. \end{aligned} \tag{3.5}$$

On the central axis $r = 0$, the values of Φ and its second derivative are

$$\Phi_d^{3D}(r = 0, z) = 1 - (1 + (d/2z)^2)^{-1/2}, \tag{3.6}$$

$$\Phi_{drr}^{3D}(r = 0, z) = -\frac{3}{2}|z|(d/2)^2(z^2 + (d/2)^2)^{-5/2}. \tag{3.7}$$

Add these together to obtain the total potential as

$$\Phi = \begin{cases} \Phi^{1D} + \phi_b(1 - \Phi_d^{2D}) + \phi_g\Phi_L^{2D} & \text{in two dimensions,} \\ \Phi^{1D} + \phi_b(1 - \Phi_d^{2D}) + \phi_g\Phi_L^{3D} & \text{in three dimensions.} \end{cases} \tag{3.8}$$

The second derivatives of the total potential on the central axis are

$$\Phi_{xx} = \begin{cases} -\phi_b \Phi^{2D}_{dxx} + \phi_g \Phi^{2D}_{Lxx} & \text{in two dimensions,} \\ -\phi_b \Phi^{2D}_{dxx} + \phi_g \Phi^{3D}_{Lrr} & \text{in three dimensions,} \end{cases} \tag{3.9}$$

$$\Phi_{yy} = \begin{cases} 0 & \text{in two dimensions,} \\ \phi_g \Phi^{3D}_{Lrr} & \text{in three dimensions.} \end{cases} \tag{3.10}$$

All the subsequent computations for the semianalytic model were performed using MATLAB programs.

3.2. Approximations for Schrödinger. The approximation for the Schrödinger eigenfunctions and eigenvalues relies on separation of variables: if m is constant and $\Phi(x, y, z) = \Phi^x(x) + \Phi^y(y) + \Phi^z(z)$, then

$$\lambda = \lambda^x + \lambda^y + \lambda^z, \tag{3.11}$$

$$\Psi(x, y, z) = \Psi^x(x)\Psi^y(y)\Psi^z(z) \tag{3.12}$$

in which

$$-(\hbar^2/2m)\Psi^x_{xx} = -\Phi^x\Psi^x + \lambda^x\Psi^x, \tag{3.13}$$

$$-(\hbar^2/2m)\Psi^y_{yy} = -\Phi^y\Psi^y + \lambda^y\Psi^y, \tag{3.14}$$

$$-(\hbar^2/2m)\Psi^z_{zz} = -\Phi^z\Psi^z + \lambda^z\Psi^z. \tag{3.15}$$

Use separation of variables to find eigenvalues in a channel of width w and center z. Neglect variation of Φ across the well and approximate the x-dependence for Φ^{2D} (i.e., for a gate that is a strip in three dimensions) or the (x, y)-dependence for Φ^{2D} (i.e., for a gate that is a circle in three dimensions) as parabolic with

$$\Phi^{2D} \approx \Phi^x_{2D}(x) \qquad = .5\Phi_{xx}(x = 0, z)\ x^2, \tag{3.16}$$

$$\Phi^{3D} \approx \Phi^x_{3D}(x) + \Phi^y_{3D}(y) = .5\Phi_{rr}(r = 0, z)(x^2 + y^2). \tag{3.17}$$

Both the two-dimensional and three-dimensional problems have been written as a sum of one-dimensional parabolic potentials. The eigenvalue and eigenvalue spacing for a one-dimensional parabolic potential $\Phi(x) = \phi_b x^2$ are

$$\lambda^p_1 = (\Phi_{xx}\hbar^2/4m)^{1/2}, \tag{3.18}$$

$$d\lambda^p = 2\lambda^p_1. \tag{3.19}$$

We denote λ^{px} and λ^{py} for the eigenvalues due to the parabolic potential in the x- and y-directions, respectively, and $d\lambda^{px}$ and $d\lambda^{py}$ for the corresponding eigenvalue spacing.

In the z-direction, the potential is approximately a square well, since

$$\Phi^z(z) = \begin{cases} 0 & |z - z_0| > L/2, \\ -U & |z - z_0| < L/2 \end{cases} \tag{3.20}$$

in which U is the offset in wells, neglecting variation across the well. The eigenvalues for this square well are solutions of

$$\lambda^{sw} = c_1 k_1^2 - U, \tag{3.21}$$

$$k_1^2(1 + (c_1/c_0)\tan^2(k_1 w/2)) = U/c_1 \tag{3.22}$$

in which c_0 and c_1 are the values of $\hbar^2/2m$ outside the well and in the well, respectively; i.e., c_0 is the value for material A (InP) and c_1 is the value for material B (InGaAs).

In summary, the eigenfunctions, lowest eigenvalue and eigenvalue spacing are

$$\Psi = \Psi^{sw}(z)\Psi^{px}(x)\Psi^{py}(y), \tag{3.23}$$

$$\lambda = \lambda^{sw} + \lambda^{px} + \lambda^{py}, \tag{3.24}$$

$$d\lambda = \begin{cases} d\lambda^{px} & \text{in two dimensions,} \\ \min(d\lambda^{px}, d\lambda^{py}) & \text{in three dimensions.} \end{cases} \tag{3.25}$$

In the simplest model, we also take the energy for two electrons to be the same as the second eigenvalue in the quantum dot; i.e., set $\tilde{E}_2^d = 0$ in (2.15).

3.3. Generalizations. Two generalizations of the semianalytic model of the previous section are formulated here to include effects of Coulomb interactions and self-consistent terms.

An approximation to the Coulomb correction $\tilde{E}_2^d$ for two electrons in the quantum dot has been developed by Gyure [7]. He computed the energy for two electrons in a one-dimensional parabolic potential using an iterative projection method, then fit the result to the following simple formula:

$$E_2^d = \lambda_1^d + c r_d \gamma_y^\kappa \tag{3.26}$$

in which $r_d = 0.00289\ eV$ is the Rydberg energy, $\gamma_y = \lambda_1^d$, and the (dimensionless) fitting parameters are $c = 3.5213$ and $\kappa = 0.75654$. The one-dimensional approximation was justified by two-dimensional calculations that showed the anisotropy of the potential is large enough in most cases to ignore the smaller dimension. The error induced is relatively small and decreases rapidly with anisotropy ratio.

The most significant self-consistent terms are the effect of the charge in the wire on the potential in the dot. For a wire defined by a parabolic potential of width a_x and a square well of depth a_z, approximate the charge in the wire as being uniformly distributed over an ellipse with a_x and a_z as the principal axes. Define elliptic coordinates (u, v) in the (x, z) plane as

$$x = b\cosh(u)\cos(v), \tag{3.27}$$

$$z = b\sinh(u)\sin(v) \tag{3.28}$$

in which $b = \sqrt{a_x^2 - a_z^2}$, so that the ellipse corresponds to $u = u_e = \cosh^{-1}(a_x/b)$. As an approximation to the potential for an elliptical charge, use

$$\tilde{\Phi} = \begin{cases} \alpha u - \gamma & \text{for } u > u_e, \\ \beta(a_x x^2 + a_z z^2) - \kappa & \text{for } u < u_e. \end{cases} \tag{3.29}$$

In the limit $u \to \infty$, $u \approx \log(r/c)$, which implies that $\alpha = \bar{\rho}/2\pi$ in which $\bar{\rho}$ is the total charge on the ellipse. At the top of the layered material, the correction $\tilde{\Phi}$ should vanish. Apply this at the value $u_0 = u(z = 0, x = 0)$ to get $\gamma = u_0\bar{\rho}/2\pi$. The total charge in the wire $\bar{\rho}$ is approximately given by

$$\bar{\rho} = \pi^{-1}\sqrt{d\lambda^w/c}((2/3)N^{3/2} + N) \tag{3.30}$$

in which $c = \frac{\hbar^2}{2m}$ and $N = -\lambda_1^w/d\lambda^w$ is the number of transverse eigenvalues that are less than the Fermi energy. Formula (3.30) comes from the number of longitudinal states below the Fermi energy for each transverse state. The potential corrections in (3.29) are used to correct the eigenvalue λ_1^w and eigenvalue spacing $d\lambda^w$, which are then used in (3.30). These two equations are solved iteratively to determine $\bar{\rho}$.

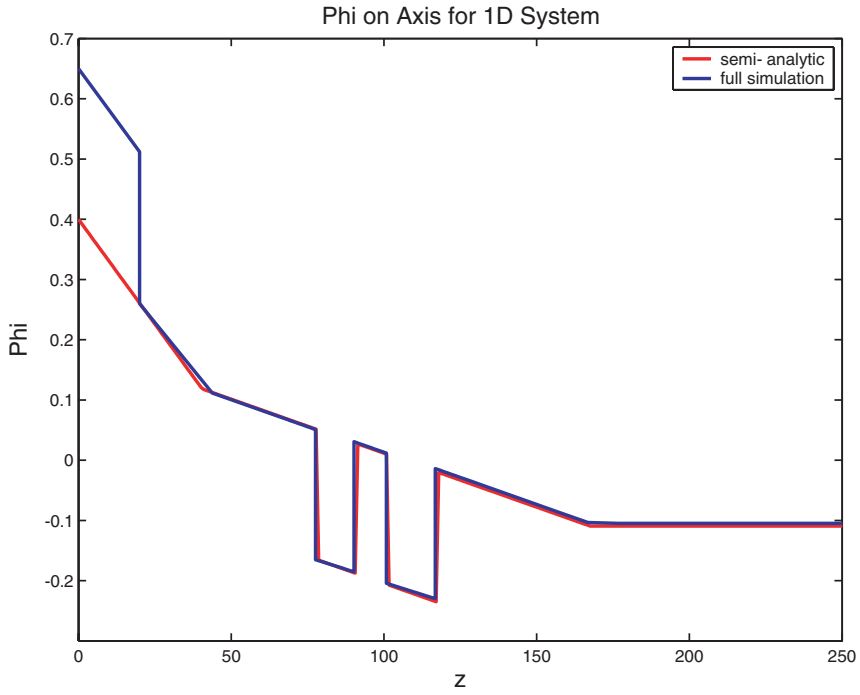


FIG. 3. *Comparison of potential* Φ *for a one-dimensional geometry from the full scale simulation method (blue) and the semianalytic model (red), with no top gates, no Coulomb interactions, self-consistent effects omitted, and no background doping, showing excellent agreement. Disagreement at the top is due to use of a different material layer in the full-scale simulation that does not influence the potential outside that layer and was not included in the semianalytic model.*

4. Validation of semianalytic model. Validation of the semianalytic model is performed by comparison to a full-scale numerical solution of the Schrödinger–Poisson equations by Anderson [2]. Consider a system with design parameters

$$\begin{aligned} v_d &= (dz_1, dz_2, dz_3, dz_4, dz_5, dz_6, R_g, R_b, \sigma_1, \sigma_2) \\ &= (40.5, 37.1, 12.6, 10.6, 16, 50.7, 61, 219.5, 3.6 \times 10^{11}, 1.25 \times 10^{11}) \end{aligned} \tag{4.1}$$

and with operation parameters $v_o = (\phi_g, \phi_b) = (0, 0.53)$. Figures 3, 4, and 5 show the potential on the central line (through the center of the quantum dot) in one, two, and three dimensions, respectively, with no Coulomb interactions, self-consistent effects omitted, and no background doping. In one dimension there are no gates on the top of the system, so that the potential Φ is a function of z only. In two dimensions the central gate is an interval (i.e., a strip in three dimensions); while in three dimensions the central gate is a circular dot. Figure 6 is the same as Figure 4, except that self-consistent effects are included in both the full-scale numerical computation and the semianalytic model. The first eigenvalue is shown for each of these problems in Table 3, in meV.

These results show excellent agreement for the case with no Coulomb interactions and no self-consistent effects. In this case the energy errors in the semianalytic method are all within 6 meV of those for the full simulation. With self-consistent effects, the agreement is still good, with energy errors of size 20 meV.

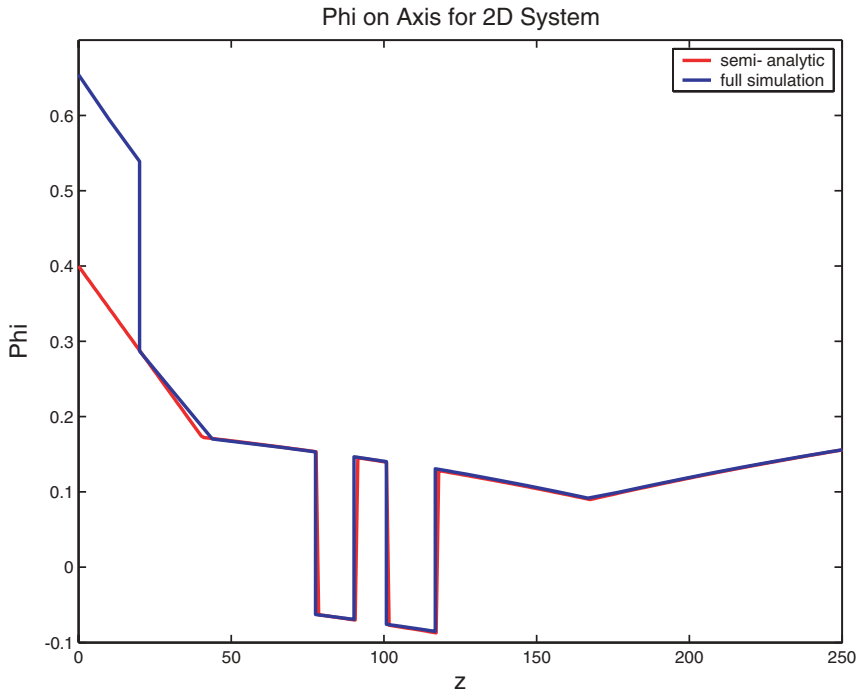


FIG. 4. *Same as Figure* 3 *except that the plot is for the potential* Φ *on the central axis* $x = 0$ *for a two-dimensional geometry.*

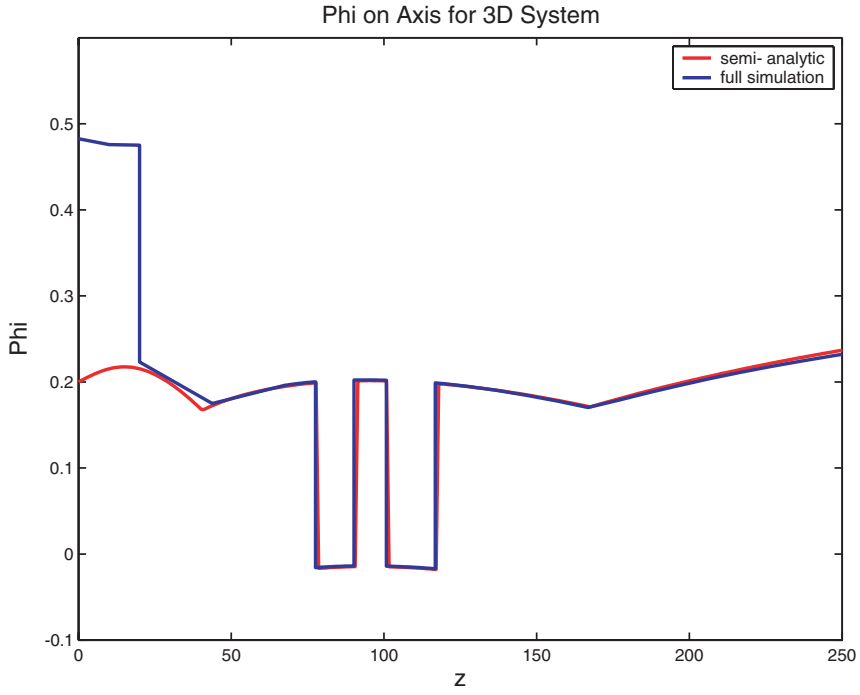


FIG. 5. *Same as Figure* 3 *except that the plot is for the potential* Φ *on the central axis* $r = 0$ *for a three-dimensional geometry.*

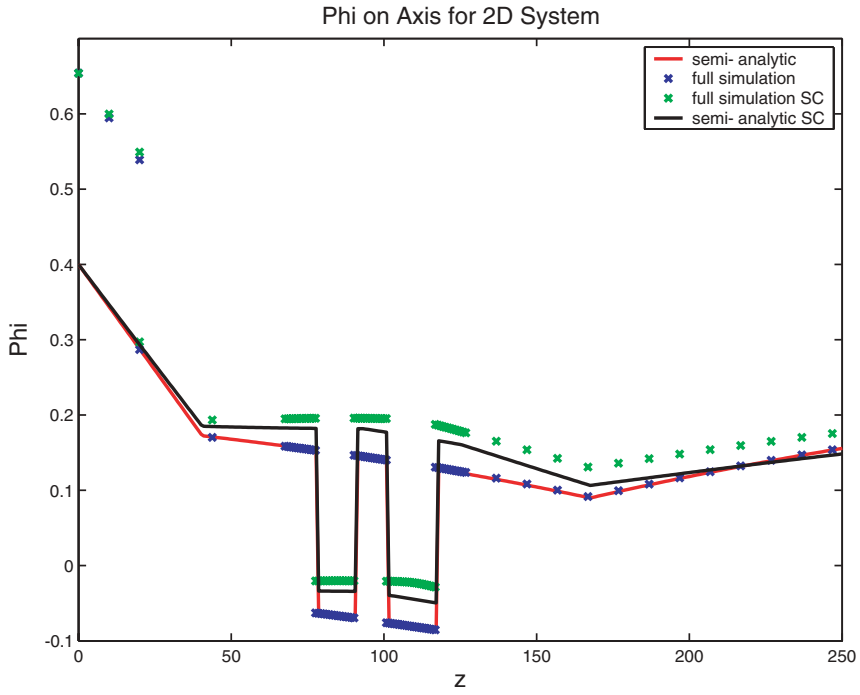


FIG. 6. *Same as Figure 4 but showing the full-scale simulation method with (green) and without (blue) self-consistent terms and the semianalytic model with (black) and without (red) self-consistent terms.*

TABLE 3
Lowest eigenvalues λ_1.

	Dot energies (meV)		Wire energies (meV)	
	simulation	semianalytic	simulation	semianalytic
1D	−144.1	−148.3	−195.9	−201.1
2D	−33.8	−37.3	−58.8	−60.8
3D	22.5	16.6	9.88	7.1
2D SC	11.6	−4.7	−3.7	−23.3

5. Double pinchoff designs. The principal virtue of the semianalytic model is that the eigenvalues in the design criteria (2.15) and (2.16) can be quickly computed, enabling rapid throughput as required for a design study or optimization exercise. In this section we describe a design that achieves the strictest design criterion (2.15) and (2.16) with $k = 1$, i.e., double pinchoff with a single electron in the dot and a single conductive state in the wire.

An example of a system achieving double pinchoff has design parameters

$$v_d = (dz_1, dz_2, dz_3, dz_4, dz_5, dz_6, R_g, R_b, \sigma_1, \sigma_2) \tag{5.1}$$
$$= (40, 94.3, 22.5, 30.3, 17.7, 50, 53.7, 876.7, 3.4 \times 10^{11}, 1.4 \times 10^{11})$$

and operation parameters $v_o = (\phi_g, \phi_b) = (-0.16, 1.47)$. The eigenvalues for this design are $(\lambda_1^d, \lambda_2^d, \lambda_1^w(0), \lambda_2^w) = (-0.67, 2.2, -0.92, 0.27)\ meV$. This system was found by a random search over the full space of possible design parameters and operation

parameters. In a search involving 10 million trial designs, 7 successful designs were found. This establishes existence (within the simulation) of a design meeting the double pinchoff goal. In section 7, however, we show that these double pinchoff designs are not robust with respect to fabrication errors.

6. Design robustness. When a prescribed design is implemented, the outcome will differ from the prescription due to errors and uncertainties in growth and fabrication, including variability in the layer thicknesses and gate sizes and variability in charge density in delta-doped layers due to uncertainties in both the doping level and the ionization fraction. Additional modeling uncertainties, such as uncertainties in the correct boundary conditions at the top of the device and additional physics such as self-consistent terms, are not accounted for in this analysis.

To find a design whose success is insensitive to the growth and fabrication uncertainties, we formulate a measure of design robustness. Assume that the errors in each of the various design parameters are independent and normally distributed and define α_k to be the standard deviation of the kth design parameter. Define a distance function d between two design vectors v and w as

$$d(v,w) = \left(\sum_{k=1}^{K} ((v_k - w_k)/\alpha_k)^2 \right)^{1/2}, \tag{6.1}$$

i.e., $d(v,w)$ is a measure of the distance between v and w in standard deviations. Next fix a design criterion by choosing the number K of allowed conduction states in the wire, and define the robustness R of a successful design v_s as the distance to the nearest unsuccessful design v_u, i.e.,

$$R(v_s) = \min_{v_u} d(v_s, v_u). \tag{6.2}$$

The design robustness optimization problem is to find the most robust design within the constraint set C from (2.1)–(2.6), i.e., v_s is chosen to be the successful design that achieves the following max-min:

$$\max_{v_s \in C} R(v_s) = \max_{v_s \in C} \min_{v_u \in C} d(v_s, v_u). \tag{6.3}$$

As an example for standard deviation of the fabrication, we take standard deviation of the growth processes (layer thicknesses) to be 3% (relative error), the standard deviation of the fabrication (gate sizes) to be 10 nm (absolute error), and the standard deviation of the charge density in the delta-doped layers to be 40% (relative error).

7. Analysis of failure modes. A direct random search for the design v_s that achieves the max-min in (6.3) would involve a double random search over two designs v_s and v_u. This can be considerably improved by analysis of the failure modes, i.e., the closest failed designs v_u for a given successful design v_s. This analysis relies on a linear approximation for the dependence of the eigenvalues λ in the design criteria (2.15) and (2.16), as functions of the gate voltages ϕ_g and ϕ_b.

A successful design v_s is one for which the four design inequalities in (2.15) and (2.16) form a quadrilateral (or triangular) set that has a nonempty intersection with the rectangular constraint set defined by (2.5) and (2.6), in the operation space (ϕ_g, ϕ_b). As the design parameter vector v is (smoothly) varied, the sides of the quadrilateral (or triangle) will (smoothly) vary. The first unsuccessful design v_u

is reached when the intersection becomes just a point. This characterizes the design v_u that occurs in (6.3).

The intersection of the operation window (i.e., the quadrilateral or triangle defined by (2.15) and (2.16)) and the constraint set (defined by (2.5) and (2.6)) can shrink to a point in either of two ways: First, the operation window can shrink to a point in the interior of the constraint set. Second, the operation window can move outside the constraint set with one vertex on the boundary of the constraint set.

We draw the operation window with coordinates $(-\phi_g, \phi_b)$, so that both coordinates are positive. Denote the boundaries of the operation window as follows:

$$a = \{(-\phi_g, \phi_b) : \lambda_1^d(\phi_g, \phi_b) = 0\}, \tag{7.1}$$

$$b = \{(-\phi_g, \phi_b) : \lambda_2^d(\phi_g, \phi_b) = 0\}, \tag{7.2}$$

$$c = \{(0, \phi_b) : \lambda_1^w(0, \phi_b) = 0\}, \tag{7.3}$$

$$d = \{(-\phi_g, \phi_b) : \lambda_2^w(\phi_g, \phi_b) = 0\}. \tag{7.4}$$

Also denote ac to be the point of intersection of the lines a and c if it exists, with coordinates $\phi_g(ac)$ and $\phi_b(ac)$, and similarly for the other intersections. Also denote $0a$ to be the intersection of a with the line $\phi_g = 0$. They have the following properties:

1. c is a horizontal line.
2. a, b, and d are lines with positive slope, with a and b steeper than d.
3. a and b cannot intersect (for $-\phi_g > 0$) and a is to the left of b.
4. $-\phi_g(ac) < -\phi_g(bc)$.
5. The operation window is nonempty if and only if $-\phi_g(ac) < -\phi_g(cd)$.
6. ad is the leftmost and the lowest point of the operation window.

From these properties, it follows that a nonempty operation window can have two possible configurations. If $-\phi_g(bc) < -\phi_g(cd)$, it is a quadrilateral with vertices ad, bd, bc, and ac, which is denoted as Type I. If $-\phi_g(bc) > -\phi_g(cd)$, a nonempty operation window is a triangle with vertices ad, cd, and ac, which is denoted as Type II. These two possibilities are shown in Figure 7.

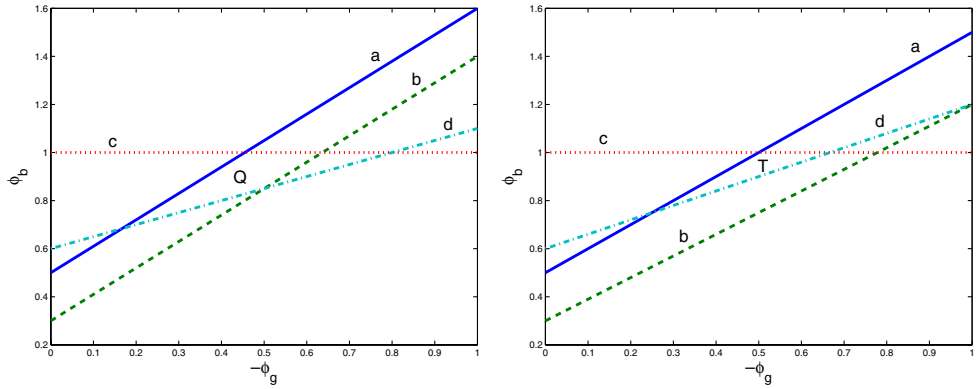


FIG. 7. *A schematic drawing of the operation windows in $(-\phi_g, \phi_b)$. The four lines that define the operation window are a, b, c, and d. The successful operation vectors are those in quadrilateral region labeled Q for the configuration on the left or in the triangular region labeled T for the configuration on the right.*

This information allows characterization of the failure modes:

A Collapse of the operation window can occur only as a transition from Type II, in which the three vertices of the triangle meet as one point acd. Failure mode A is characterized by existence of a triple intersection point acd, which is denoted as point A.

B If the operation window leaves the constraint region through the upper boundary, $\phi_b = \overline{\phi}_b$, then the final point of intersection of the two regions is ad. Failure mode B is characterized by existence of a triple intersection point ad with $\phi_b = \overline{\phi}_b$, which is denoted as point B.

C If the operation window leaves the constraint region through the right boundary, $-\phi_g = \overline{\phi}_g$, then the final point of intersection of the two regions is ad. Failure mode C is characterized by existence of a triple intersection point ad with $-\phi_g = \overline{\phi}_g$, which is denoted as point C.

D If the operation window leaves the constraint region through the lower boundary, denoted as failure mode D, then line c coincides with the lower constraint $\phi_b = \underline{\phi}_b$.

E If the operation window leaves the constraint region through the left boundary, $-\phi_g = \underline{\phi}_g$, then the final point of intersection of the two regions is bc in Type I or cd in Type II. Failure mode E is characterized by existence of a triple intersection point bc or cd with $-\phi_g = \underline{\phi}_g$, which is denoted as point E.

The three failure points A, B, and C, which are the ones that most frequently occur, are illustrated in Figure 8

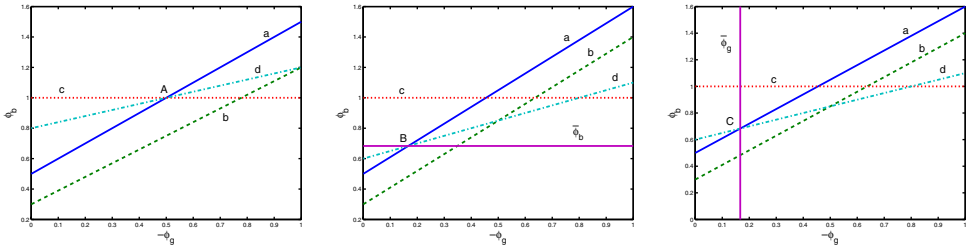


FIG. 8. *A schematic drawing of the failure modes in $(-\phi_g, \phi_b)$, including the four lines a, b, c, and d that define the operation window, and the constraining lines $\phi_b = \bar{\phi}_b$ in the middle and $-\phi_g = \bar{\phi}_g$ in the right. In mode A (left), the operation window has collapsed to a point A in the interior of the constraint set. In mode B (middle), the operation window intersects the constraint set in only a single point B on the upper boundary. In mode C (right), the operation window intersects the constraint set in only a single point C on the right boundary.*

The distance from a successful design v_s to one of the failure points A, B or C can be estimated through a linear approximation. For point A, let $(\Phi_g^A(v), \Phi_b^A(v))$ solve

$$\lambda_1^d(\Phi_g^A(v), \Phi_b^A(v), v) = 0, \tag{7.5}$$

$$\lambda_1^w(0, \Phi_b^A(v), v) = 0. \tag{7.6}$$

Then to leading order, since $\lambda_2^w(\Phi_g^A(A), \Phi_b^A(A), A) = 0$,

$$\begin{aligned} \lambda_2^w(\Phi_g^A(v), \Phi_b^A(v), v) &= \lambda_2^w(\Phi_g^A(v), \Phi_b^A(v), v) - \lambda_2^w(\Phi_g^A(A), \Phi_b^A(A), A) \\ &= (v - A) \cdot \nabla_v \lambda_2^w(\Phi_g^A(v), \Phi_b^A(v), v). \end{aligned} \tag{7.7}$$

At a minimum point, the design difference $v - A$ is parallel to the gradient in (7.7), so that

$$\min |v - A| = |\lambda_2^w(\Phi_g^A(v), \Phi_b^A(v), v)|/|\nabla_v \lambda_2^w(\Phi_g^A(v), \Phi_b^A(v), v)|. \tag{7.8}$$

A similar analysis can be carried out for B and C. For B, $\Phi_b^B(v) = \overline{\phi}_b$ and let $\Phi_g^B(v)$ solve

$$\lambda_1^d(\Phi_g^B(v), \overline{\phi}_b, v) = 0. \tag{7.9}$$

Then to leading order, since $\lambda_2^w(\Phi_g^B(B), \overline{\phi}_b, B) = 0$, it follows that

$$\min |v - B| = |\lambda_2^w(\Phi_g^B(v), \overline{\phi}_b, v)|/|\nabla_v \lambda_2^w(\Phi_g^B(v), \overline{\phi}_b, v)|. \tag{7.10}$$

For C, $\Phi_g^C(v) = \overline{\phi}_g$ and let $\Phi_b^C(v)$ solve

$$\lambda_1^d(\overline{\phi}_g, \Phi_g^C(v), v) = 0. \tag{7.11}$$

Then to leading order, since $\lambda_2^w(\overline{\phi}_g, \Phi_g^C(C), C) = 0$, it follows that

$$\min |v - C| = |\lambda_2^w(\overline{\phi}_g, \Phi_g^C(v), v)|/|\nabla_v \lambda_2^w(\overline{\phi}_g, \Phi_g^C(v), v)|. \tag{7.12}$$

The robustness R and the design robustness optimization problem can now be rephrased as

$$R(v_s) = \min\{|v - A|,\ |v - B|,\ |v - C|\}, \tag{7.13}$$

$$\max_{v_s \in C} R(v_s) = \max_{v_s \in C} \min\{|v - A|,\ |v - B|,\ |v - C|\} \tag{7.14}$$

in which $|v - A|$, $|v - B|$, and $|v - C|$ are defined by (7.8), (7.10), and (7.12). This has the advantage over the formulation (6.3) that it requires only a single random search for successful designs v_s rather than a double random search for v_s and v_d. For each v_s, the min is found by evaluation of the three quantities $|v - A|$, $|v - B|$ and $|v - C|$ from (7.8), (7.10), and (7.12).

For the design v_{dp} that achieved double pinchoff, as described in section 5, the design robustness distance (from (7.13)) is $R(v_{dp}) = 0.3$, which corresponds to probability of about 0.2 of successful design. The search for a more robust design through maximization of $R(v_s)$ as in (7.14) is described in section 8.

8. Design optimization. The search for a maximally robust design v_s in (7.14) can be accelerated by decomposition and some analysis. First select values for the geometrical design parameters $v' = (dz_1, dz_2, dz_3, dz_4, dz_5, dz_6, R_g, R_b)$.

For a given choice of geometrical parameters v', the possible values of the δ-doping densities σ_1 and σ_2 can be determined using the linear dependence of the eigenvalues λ_1^d and λ_1^w on σ_1 and σ_2, as well as on ϕ_g and ϕ_b. The operation window can be characterized as having the point ac inside the constraint set. The two equations (7.1) and (7.3) defining ac can be used to define a mapping between the operation vector $(\phi_g,\ \phi_b)$ and the density vector $(\sigma_1,\ \sigma_2)$. Then the constraint set defined by (2.5) and (2.6) can be mapped to a constraint set in the space of density vectors, which may need to be cut off to accommodate the constraints (2.4). To simplify, we choose a value of $(\phi_g,\ \phi_b)$ that is approximately in the center of the resulting polygon. This is illustrated in Figure 9.

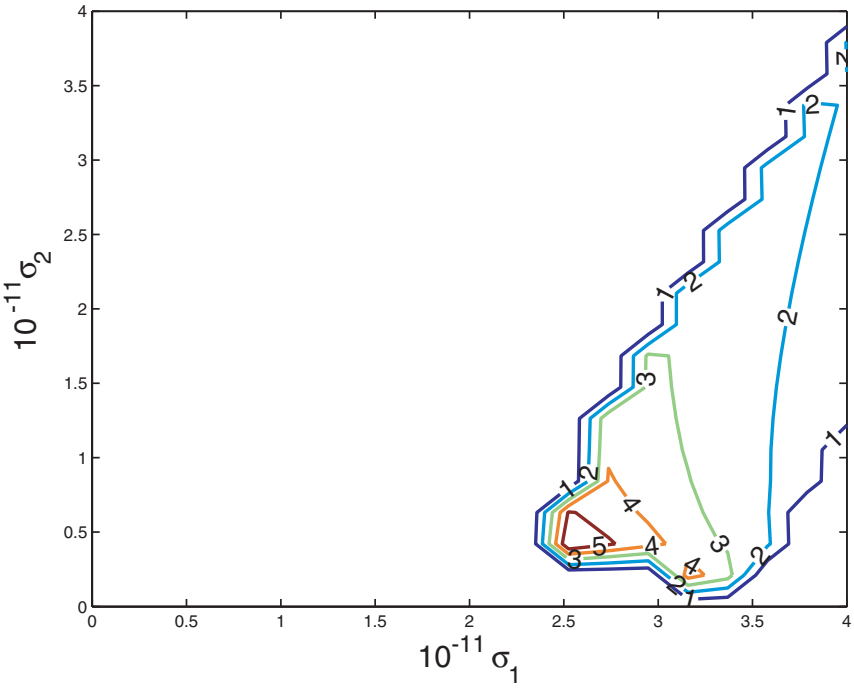


FIG. 9. *Design window in the space of doping densities* (σ_1, σ_2) *for one of the optimally robust designs.*

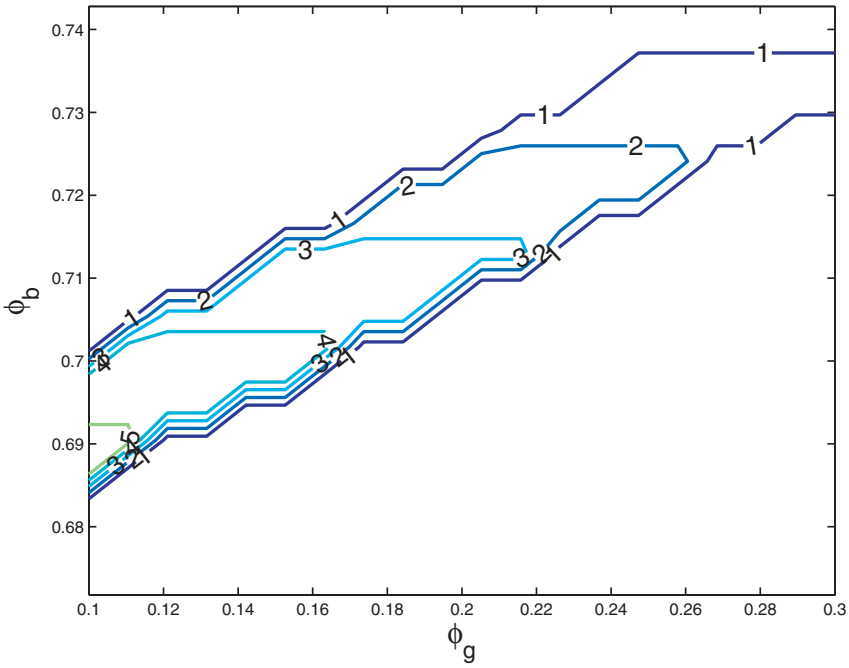


FIG. 10. *Design window in the space of operation vectors* $(-\phi_g, \phi_b)$ *for one of the optimally robust designs.*

Following this procedure for $K = 7$, we have found designs with robustness values of 2.5 or more. A typical result is

$$(8.1)\quad v = (dz_1, dz_2, dz_3, dz_4, dz_5, dz_6, R_g, R_b, \sigma_1, \sigma_2)$$
$$= (47.25, 49.31, 8.24, 16, 23.62, 50, 50.38, 608.28, 3.78 \times 10^{11}, 2.45 \times 10^{11}).$$

The operation vector is $(\phi_g, \phi_b) = (0.24, 1.91)$. The eigenvalues for this design are $(\lambda_1^d, \lambda_2^d, \lambda_1^w(0), d\lambda^w) = (-0.559, 3.60, -33.0, 4.0)$. There are seven transverse states in the wire and the robustness is 2.8, which corresponds to more than 99% probability of a successful design. The resulting operation window is shown in Figure 10.

9. Conclusions. In this work, we have developed a mathematical model for an electron spin qubit system, for the successful design of the system, and for optimization of the design robustness. In addition, we have developed a simple semianalytic model that is both sufficiently accurate to provide relevant results for the system and sufficiently fast to allow for the high throughput required by design and optimization studies. After some analysis to simplify the computation of design robustness, we have performed a random search for designs that satisfy the design criteria and for designs that are maximally robust.

From this search, we have found system designs that achieve double pinchoff, in the sense that they have a single electron in the quantum dot and a single conduction state in the quantum wire. These designs are not sufficiently robust to be practical, having a design robustness of only about 0.3, in terms of standard deviation using a current assessment of design uncertainties. By relaxing the design criterion to allow for a small number (e.g., $K = 7$) of conduction states in the wire, we have found designs that are more than 2.8 standard deviations from an unsuccessful design. Currently these designs are being built and tested for their electronic properties.

Several conclusions can be drawn from the present study. First is the importance of models at different levels of complexity. A full-scale model, as in [2], is needed to give reliable values for the system properties and to provide validation for simpler models. Simpler models that are less computationally intensive are also needed, however, to enable design and optimization studies on a reasonable time scale. In addition, we have been using nextnano^3 [13], a computational physics software package, which includes a much wider set of physics in order to check and validate the results from full-scale numerical solver for the Schrödinger–Poisson equation [2]. Second is the importance of analysis as a method for accelerating the random search that is often required in a design and optimization study. In the present study, the search for an optimally robust design was greatly aided by analysis of the failure modes for a design (i.e., the closest unsuccessful designs to a given successful design) and elimination of the charge variables using their special (linear) occurrence in the model.

REFERENCES

[1] G. Burkard, D. Loss, and D. P. DiVincenzo, *Coupled quantum dots as quantum gates*, Phys. Rev. B, 59 (1999), pp. 2070–2078.

[2] C. R. Anderson, *private communication*, 2004.

[3] I. A. Fedorov, K. W. Kim, R. E. Caflisch, and E. Yablonovitch, *A Scalable Quantum Gate Design for Quantum Computation*, preprint, 2004.

[4] T. Fujisawa, D. G. Austing, Y. Tokura, Y. Hirayama, and S. Tarucha, *Allowed and forbidden transitions in artificial hydrogen and helium atoms*, Nature, 419 (2002), pp. 278–281.

[5] M. Fujiwara, M. Takeoka, J. Mizuno, and M. Sasaki, *Exceeding classical capacity limit in quantum optical channel*, Phys. Rev. Lett., 90 (2003), 167906.

[6] L. K. Grover, *Quantum mechanics helps in searching for a needle in a haystack*, Phys. Rev. Lett., 79 (1997), pp. 325–328.
[7] M. Gyure, *private communication*.
[8] X. Hu and S. Das Sarma, *Hilbert-space structure of a solid-state quantum computer: Two-electron states of a double-quantum-dot artificial molecule*, Phys. Rev. A, 61 (2000), 062301.
[9] A. Imamoglu, D. D. Awschalom, G. Burkard, D. P. DiVincenzo, D. Loss, M. Sherwin, and A. Small, *Quantum Information Processing Using Quantum Dot Spins and Cavity QED*, Phys. Rev. Lett., 83 (1999), pp. 4204–4207.
[10] B. E. Kane, *A silicon-based nuclear spin quantum computer*, Nature, 393 (1998), pp. 133–137.
[11] J. Levy, *Quantum-information processing with ferroelectrically coupled quantum dots*, Phys. Rev. A, 64 (2001), 052306.
[12] D. Mozyrsky, V. Privman, and M. L. Glasser, *Indirect interaction of solid-state qubits via two-dimensional electron gas*, Phys. Rev. Lett., 86 (2001), pp. 5112–5115.
[13] *nextnano^3: Next Generation 3D Nano Device Simulator*. http://www.wsi.tum.de/nextnano3/.
[14] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Communication*, Cambridge University Press, Cambridge, UK, 2000.
[15] P. W. Shor, *Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer*, SIAM J. Comput., 26 (1997), pp. 1484–1509.
[16] S. Somaroo, C. H. Tseng, T. F. Havel, R. Laflamme, and D. G. Cory, *Quantum simulations on a quantum computer*, Phys. Rev. Lett., 82 (1999), pp. 5381–5384.
[17] R. Vrijen, E. Yablonovitch, K. L. Wang, H. W. Jiang, A. Balandin, V. Roychowdhury, T. Mor, and D. DiVincenzo, *Electron spin resonance transistors for quantum computing in silicon-germanium hetero-structures*, Phys. Rev. A, 62 (2000), 012306.
[18] L.-X. Zhang, P. Matagne, J. P. Leburton, R. Hanson, and L. P. Kouwenhoven, *Single-electron charging and detection in a laterally coupled quantum-dot circuit in the few-electron regime*, Phys. Rev. B, 69 (2004), 245301.

SIAM J. APPL. MATH.
Vol. 65, No. 4, pp. 1305–1327

THE EFFECT OF DISPERSAL PATTERNS ON STREAM POPULATIONS*

FRITHJOF LUTSCHER†, ELIZAVETA PACHEPSKY‡, AND MARK A. LEWIS§

Abstract. Individuals in streams are constantly subject to predominantly unidirectional flow. The question of how these populations can persist in upper stream reaches is known as the "drift paradox." We employ a general mechanistic movement-model framework and derive dispersal kernels for this situation. We derive thin- as well as fat-tailed kernels. We then introduce population dynamics and analyze the resulting integrodifferential equation. In particular, we study how the critical domain size and the invasion speed depend on the velocity of the stream flow. We give exact conditions under which a population can persist in a finite domain in the presence of stream flow, as well as conditions under which a population can spread against the direction of the flow. We find a critical stream velocity above which a population cannot persist in an arbitrarily large domain. At exactly the same stream velocity, the invasion speed against the flow becomes zero; for larger velocities, the population retreats with the flow.

Key words. nonlocal dispersal, critical domain size, spread speed, drift paradox

AMS subject classifications. 92B05, 45C05, 34K05

DOI. 10.1137/S0036139904440400

1. Introduction.

Many organisms, ranging from river-dwelling flora and fauna to gut-dwelling bacteria, live in environments with predominantly unidirectional flow. As with simple chemostat residents [35], organisms that persist in the presence of such unidirectional flow must resist being washed out by their moving surroundings. The success of many organisms in maintaining a foothold, even at high flow rates, has given rise to the so-called drift paradox of persistence in unidirectional flow [27, 28].

While possible solutions of the drift paradox have been discussed in the ecological literature [27, 28, 41, 22], until recently the discussion has lacked quantitative scrutiny in the form of models that can be used to predict the effect of environmental variables on maintaining the population. Two recent papers have begun to remedy this lack and have analyzed conditions for species persistence and population spread into upstream environments, both analytically and numerically. The models used there are PDE systems, such as a single compartment model with growth, advection, and diffusion [36], or a two-compartment model with separate mobile and stationary states corresponding to aquatic and benthic populations [32].

Flows in river systems are very complex and include, for example, up- and downriver currents as well as turbulent long-distance movement of biota [1]. Although systems of PDEs are the workhorse for spatial ecology models in continuous space [15],

*Received by the editors February 2, 2004; accepted for publication (in revised form) August 20, 2004; published electronically April 14, 2005.
http://www.siam.org/journals/siap/65-4/44040.html
†Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, T6G2G1, Canada (flutscher@math.ualberta.ca). This author was supported as a postdoctoral fellow through the Pacific Institute for the Mathematical Sciences.
‡Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, CA 93106 (pachepsk@lifesci.ucsb.edu). This author was supported by National Science Foundation grant DEB01-08450.
§Department of Mathematical and Statistical Sciences and Department of Biological Sciences, University of Alberta, Edmonton, T6G2G1, Canada (mlewis@math.ualberta.ca). This author was supported by NSERC, Collaborative Research Opportunity, and Canada Research Chair grants.

their application is limited as they depict the complex asymmetrical spatial flow in a river through simple advection and diffusion.

Integrodifferential equations [16] are related to PDEs but encompass more general movement patterns than diffusion and advection. In particular, the modeling formalism can allow for a detailed description of the complicated dispersal that arises through river flow. The added realism of integrodifferential models comes at a price: much of the theory for PDEs on problems such as critical domain size for species persistence [34] or population spread [18] has not yet been formulated for their integrodifferential cousins, but see [26] for invasion speeds. We develop some of the theory needed for analysis in this paper.

In this paper we revisit the drift paradox, employing integrodifferential models that allow us to include long-distance dispersal. We show how the long-distance dispersal changes previous washout predictions [32, 36]: populations can always persist under high flow rates providing rare, long-distance dispersal events are sufficient to allow maintenance of a foothold in the river. Our results contrast with those of Lockwood, Hastings, and Botsford [24], where long-distance dispersal is discounted as playing a role in determining population persistence. While our model and application are new, we draw on theoretical ideas that have a distinguished history in the theory of spatial ecology.

The critical domain size is a fundamental ecological quantity that gives the minimal size of a habitable area required for species survival. In turn, it provides an important tool in reserve design and conservation [6, 8]. The first models for the critical domain size using diffusion equations date back to the 1950s [34, 17]. The analysis has since been extended to cover more complex spatial domains [9], the influence of advection [29, 32], and discrete-time integrodifference equations [20, 40, 25].

Another relevant ecological metric is the speed of spread, which is important in a wide range of ecological applications. While some invasions are intended, such as the introduction of biological control agents [4], others can be devastating for native species being out-competed by invaders and for species diversity. The spread of diseases is a worldwide problem and can be treated in the same modeling framework [26].

While the idea of having stationary and mobile compartments has recently been used by numerous authors, for example, to model protein movement in a cell nucleus [10], population dynamics with diffusive movement [23, 13], or wavelike movement [14], the idea of coupling such models to asymmetric spatial flow dynamics via advection and diffusion, as in [32], is a recent one (but see [5]).

We start our investigation by presenting a general framework to derive dispersal kernels from mechanistic movement models, and we apply this framework to derive a thin-tailed and a fat-tailed kernel. In section 3 we present the general integrodifferential model and develop the theoretical results on critical domain size and invasion speeds. The following three sections contain the application of the general theory to persistence and spread in streams. Three cases for dispersal kernels are considered: thin-tailed (section 4), a weighted sum of thin-tailed kernels, accounting for short- and long-distance dispersal (section 5), and, finally, fat-tailed (section 6).

2. Modeling dispersal. In this section, we use a mechanistic approach for individual movement to derive theoretical forms of dispersal kernels. A *dispersal kernel* describes the probability that an individual moves from one location to another in a certain time interval. Such dispersal kernels, also referred to as redistribution kernels or seed shadows, have been measured for many organisms [30]. The mechanistic approach taken here allows for explicit description of the movement process and be-

havior. We assume that population dynamics happen on a much slower time scale than individual movement and hence can be neglected while deriving the kernel. This separation of time scales occurs frequently, and it is certainly true for stream insects, where dispersal can occur over daily time scales, while significant growth typically requires monthly or yearly time scales. The general theory presented here follows, but significantly extends, the results in [30] and is applied to derive a thin-tailed and a fat-tailed dispersal kernel as specific examples for analysis and further development later in the paper.

We denote $\omega(t, x; y)$ as the probability density of the location of a mobile individual with initial location $x = y$. We assume that the individual moves for a random length of time, T, after which it settles, and that the random variable T has a given probability density $p(t)$. The dispersal kernel is now defined as the probability density of stopping points from given initial location, i.e.,

$$\kappa(x, y) = \int_0^\infty p(t)\omega(t, x; y)dt. \tag{2.1}$$

If $\omega(t, x; y)$ depends only on the signed distance from the starting point $\xi = x-y$ rather than the exact location, we simply write $w(t, \xi) = \omega(t, x; y)$ and $k(\xi) = \kappa(x, y)$. Most dispersal kernels in this paper are of this form. For an exception, see Appendix E. When the individual moves by Brownian motion with diffusion coefficient D, the function $w(t, x)$ is the fundamental solution of the heat equation on the real line,

$$w(t, x) = \frac{\exp\left(\frac{-x^2}{4Dt}\right)}{\sqrt{4\pi Dt}}. \tag{2.2}$$

When drift at rate v is included with the Brownian motion, the function $w(x, t)$ is given by (2.2) with x replaced by $x - vt$.

However, if dispersing individuals can jump long distances in short time intervals, the Brownian motion model may not be valid. For example, the Lévy flight model [11] assumes that arbitrarily large jumps can occur over short time scales. The result is a distribution of jump distances which has no variance. In this "anomalous diffusion" case, a typical form for w is the Cauchy distribution

$$w(t, x) = \frac{t}{\rho\pi}\left[\left(\frac{x}{\rho}\right)^2 + t^2\right]^{-1}. \tag{2.3}$$

The parameter ρ has dimension [space/time] and stands for an effective speed. Details of how (2.3) can be derived from a random walk model for individuals are given in Appendix A. As above, we introduce drift at rate v by replacing x with $x - vt$.

We now turn to modeling the stopping time T. The simplest possible assumption is that all individuals disperse for the same, fixed, length of time t_0. In this case

$$p(t) = \delta(t - t_0), \tag{2.4}$$

so that (2.1) yields $k(x) = w(x, t_0)$. Thus, for a fixed dispersal time t_0, the dispersal kernel (2.1) is simply the Gaussian (2.2), possibly shifted if $v \neq 0$, or the Cauchy distribution (2.3), again possibly shifted if $v \neq 0$, evaluated at time t_0. In Figure 2.1, we plot the shapes of these kernels.

A more general form of stopping times comes from defining $\alpha(t)$ as the *settling* or *failure rate* [37], i.e., $\alpha(t)\,dt$ as the probability that the individual ends its movement

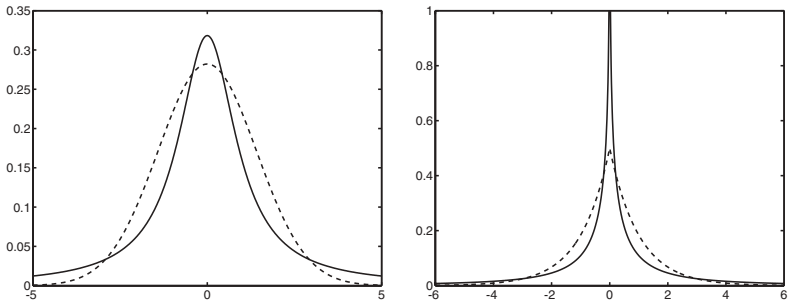


FIG. 2.1. *The plot on the left shows the thin-tailed Gaussian (dashed) and fat-tailed Cauchy (solid) distribution as given in* (2.2) *and* (2.3) *for* $t_0 = 1$ *with parameters* $D = 1$ *and* $\rho = 1$*, respectively. The plot on the right shows the Laplace distribution* (2.6) *(dashed) and the fat-tailed distribution* (2.7) *(solid). Parameters are as above and the settling rate is* $\alpha = 1$*. Note that the fat-tailed distribution has a singularity at the origin.*

during $[t, t + dt)$. The probability density for the stopping times of the individual, also called the lifetime probability density, is then

$$p(t) = \alpha(t) \exp\left(-\int_0^t \alpha(s)\, ds\right). \tag{2.5}$$

The argument of the exponential function is known as the *hazard function* [37].

For constant settling rate α, the dispersal kernel (2.1) is the Laplace transform of the probability density ω with respect to time. In the case of Brownian motion (2.2), the kernel (2.1) becomes the Laplace distribution [7],

$$k(\xi) = \sqrt{\frac{\alpha}{4D}} \exp\left(-\sqrt{\frac{\alpha}{D}}|\xi|\right). \tag{2.6}$$

For constant settling rate α and the Cauchy redistribution function (2.3), the kernel (2.1) becomes the fat-tailed kernel

$$\begin{aligned} k(\xi) &= \theta\, \Re\left\{E_1(i\theta\xi) \exp(i\theta\xi)\right\}/\pi \\ &= -\theta\left(\cos(\theta\xi)\,\mathrm{ci}(\theta\xi) + \sin(\theta\xi)\,\mathrm{si}(\theta\xi)\right)/\pi, \end{aligned} \tag{2.7}$$

where $\theta = \alpha/\rho$. The functions E_1, ci, and si are the exponential, cosine, and sine integrals, respectively,

$$\mathrm{E}_1(x) = \int_1^\infty \frac{\exp(xz)}{z}\, dz,\ \mathrm{ci}(x) = -\int_1^\infty \frac{\cos(xz)}{z}\, dz,\ \mathrm{si}(x) = -\int_1^\infty \frac{\sin(xz)}{z}\, dz. \tag{2.8}$$

The kernels given by (2.6) and (2.7) are plotted in Figure 2.1.

Adding drift into the last two scenarios does *not* simply shift the kernels (2.6) and (2.7) as it did above but instead causes a different kind of asymmetry in dispersal, as we show later. In section 4.1, we employ a somewhat simpler method to derive the kernel for Brownian motion with drift. The case of Lévy flight with drift is done in section 6.

In Appendix B we generalize the simple model of Brownian motion to the case of two (and potentially more) dispersal modes. Individuals switch between these modes. We show that corresponding dispersal kernels can be derived explicitly for constant settling rate.

3. The model equation, critical domain size, and spread speed. In this section, we present the general model for a population subject to population dynamics and spatial movement. It has the form of an integrodifferential equation, for which we give alternative derivations. We then state the main assumptions and prove formulas for the critical domain size and the spread speed of the population.

We consider a single population, which is described by its density $u(t,x)$. Population dynamics such as birth and death of individuals are summarized in the function $f(u)$. Then the dispersal time scale is small compared to the population dynamics time scale; dispersal can be modeled by a position-jump process with jumping rate μ [31]. If an individual jumps, the dispersal kernel $\kappa(x,y)$, as discussed in section 2, describes the probability that the individual moves from some point y to x. Then the evolution of the population density is governed by the following integrodifferential equation:

$$u_t(t,x) = f(u(t,x)) - \mu u(t,x) + \mu \int_\Omega \kappa(x,y)u(t,y)dy. \tag{3.1}$$

The domain of integration Ω will depend on the question we study. In the case of the critical domain size, it will be a bounded interval; in the case of invasion speeds, it will be the real line. Although the model formulation is valid in spatial domains of any dimension, we will restrict ourselves to the one-dimensional case since the applications below will be to systems with unidirectional flow. We assume that the function f is a single-hump function, i.e., $f(0) = f(\bar{u}) = 0$, and $f > 0$ on $(0,\bar{u})$. To prove Theorem 3.2, we will need more assumptions on f and k, which we state then.

There are several ways to derive (3.1). We present a novel approach emphasizing the separation of time scales. Then we present the necessary theoretical results about the critical domain size and invasion speeds.

3.1. Model derivations. Besides the derivation in [31], (3.1) is derived in the ecological literature from a random walk process with variable move length [39]. Reaction and movement are assumed to be on the same time scale [12]. Recently, a very careful derivation of (3.1) has been presented where some scaling issues have been avoided [16].

Here, we present an alternative derivation that respects and even relies on the fact that movement often happens on a much faster time scale than population dynamics. We start by dividing the population into mobile and stationary classes, u and v, respectively, and assume that birth and death processes affect only stationary individuals. Stationary individuals start moving with rate μ, and mobile individuals settle with rate σ. Then we obtain the system

$$u_t = f(u) - \mu u + \sigma v, \qquad v_t = G[v] + \mu u - \sigma v, \tag{3.2}$$

where G is a differential operator describing movement, e.g., $G = D\Delta$ (diffusion) or $G = D\Delta - V\nabla$ (advection and diffusion). Recently, there has been increasing interest in this or similar systems [5, 10, 13, 23, 32]. To apply the quasi steady-state assumption that movement happens on a much faster time scale than population dynamics, we introduce the scaling parameter $\varepsilon = \mu/\sigma$ and rescale v and G in (3.2) to obtain

$$u_t = f(u) - \mu u + \mu\tilde{v}, \qquad \varepsilon\tilde{v}_t = \tilde{G}[\tilde{v}] + \mu u - \mu\tilde{v}, \tag{3.3}$$

where $\tilde{}$ denotes the rescaled quantities. Under the quasi steady-state assumption

$\varepsilon \to 0$, the equation for $\tilde{v}$ gives the linear differential operator

$$\mu u = \left(\mu - \tilde{G}\right) \tilde{v}. \tag{3.4}$$

System (3.2) becomes (3.1), with $\kappa(x, y)$ denoting the Green's function of (3.4), i.e.,

$$\tilde{v}(x) = \int \kappa(x, y) u(y) dy. \tag{3.5}$$

3.2. Critical domain size. As a first step in the analysis of (3.1), we now study the critical domain size problem. We find that parameter space can be divided into two parts, one that allows persistence independently of domain size and dispersal kernel, and one in which persistence depends on these two factors. We assume that there is no immigration into the domain. A population will persist if it grows at low density; therefore, we study conditions such that the zero steady state is unstable. The linearization of (3.1) on the interval $[0, L]$ is given by

$$u_t(t, x) = (r - 1)u(t, x) + \int_0^L \kappa(x, y) u(t, y) dy, \tag{3.6}$$

where we have rescaled time by the rate of movement μ and abbreviated $r = f'(0)/\mu$ as the rescaled growth rate at low density. From (3.6), we immediately see that if $r > 1$, then the zero steady state is unstable *independently* of the domain size and the kind of movement individuals perform. On the other hand, if $r < 1$, then the stability of the zero solution depends on the integral expression in (3.6). We assume that the integral operator

$$I[\phi](x) = \int_0^L \kappa(x, y) \phi(y) dy \tag{3.7}$$

has a unique simple dominant eigenvalue ν for an appropriate choice of function space. In Appendix C, we discuss possible choices and show the following result.

THEOREM 3.1. *Assume that κ is independent of L. The unique simple dominant eigenvalue ν of* (3.7) *is a strictly increasing function of the domain length L. Next, assume $f(0) = 0$ and $f'(0) > 0$. Then the zero steady-state solution of* (3.1) *is unstable provided $\nu(L) > 1 - r$.*

The condition that κ be independent of L means that dispersing individuals do not perceive domain boundaries or at least do not alter their movement behavior there. For example, aquatic individuals in a river stretch without breaks (source, mouth, waterfall), or in a no-fishing zone, or wind-dispersed seeds. In those cases, the dispersal kernel derived on the infinite domain is simply cut off at the domain boundaries [40]. If the movement behavior is altered at the boundary, then the kernel will depend on L (see section E). Then the first statement of the theorem is shown by showing that the smallest eigenvalue of the differential operator (3.4) is a decreasing function of L. In general, this follows from standard arguments; however, in the special case of zero-flux boundary conditions at both ends (i.e., no loss from the domain) this eigenvalue is independent of L.

According to the theorem, the critical domain size is given by $\nu(L) = 1 - r$. In the original nonscaled parameters, the population can persist if

$$f'(0) > \mu(1 - \nu). \tag{3.8}$$

Condition (3.8) is a refinement of the unconditional persistence in case $r > 1$, which we found above. Its interpretation gives a possible explanation of the drift paradox as follows. If the population growth rate at low density, $f'(0)$, exceeds the rate at which individuals move, μ, then the population will always persist, independently of the length of the domain and the kind of movement. In particular, the population can persist in an environment with unidirectional flow. This conclusion was also reached as one possible explanation of the drift paradox in [32]. If $f'(0)$ is smaller than μ, then persistence depends on the term $(1 - \nu)$. As the leading eigenvalue, ν asymptotically gives the fraction of individuals that remains in the domain during dispersal, and consequently, $(1 - \nu)$ is the fraction of individuals leaving the domain due to dispersal. Therefore, if the rate at which individuals move times the probability that they leave the domain during dispersal exceeds the population growth rate, then the population will go extinct. A similar switch from conditional to unconditional persistence in a PDE system was found in [13] (without advection) and [32] (with advection).

3.3. Spread speed. In the previous section, we analyzed population persistence on a bounded domain. Here, we look at population spread into an unbounded, previously uninhabited domain. We first derive the *minimal speed of a traveling wave* of the linearized system (3.6). We follow the usual line of argument, emphasizing the direction in which the wave is moving [26]. In systems with unidirectional flow, the spread in the direction of the drift will be faster than against the drift. This asymmetry requires some modification in the definition of the *asymptotic spreading speed* [3] for the nonlinear model. After we give the modified definition, we show in Theorem 3.2 that the minimal traveling wavespeed and the asymptotic spreading speed coincide.

To determine the wave speed of the linear system, we assume that the kernel is of the form $\kappa(x, y) = k(x - y)$ and change to traveling wave coordinates, $z = x - ct$, where c is the speed of a traveling wave. Then (3.6) gives the following equation for the profile ψ of a traveling wave:

$$-c\psi'(z) = (r - 1)\psi(z) + \int k(z - w)\psi(w)dw. \tag{3.9}$$

In this linear equation, we make the exponential ansatz $\psi(z) = e^{-sz}$, with $s > 0$ ($s < 0$), such that asymptotically, $\psi \to 0$ as $z \to \infty$ ($z \to -\infty$). After canceling equal terms on both sides, we get the characteristic equation

$$sc + 1 - r = \int_{-\infty}^{\infty} k(w)e^{sw}dw =: M(s) \tag{3.10}$$

for $s \neq 0$, where M stands for the moment generating function of k. We will always assume that advection points to the right. Therefore, waves with positive c travel in the direction of advection, and waves with negative c travel against the advection. From (3.10), which will be of use later, the minimal wave speeds are derived as in [26] and given by

$$c^+ = \inf_{s>0} \frac{r - 1 + M(s)}{s}, \quad c^- = \sup_{s<0} \frac{r - 1 + M(s)}{s} \tag{3.11}$$

for waves with decreasing (c^+) and increasing (c^-) profile. Here, we assume that the moment generating function exists at least for some interval containing zero. In

section 6, we discuss the case of a kernel whose moment generating function does not exist except at $s = 0$.

The representation (2.1) of the dispersal kernel for arbitrary settling rate (see (2.5)) is particularly useful in connection with formula (3.11) because the moment generating function of the Gaussian distribution is known. Since the moments of k involve integration in the spatial variable only and since the stopping times are independent of the spatial location, the moment generating function of k is given by

$$M(s) = \int_0^\infty p(t) \exp(Dts^2) dt. \tag{3.12}$$

The concept of the asymptotic spreading speed (henceforth simply referred to as spread speed) for the nonlinear equation was introduced by Aronson and Weinberger [3] and has since been explored in many publications; see [38]. To accommodate for asymmetric spread, we define spread speeds $c_\pm^*$ by the condition

$$\lim_{t\to\infty} u(t, x + ct) = \begin{cases} \bar{u}, & c_-^* < c < c_+^*, \\ 0, & c < c_-^* \text{ or } c > c_+^*, \end{cases} \tag{3.13}$$

where $\bar{u} > 0$ is the positive zero of f, i.e., $f(\bar{u}) = 0$.

THEOREM 3.2. *Assume that f satisfies $f(0) = 0 = f(\bar{u})$ for some $\bar{u} > 0$, $f'(0) > 0$, and the subtangential condition $f(u) \le f'(0)u$. Assume that the kernel satisfies the technical conditions stated in Appendix* D. *Then the spread speeds of the nonlinear equation* (3.1) *are given by* (3.11), *i.e., $c_\pm^* = c^\pm$.*

The proof of this theorem in Appendix D uses the upper bound for the spread speed from [26]. To show that the upper bound equals the lower bound, we construct subsolutions of (3.11) adapting the proof in [2] for a simple epidemic model.

4. A model with unidirectional flow. We now apply the general model (3.1) to study systems with unidirectional flow and the influence of the flow on the critical domain size and the spread speed. The biological system motivating our study is a population of aquatic insects in streams, and our results give possible explanations of the drift paradox. At first, we derive an appropriate dispersal kernel. Then we compute the critical domain size as well as the spread speeds with and against the flow direction. We show that these two important ecological characteristics are related as follows. The spread speed against the flow decreases as the advection increases until, at some critical advection speed, there is no spread against the flow direction. On the other hand, the critical domain size increases with the advection speed until, at some critical advection speed, it becomes infinite, i.e., the population cannot persist in a domain of any size. We show that the two critical advection speeds, indeed, coincide.

4.1. A dispersal kernel with advection. We derive a dispersal kernel that represents the movement of aquatic insects in streams. The larvae of these insects reside on the bottom of the stream, from where they periodically jump into the water column, where they are subject to the flow. Our submodel for individual movement consists of diffusion and advective flow, and we assume constant settling rate. We think of advection as representing the drift velocity experienced by the larvae and of diffusion as a first approximation to the variability in flow speed and direction. Denoting $z(t, x)$ as the density of moving individuals, we obtain the equation

$$z_t = Dz_{xx} - vz_x - \alpha z, \tag{4.1}$$

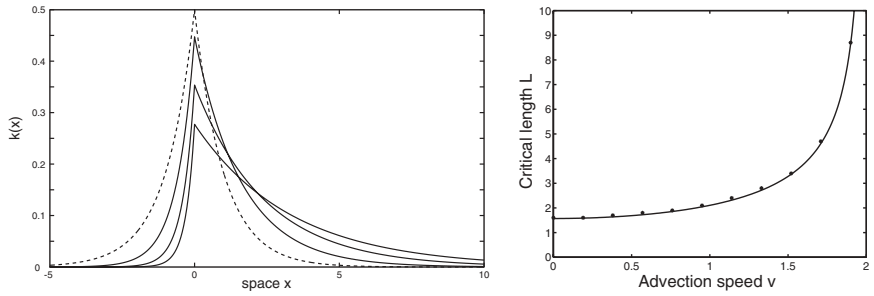


FIG. 4.1. *The picture on the left shows the dispersal kernel* (4.4) *with* $D = 1$, $\alpha = 1$, *and* $v = 1, 2, 3$ *in decreasing height of the peak, solid lines. For comparison, the symmetric kernel for* $v = 0$ *is plotted as the dashed line. The plot on the right gives the critical domain size as a function of the advection as in* (4.6). *The parameters are* $D = 1$, $\alpha = 1$, *and* $r = 0.5$. *The solid line is the analytical expression* (4.6), *stars are numerical results computing the eigenvalue of the integral operator* (C.1), *using Simpson's rule on* 1401 *data points in the interval* $[0, 1]$.

where D is the diffusion constant, v is the advection velocity, and α is the settling rate. Integrating (4.1) over $0 \le t \le \infty$ and applying the initial condition $z(0, x) = \delta(x)$ as well as (B.2), we observe that the dispersal kernel k satisfies

$$\frac{D}{\alpha} k_{xx} - \frac{v}{\alpha} k_x - k = -\delta, \tag{4.2}$$

i.e., k is the Green's function from (3.2). The characteristic equation of (4.2) is $Da^2 - va - \alpha = 0$ with solutions $a_1 > 0$ and $a_2 < 0$, given by

$$a_{1,2} = \frac{v}{2D} \pm \sqrt{\frac{v^2}{4D^2} + \frac{\alpha}{D}}. \tag{4.3}$$

Using the asymptotic boundary conditions for $x \to \pm\infty$ and the matching condition at $x = 0$, we find that k is of the form

$$k(x) = A \exp(a_1 x), \quad x \le 0, \quad \text{and} \quad k(x) = A \exp(a_2 x), \quad x \ge 0. \tag{4.4}$$

The value of the constant A is determined by the condition $\int_{-\infty}^{\infty} k(x)dx = 1$, which leads to

$$A = \frac{a_1 a_2}{a_2 - a_1} = \frac{\alpha}{\sqrt{D(v + 4\alpha)}}. \tag{4.5}$$

Alternatively, this kernel can be expressed by substituting $x \to x - vt$ in (2.2), $\alpha =$ const. in (2.5), and inserting the result in (2.1). In the special case $v = 0$, the Laplace kernel (2.6) results. We plot the shape of k in Figure 4.1 for different values of v while keeping D, α constant. In Appendix E, we contrast the kernel derived here for an infinite domain with a kernel on a finite domain with mixed boundary conditions of the same type as in [32, 36].

4.2. Critical domain size. From the previous section we know that the population persists unconditionally if $r > 1$. For $r < 1$, we have to find L such that $\nu(L) = 1 - r$; see (3.8). This can be calculated analytically. In Appendix F, we convert the integral equation (C.1) into a differential equation, extending earlier work

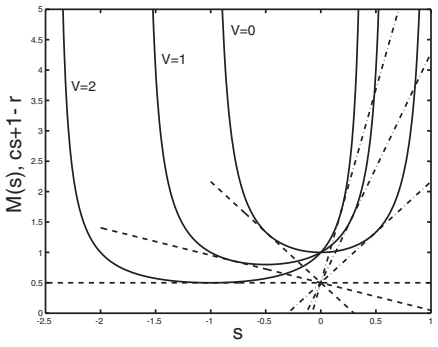


FIG. 4.2. *The hyperbolas are the moment generating function $M(s)$ (4.8) for three different values of v with $D = 1$ and $\alpha = 1$. The straight lines correspond to the left-hand side of (3.10). The slopes of these lines correspond to the spread speeds $c^{\pm}$ according to (3.11). For a more thorough explanation, see corresponding text. For $v = 0$ upstream and downstream spread speed are the same. For $v = 1$ the speed is faster downstream than upstream. For $v = 2$ the upstream spread stops.*

for symmetric kernels [20, 40], and obtain the following expression for L in terms of the eigenvalue ν and the dispersal related constants $a_{1,2}$ from (4.3):

$$L = \frac{4 \arctan\left(\sqrt{\frac{4a_1|a_2|}{\nu(a_1-a_2)^2} - 1}\right)^{-1}}{(a_1 - a_2)\sqrt{\frac{4a_1|a_2|}{\nu(a_1-a_2)^2} - 1}}. \tag{4.6}$$

Setting $\nu = 1 - r$, we can hence determine the critical domain size, which we plot in Figure 4.1 as a function of the advection speed v. As expected, the critical domain size is an increasing function of advection speed. From the plot, it appears that $v = 2$ is the critical advection speed, above which the population cannot persist in a domain of any length. In (4.6), L approaches infinity as the square root in the denominator approaches zero. Hence, the critical advection speed is defined by

$$\nu = 1 - r = \frac{4\frac{\alpha}{D}}{\frac{v^2}{D^2} + 4\frac{\alpha}{D}}. \tag{4.7}$$

For the set of parameters above, $v = 2$ is indeed the critical advection speed.

4.3. Spread speed. We use formulas (3.10) and (3.11) and Theorem 3.2 to determine the speed of spread. The moment generating function for the generalized Laplace kernel (4.4) is given by

$$M(s) = \frac{a_1 a_2}{(a_1 + s)(a_2 + s)}, \qquad -a_1 < s < -a_2. \tag{4.8}$$

In Figure 4.2, we plot the hyperbola $M(s)$ with y-intercept $M(0) = 1$ for three different values of the advection speed v. According to (3.10), we also plot straight lines with slope c, the propagation speed, and y-intercept $1 - r < 1$. As given in (3.11), we plot these straight lines for minimal values of $|c|$, such that the straight line and the hyperbola have a point in common, i.e., we plot the case that the line is tangent to the hyperbola. The resulting slopes give the minimal wave speed.

We find exactly two tangent lines. One of them (dash-dot line) always has positive slope, independent of the advection speed $v \geq 0$. This slope is the spread speed c^+

in the direction of advection. It increases with advection. For the other tangent line, we distinguish two cases. First note that the hyperbola is always positive since we assume k to be nonnegative. If now $r > 1$, then the y-intercept of the straight line is negative, and hence the (dashed) tangent line will always have negative slope. This slope corresponds to c^-, the spread speed against the advection. That means if $r > 1$, then the population can always invade against the advection. If, on the other hand, $r < 1$, then the tangent line has zero slope if the minimum of $M(s)$ equals $1-r$. If the minimum is smaller than $1-r$, then also the dashed tangent line has positive slope. Since the slope corresponds to c^-, and since the minimum of $M(s)$ is decreasing with increasing advection, we find a switch in the population's ability to invade against the advection. For small values of $v > 0$, the population can invade against the advection; for large values of $v > 0$, the population retreats with the advection.

To compute the critical advection velocity at which the switch happens, we compute the minimum of $M(s)$ as

$$M\left(-\frac{a_1+a_2}{2}\right) = -\frac{4a_1a_2}{(a_1-a_2)^2} > 0. \tag{4.9}$$

Therefore, the critical advection speed is given by

$$1-r = -\frac{4a_1a_2}{(a_1-a_2)^2} \qquad \text{or} \qquad v^2 = 4\frac{r}{1-r}\,\alpha D. \tag{4.10}$$

After some rearranging, we find that (4.10) is exactly the same as (4.7). Hence, the advection velocity above which a population cannot persist in a domain of arbitrary length is exactly the same as the advection velocity at which the population stops spreading upstream and starts retreating downstream. This connection between the two ecologically important quantities critical domain size and invasion speed in systems with advection was first hinted at in [36] and then demonstrated in the context of the PDE system (3.2) in [32].

4.4. Upstream settling probability. The probabilities that, after a dispersal event, an individual settles down- or upstream from its initial location are given by

$$P_{\text{down}} = \int_0^\infty k(x)dx = \frac{a_1}{a_1+|a_2|}, \qquad P_{\text{up}} = 1 - P_{\text{down}}. \tag{4.11}$$

For $r < 1$, we compute a critical upstream-settling probability, below which the population cannot persist or spread against the advection. We insert the critical advection velocity (4.10) into (4.3) and find

$$P^*_{\text{down}} = \frac{1+\sqrt{r}}{2}, \qquad P^*_{\text{up}} = \frac{1-\sqrt{r}}{2} \tag{4.12}$$

as the critical downstream and upstream probabilities, respectively. This result is surprising since the two quantities depend only on the population dynamics parameter and *not* on the movement related parameters α and D. Here lies a chance to test the predictions of the model *without* having to estimate α and D, provided we can estimate P_{up}. Later in the paper (Figure 6.2), we plot the critical domain size as a function of the downstream settling probability and compare it to the case of a fat-tailed kernel.

5. Two modes of dispersal. In the case *without advection*, it is known that the shape of the tail of the dispersal kernel has virtually no influence on the critical domain size [24]. On the other hand, the invasion speed for systems without advection crucially depends on the shape of the tail of the dispersal kernel [19]. Even a tiny fraction of long-distance dispersers can have a huge effect on the invasion speed. In the dispersal model of diffusion and settling, longer dispersal distances result from higher diffusion rate or lower settling rate. In the previous section, we showed that in systems *with advection*, there is a close relationship between critical patch size, critical advection velocity, and invasion speed. In this section, we explore how this relationship depends on the shape of the tail of the kernel.

We assume that individuals have two different dispersal modes and choose between those with probabilities p and $1-p$, respectively. We assume that both dispersal modes can be described by the simple advection-diffusion-settling model (4.1), but with possibly different parameters. Hence, the movement model is given by

$$\begin{aligned} z_{1,t} &= D_1 z_{1,xx} - v_1 z_{1,x} - \alpha_1 z_1, \\ z_{2,t} &= D_2 z_{2,xx} - v_2 z_{2,x} - \alpha_2 z_2 \end{aligned} \tag{5.1}$$

with initial conditions $z_1(0,x) = (1-p)\delta(x)$, $z_2(0,x) = p\delta(x)$. Since there is no interaction between the two different dispersal modes, the resulting kernel is simply the weighted sum of the kernels associated with each mode, i.e.,

$$k = (1-p)k_1 + pk_2, \tag{5.2}$$

where $k_{1,2}$ are given in (4.4) with the appropriate parameters. We are thinking of the z_2-compartment as the long-distance dispersers, i.e., we want k_2 to have fatter tails than k_1, and we assume that p is small. All other parameters being equal, k_2 will have fatter tails than k_1 if either $D_2 > D_1$ or $\alpha_2 < \alpha_1$. The effect of varying $v_{1,2}$ depends on whether we are looking at the upstream or the downstream direction. For simplicity and to compare the results of this section with those of the previous section, we restrict ourselves to the case $v_1 = v_2$.

We first explore the case of varying D_2 at equal settling rates $\alpha_1 = \alpha_2$. In Figure 5.1 we plot the critical domain size as a function of the advection speed for three different values of D_2 and for fixed $p = 0.1$ We also plot the critical advection speed at which the upstream spread is zero as a vertical line. We observe the following. At low advection speeds, the critical domain size is indeed insensitive to changes in D_2; i.e., it does not depend strongly on the tail of the dispersal kernel. The critical domain size increases with increasing D_2, reflecting higher loss at higher diffusion rates. At higher advection speeds, the picture is different. The critical domain size does depend crucially on D_2 and it decreases with increasing D_2. Whereas increasing D_2 increases the loss from the domain downstream, it also increases the probability that a few individuals move upstream. Summarizing in biological terms, at small advection speeds it is important to keep many individuals in the domain; at large advection speeds it is more important to have a few individuals dispersing against the advection. The critical advection speed increases with increasing D_2, which was to be expected since the tails of k get fatter. The curves for the critical domain size approach the straight lines for the critical advection speed for upstream spread, and hence the critical advection speed for persistence and invasion agree, as in the previous section.

Next, we vary the settling rate α_2 at equal diffusion coefficients $D_1 = D_2$. The results are plotted in Figure 5.1, which includes the critical domain size and the

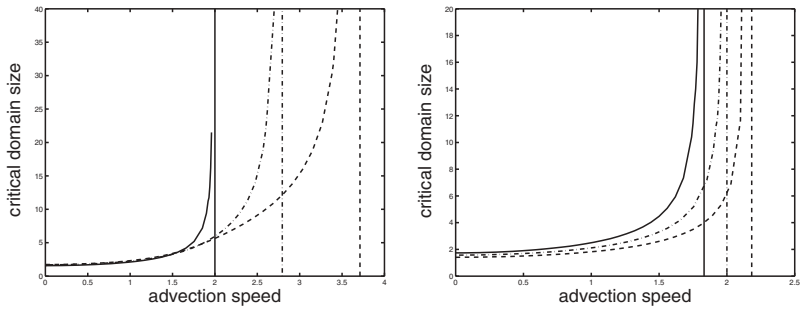


FIG. 5.1. *Left: The critical domain size as a function of the advection speed with dispersal kernel* (5.2). *The parameters are* $D_1 = 1$, $\alpha_1 = \alpha_2 = 1$, $r = 0.5$, $p = 0.1$. *The varying parameter is* $D_2 = 1$ *(solid),* $D_2 = 5$ *(dash-dot), and* $D_2 = 10$ *(dashed). The vertical lines give the critical advection speed for upstream invasion from formula* (3.10). *The values are* $v = 2$, $v = 2.7948$, $v = 3.7114$ *for* $D_2 = 1$, $D_2 = 5$, $D_2 = 10$, *respectively. Right: The critical domain size as a function of the advection speed with dispersal kernel* (5.2). *The parameters are* $D_1 = D_2 = 1$, $\alpha_1 = 1$, $r = 0.5$, $p = 0.1$. *The varying parameter is* $\alpha_2 = 0.1$ *(solid),* $\alpha_2 = 1$ *(dash-dot), and* $\alpha_2 = 10$ *(dashed). The vertical lines give the critical advection speed for upstream invasion from formula* (3.10). *The values are* $v = 1.8321$, $v = 2$, $v = 2.1833$ *for* $\alpha_2 = 0.1$, $\alpha_2 = 1$, $\alpha_2 = 10$, *respectively.*

critical advection speed for upstream spread just as in the previous plot. The two most important observations are that the curves for different α_2 do not intersect and that the curve with the higher α_2 is always the lower one. Hence, independently of the strength of advection, higher settling rate always promotes species persistence and ability to spread upstream. In view of our earlier considerations, this is a surprising result, since decreasing α_2 gives fatter tails of k, yet it reduces the critical advection velocity instead of increasing it as above when we varied D_2.

There are several ways to explain why increasing D_2 and decreasing α_2, which both produce fatter tails of k_2, have opposite effects on the domain length and the invasion speed. Whereas settling rate and diffusion coefficient appear as a quotient in formulas (2.6), (4.3), which determine the tail of the kernel, they appear as a product in formula (4.10) for the critical velocity of upstream propagation. Increasing D_2 in (4.3) decreases both $a_1, |a_2|$ to zero, whereas decreasing α decreases $|a_2|$ to zero and a_1 to v/D. Therefore, increasing D_2 makes the kernel more symmetric, whereas decreasing α_2 makes it less symmetric. This can also be seen by computing the skewness of k from (4.4) as

$$-2\frac{v(v^2+\alpha D)}{(v^2+2\alpha D)\sqrt{v^2+2\alpha D}}, \tag{5.3}$$

which is a decreasing function in the product αD. In more biological terms, in systems with advection, the probability of moving downstream is higher than the probability of moving upstream. Increasing the diffusion rate increases the probability of moving upstream, increasing the settling rate decreases it. Last, dimensional analysis gives the same result. Characteristic length scales are $\sqrt{D/\alpha}$ for a system without advection and v/α for a system without diffusion. The balance between up- and downstream movement is hence given as

$$\sqrt{D/\alpha} \sim v/\alpha \qquad \text{or} \qquad \alpha D \sim v^2. \tag{5.4}$$

6. Dispersal by extremes. In this last section, we explore the ideas from the previous paragraphs in the context of a dispersal kernel whose tails are not expo-

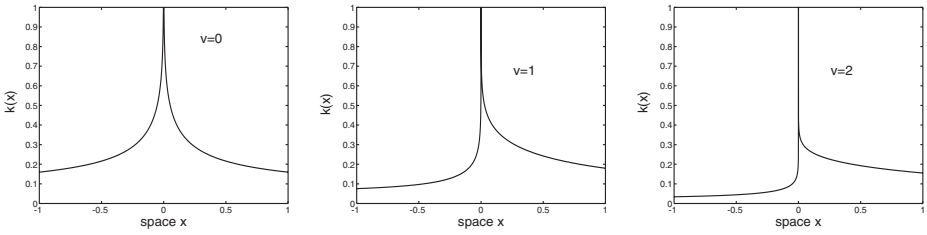


FIG. 6.1. *The fat-tailed kernel from* (6.1) *with parameters* $\mu = 1$, $\alpha = 0.5$ *for different values of advection velocity* v.

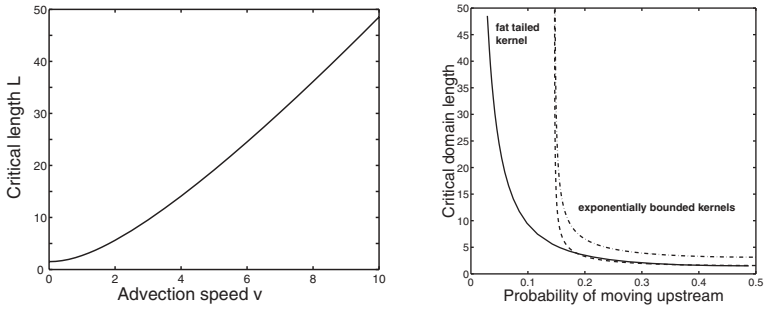


FIG. 6.2. *On the left, the critical domain length for the fat-tailed kernel* (6.1) *is plotted as a function of the advection velocity. The parameters are* $\rho = 1$, $r = 1$, $\alpha = 1$. *On the right, the critical domain length is given as a function of the upstream settling probability. The solid line represents the fat-tailed kernel* (6.1) *with parameters* $\rho = 1$, $r = 1$, $\alpha = 1$. *The dashed and dash-dot line are for the exponential kernel* (4.4) *with parameters are* $\alpha = 1$, $r = 0.5$. *The dashed line corresponds to* $D = 1$, *the dash-dot line to* $D = 4$.

nentially bounded. Such kernels are also known as fat-tailed kernels and describe a situation where long-distance dispersal events are not rare. Different phenomena, such as accelerating invasions, have been shown to occur in that case [19]. We follow the ideas from section 2 to incorporate unidirectional flow in such kernels. Then we numerically investigate how the critical domain size depends on the strength of the flow.

As described in section 2, we compute the appropriate fat-tailed kernel by integrating the Cauchy distribution (2.3) with x replaced by $x - vt$, multiplied with the probability of stopping times (2.5) according to (2.1). This integration yields the asymmetric fat-tailed dispersal kernel

$$k(x) = \frac{\alpha}{(\mu^2 + v^2)\pi} \Re\left((\mu + vi) E_1\left(-\frac{\alpha(v - \mu i)x}{\mu^2 + v^2} \right) \exp\left(-\frac{\alpha(v - \mu i)x}{\mu^2 + v^2} \right) \right). \tag{6.1}$$

In Figure 6.1 we plot this kernel for various values of v. The critical domain length for the fat-tailed kernel (6.1) is plotted as a function of advection speed in Figure 6.2. As expected, it increases with advection speed but it seems to remain finite even for large v. To compare the results for the fat-tailed kernel here with the results from the asymmetric exponential kernel from section 4, we plot the critical domain length in both cases as a function of the probability of settling upstream from the point of release; see section 4.4. If the advection speed is zero, then the probability of settling upstream from the point of release is 0.5. As the advection speed increases, the probability of settling upstream decreases. In the limit as the advection speed approaches

infinity, the upstream probability goes to zero. From the plot in Figure 6.2 we make two observations. The fat-tailed kernel (6.1) produces finite critical domain lengths for smaller upstream probabilities than the exponential kernel; i.e., the population can persist for larger advection speeds. Second, the critical upstream probability for the exponential kernel as computed in (4.12) is independent of the dispersal parameters D and α and depends only on the population growth rate r.

7. Discussion. In this work, we consider integrodifferential models that incorporate population dynamics and individual movement described by dispersal kernels. Extending previous work [30, 32], we consider not only kernels arising from simple random walks but also including (1) unidirectional flow, producing asymmetric kernels, and (2) long-distance jumps (Lévy flight motion), producing fat-tailed kernels. These derivations contribute to the effort to incorporate mechanistic descriptions of individual movement into population models in order to understand the impact of details of individual movement on population dynamics under different conditions.

We obtain general criteria for persistence of a population by deriving the critical domain size for integrodifferential equations. We also extend existing work on the rate of spread [26] and prove that the linear conjecture holds for these systems. Further, we show that in systems with advection there exists a critical advection speed that links population persistence and spread as follows. At a critical advection speed, the population can no longer persist on any finite domain (i.e., the critical domain size is infinite). This critical advection speed is the same as the one that causes upstream propagation to stall (i.e., the upstream propagation speed is 0). We show this result analytically for the modified Laplace kernel and numerically for other kernels; for related results in a PDE model, see [32].

It has been shown that in systems without advection the shape of tails of the dispersal kernel have little effect on persistence [24] but may be a major determinant of the spread rate of a population [19]. Our results show that in systems with advection, the shape of the tails of kernels influences both. With fat-tailed kernels, a population is able to both persist and spread upstream in conditions with higher flow speed.

Whereas the current model gave us valuable insight in dispersal in stream populations and possible explanations for the drift paradox, we plan to continue these investigations using more realistic biological models. The techniques in this paper will be extended to cover, e.g., resource dynamics and predator-prey interaction. Most important, we plan to model a population of larvae and adult stage where adults emerge from the stream and fly upstream to deposit eggs. This mechanism is the most commonly quoted biological hypothesis to solve the drift paradox. Finally, as we are dealing with low population numbers, we intend to compare the results of these deterministic models here to stochastic simulations.

Appendix A. Derivation of the Cauchy distribution for individuals undergoing a random walk. The derivation we use follows [11]. Let Y be a random variable, assuming its values on the integer lattice and describing the number of space steps that an individual jumps each time step. The probability that the individual jumps k steps to the right ($\Pr(Y = k) = p_k$) is defined to be

$$p_k = \begin{cases} 1 - \frac{2m}{\pi} & \text{if} \quad k = 0, \\ \frac{m}{\pi|k|(|k|+1)} & \text{if} \quad k > 0. \end{cases} \tag{A.1}$$

The parameter m, restricted to $0 < m < \pi/2$, describes the likelihood of dispersing. It is straightforward to show that the p_k sum to one. We produce a random walk on

the grid of spacing h with $h > 0$ by letting the walker start at point 0 at instant 0 and defining the location of the individual after n time steps to be

$$X_n = hY_1 + hY_2 + \cdots + hY_n, \tag{A.2}$$

where Y_n are independent, identically distributed random variables, all having the same distribution as Y. We relate space steps h and time steps τ by the speed s, so $h = s\tau$. At time $t = n\tau = nh/s$ we have $h = ts/n$, so that the spreading time associated with distance X_n is to

$$\frac{X_n}{\rho} = \frac{t}{mn}\left(Y_1 + Y_2 + \cdots + Y_n\right), \tag{A.3}$$

where $\rho = ms$ is a speed of spreading. The distribution of the right-hand side of (A.3) converges to the distribution of normalized Cauchy distribution

$$\frac{1}{\pi}\frac{t}{x^2 + t^2} \tag{A.4}$$

in the limit as n approaches infinity [11]. Thus X_n approaches (2.3) in the same limit. For any given fixed time t and speed s the limit n approaches infinity is equivalent to the space step h approaching zero.

Appendix B. Dispersal kernels for multiple dispersal modes. In extending the simple diffusion model for individual movement, we assume that individuals have two different modes of dispersal and that they can switch between these modes. We show how the dispersal kernel can be computed explicitly for constant rates and that the kernel is exponentially bounded. The description for individual movement is given by

$$\begin{aligned} z_{1,t} &= D_1 z_{1,xx} - v_1 z_{1,x} - \mu z_1 + \sigma z_2 - \alpha_1 z_1, \\ z_{2,t} &= D_2 z_{2,xx} - v_2 z_{2,x} - \mu z_2 + \sigma z_1 - \alpha_2 z_2. \end{aligned} \tag{B.1}$$

The parameters D_j, v_j, and α_j are the diffusion rates, the advection speeds, and the settling rates for the different stages. The parameters μ and σ are switching rates between the stages. Initially, there is a certain fraction of the population in each stage, i.e., $z_1(0,x) = p\delta(x)$, and $z_2(0,x) = (1-p)\delta(x)$. The density of stopping points from the respective stages is given by

$$k_j(x) = \int_0^\infty \alpha_j z_j(t,x) \tag{B.2}$$

for $j = 1, 2$, and hence the kernel is given by $k(x) = k_1(x) + k_2(x)$.

The case $\alpha_1 = \sigma = 0$ can be interpreted as two successive modes of dispersal. In the case without advection, this has been treated by [30]. If we consider only movement, not settling ($\alpha_{1,2} = 0$), then we can study the shape of the spatial distribution of z_1 and z_2 as it evolves in time. For systems like (B.1) but without advection, Skalski and Gilliam [33] have constructed an explicit solution and computed asymptotic speeds of spread for a linear model.

From (B.1) we deduce that k_1, k_2 satisfy the system

$$\begin{aligned} -p\alpha_1\delta &= D_1 k_1^{(2)} - v_1 k_1^{(1)} - (\mu + \alpha_1)k_1 + \sigma\frac{\alpha_1}{\alpha_2}k_2, \\ -(1-p)\alpha_2\delta &= D_2 k_2^{(2)} - v_2 k_2^{(1)} - (\sigma + \alpha_2)k_2 + \mu\frac{\alpha_2}{\alpha_1}k_1, \end{aligned} \tag{B.3}$$

where $k_j^{(l)}$ denotes the lth derivative of k_j. Restriction to the interval $(0,\infty)$ and repeated differentiation and substitution of (B.3) yields a fourth-order equation for k_1 as follows:

$$D_1 k_1^{(4)} - \left(v_1 + v_2\frac{D_1}{D_2}\right)k_1^{(3)} - \left(\mu + \alpha_1 - \frac{v_1 v_2 D_1(\sigma+\alpha_2)}{D_2}\right)k_1^{(2)}$$

(B.4)

$$+ \frac{v_1(\sigma+\alpha_2)+v_2(\sigma+\alpha_1)}{D_2}k_1^{(1)} + \frac{(\sigma+\alpha_2)(\mu+\alpha_2)-\mu\sigma}{D_2}k_1 = 0.$$

This is a linear equation with constant coefficients; therefore the solution is readily determined and is exponentially bounded. The coefficients are determined by the usual conditions, i.e., the kernel has to integrate to unity, it has to be continuous at zero, and the jump conditions at zero have to be satisfied.

Appendix C. Proof of Theorem 3.1. The exponential ansatz $u(t,x) = \exp(\lambda t)\phi(x)$ in the linearization (3.6) leads to the eigenvalue problem

$$\nu\phi(x) = I[\phi](x) = \int_0^L k(x,y)\phi(y)dy \tag{C.1}$$

with $\nu = \lambda + 1 - r$. The solution u of (3.6) will grow if $\lambda > 0$ and decay if $\lambda < 0$. Hence, the critical value is given by $\lambda = 0$ or $\nu = 1 - r$.

We now show that the dominant eigenvalue ν^* is a monotone increasing function of domain length. For two domain lengths $L_2 > L_1$, we denote I_j as the linear operator given by (C.1) with L replaced by L_j, $j = 1, 2$. We denote $\nu_{1,2}$ as the corresponding dominant eigenvalues and $\phi_{1,2}$ as corresponding (positive) eigenfunctions. Then $I_2 \geq I_1$ and hence $\nu_2 \geq \nu_1$. We show that the inequality is in fact strict. We write

$$I_2\phi_2 = I_1\phi_2 + \int_{L_1}^{L_2} k(x,y)\phi_2(y)dy.$$

Since $\phi_2 > 0$, the last term is positive and hence there is an $\varepsilon > 0$ such that

$$f\colon = \int_{L_1}^{L_2} k(x,y)\phi_2(y)dy > \varepsilon\phi_1.$$

Then the equation $\nu_2\psi = I_1\psi + f$ has no solution for $\nu_2 \leq \nu_1$ [21]. But ϕ_2 is a solution and hence necessarily $\nu_2 > \nu_1$.

If the dispersal kernel is continuous, then the resulting integral operator on $\mathcal{L}^2[0,L]$ is completely continuous and, for positive kernel, has a unique simple dominant eigenvalue [21]. Therefore, our assumptions are valid for all kernels in sections 4 and 5. In fact, the condition that the kernel be continuous can be weakened by saying that the kernel to the power $1+q$, $q \geq 1$, has to be integrable on $[0,L]^2$ [21]. Numerically, the fat-tailed kernel (2.7) can be bounded by $x^{-0.4}$, which is square integrable, and hence the assumption holds. This is an area of future research.

Appendix D. Proof of Theorem 3.2. By scaling time, we may assume $\mu = 1$ in equation (3.1). It was shown in [26] that $c^- \leq c_-^*$ and $c_+^* \leq c^+$. To show the reversed inequalities, we follow Aronson's proof [2] and show that for all $c \in (c^-, c^+)$ there is a subsolution of (3.1) which expands at speed c. Due to a comparison principle, the true solution has to expand at speed at least c.

We make the following technical requirements on the kernel k [2]. We assume $k \geq 0$ and $\text{supp}(k) = \mathbb{R}$. We assume that the moment generating function $M(s)$ exists for $s \in (\hat{s}^-, \hat{s}^+)$ with $\hat{s}^- < 0$, $\hat{s}^+ > 0$. We assume furthermore that the function

$$A_\lambda(s) = [(M(s) + \lambda)/s], \qquad s \neq 0, \tag{D.1}$$

has exactly one minimum at $\bar{s}^+ \in (0, \hat{s}^+)$ and one maximum at $\bar{s}^- \in (\hat{s}^-, 0)$. In addition, $A_\lambda(s)$ is increasing on $(\hat{s}^-, \bar{s}^-) \cup (\bar{s}^+, \hat{s}^+)$ and decreasing on $(\bar{s}^-, 0) \cup (0, \bar{s}^+)$. Finally, we assume that the function $x \mapsto \exp(sx)k(x)$ is decreasing for large enough x. Note that with this notation, $c^\pm = A_\lambda(\bar{s}^\pm)$ with $\lambda = f'(0) - 1$.

We first switch to a moving coordinate frame and show a comparison principle for the resulting integrodifferential operator. The function $W(t, \xi) = u(t, \xi + ct)$ satisfies

$$W_t = cW_\xi + f(W) - W + \int k(\xi - \eta)W(t, \eta)d\eta =: Q_c[W], \quad W(0, \xi) = u(0, \xi). \tag{D.2}$$

LEMMA D.1 (comparison). *Let V, W be bounded and continuously differentiable functions which satisfy, on $\mathbb{R}^+ \times \mathbb{R}$,*

$$V_t - Q_c[V] \geq W_t - Q_c[W], \tag{D.3}$$

and $V(0, \xi) > W(0, \xi)$ on $\mathbb{R}$. Then $V > W$ on $(0, \infty) \times \mathbb{R}$.

Proof. Let V, W be given. The difference $Z = V - W$ satisfies

$$Z_t - cZ_\xi \geq h(t, \xi)Z + k * Z, \quad Z(0, \cdot) > 0, \tag{D.4}$$

where h is some bounded function, given by the mean value theorem. Suppose there is a first time t_0 such that $Z > 0$ on $[0, t_0) \times \mathbb{R}$, and $Z(t_0, \xi_0) = 0$ for some ξ_0. By assumption, the convolution term in (D.4) is nonnegative on $[0, t_0] \times \mathbb{R}$. Therefore, along the characteristic lines $\xi + ct$, Z is bounded below by the solution of the differential equation $\dot{\zeta} = h\zeta$, $\zeta(0, \xi) = Z(0, \xi) > 0$. Since ζ remains positive, Z has to remain positive. ∎

In the following, we will use Lemma D.1 with nonstrict inequalities, i.e., $V(0, \xi) \geq W(0, \xi)$ implies $V \geq W$ provided V satisfies a well-posed initial value problem, and still refer to that as Lemma D.1. The idea is the same as in [2]. Let V_ϵ, $\epsilon > 0$, be the solution of a well-posed initial value problem with initial value $V_\epsilon(0, \xi) = V(0, \xi) + \epsilon$. Then by the above $V_\epsilon > W$, and in the limit $\epsilon \to 0$, we have $V \geq W$.

LEMMA D.2 (subsolution). *Let $c \in (c^-, c^+)$ be given. Then there exists a function $V_0(\xi)$, which is positive on $(0, \pi/\gamma)$, such that $Q_c[\varepsilon V_0] \geq 0$ and*

$$Q_c[\varepsilon V_0] > 0 \quad \text{on} \quad (0, \pi/\gamma) \tag{D.5}$$

for all sufficiently small $\varepsilon, \gamma > 0$.

Before we prove Lemma D.2, we demonstrate how the subsolution and repeated use of the comparison principle are employed to prove the theorem. Suppose that $W(0, \xi)$ and $c \in (c^-, c^+)$ are given and $W(t, \xi)$ satisfies (D.2). We need to show that $W(t, \xi) \to \bar{u}$ as $t \to \infty$ for all $\xi \in \mathbb{R}$. At first, Lemma D.2 ensures the existence of $V_0(\xi)$, which is positive on $(0, \pi/\gamma)$ for small enough $\gamma > 0$. We apply the comparison principle to εV_0 and V, defined as the solution to

$$V_t = Q_c[V], \qquad V(0, \xi) = \varepsilon V_0(\xi), \tag{D.6}$$

to see that $V(t,\xi) \geq \varepsilon V_0(\xi)$ for all $t > 0$. Next, the comparison principle is applied to $V(t,\xi)$ and $\tilde{V}(t,\xi) = V(t+h,\xi)$ for any fixed $h > 0$. As a result, $\tilde{V} \geq V$ and therefore $V(t,\xi)$ is a nondecreasing function in t for each fixed ξ. On comparing $V(t,\xi)$ with the constant $\bar{u}$, we get that V is bounded by $\bar{u}$, and therefore $V(t,\xi) \to q(\xi)$ for each ξ. Following Aronson [2], one can actually show that $q(\xi) \equiv \bar{u}$.

Finally, for T sufficiently large, there is a bound $m > 0$ such that $W(T,\xi) \geq m > 0$ on $(0,\pi/\gamma)$. We choose $\varepsilon > 0$ such that $\varepsilon V_0 < m$. We now apply the comparison principle to $W(t,\xi)$ and the solution $V(t-T,\xi)$ of $V_t = Q_c[V]$, $V(T,\xi) = \varepsilon V_0(\xi)$, to obtain that $W(t,\xi) \geq V(t-T,\xi)$. This completes the proof.

We now prove Lemma D.2. We first look at the linear equation

$$W_t = L_c[W] \colon = cW_\xi + \lambda W + k * W, \tag{D.7}$$

where $*$ denotes the convolution. For $s \in (\bar{s}^-, \bar{s}^+)\backslash\{0\}$, we define

$$\hat{V}_0(\xi) = e^{-s\xi} \sin \gamma\xi. \tag{D.8}$$

After a little bit of algebra, we find that $L_c[\hat{V}_0](\xi)$ is given by

$$\left[-cs + \lambda + \int e^{s\eta} k(\eta) \cos(\gamma\eta) d\eta\right] \hat{V}_0 + \left[c\gamma - \int e^{s\eta} k(\eta) \sin(\gamma\eta) d\eta\right] e^{-s\xi} \cos \gamma\xi.$$

Therefore, $L_c[\hat{V}_0] > 0$ on $(0,\pi/\gamma)$ if the following two conditions are satisfied:

$$c < \frac{1}{s}\left[\lambda + \int e^{s\eta} k(\eta) \cos(\gamma\eta) d\eta\right] =: \mathcal{A}_\lambda(s,\gamma), \qquad s > 0, \tag{D.9}$$

$$c > \mathcal{A}_\lambda(s,\gamma), \qquad s < 0, \tag{D.10}$$

$$c = \frac{1}{\gamma}\left[\int e^{s\eta} k(\eta) \sin(\gamma\eta) d\eta\right] =: \mathcal{B}(s,\gamma). \tag{D.11}$$

We first establish some properties of the functions $\mathcal{A}_\lambda$ and $\mathcal{B}$. As $\gamma \to 0$, we have uniform convergence on compact subsets of $(\bar{s}^-, \bar{s}^+)\backslash\{0\}$ of

$$\mathcal{A}_\lambda(s,\gamma) \to A_\lambda(s), \quad \mathcal{B}(s,\gamma) \to B(s) \colon = \int \eta e^{s\eta} k(\eta) d\eta.$$

The function $B(s)$ is increasing. Differentiation gives $A'_\lambda(s) = (B(s) - A_\lambda(s))/s$. Hence, due to the assumptions on A_λ, we furthermore see that $B < A_\lambda$ on $(0, \bar{s}^+)$, $B > A_\lambda$ on $(\bar{s}^-, 0)$, and $B(\bar{s}^\pm) = A_\lambda(\bar{s}^\pm)$. Note that $B(0)$ is the average dispersal distance, and since B is an increasing function, $c^- < B(0) < c^+$; i.e., the interval (c^-, c^+) is never empty.

We now return to the construction of $\hat{V}_0$; i.e., we show that conditions (D.9)–(D.11) can be satisfied simultaneously. Without loss of generality, we may assume $c > B(0)$, and hence we restrict ourselves to $s > 0$. First, we can choose $\lambda < f'(0) - 1$ such that $c < A_\lambda(\bar{s}^+)$. Then we can choose $s_0, s_1, \delta, \gamma > 0$, such that

$$B(s_0) + \delta < c < B(s_1) - \delta \quad \text{and} \quad |\mathcal{B}(s,\gamma) - B(s)| < \delta.$$

By continuity, there is a value $s(\gamma)$ such that $\mathcal{B}(s(\gamma),\gamma) = c$ for all sufficiently small γ. Obviously, we can choose γ small enough such that $\mathcal{A}_\lambda(s(\gamma),\gamma) > c$. Hence, the two conditions (D.9), (D.11) can be satisfied simultaneously.

By the same argument as [2], one can show that the modified function

$$V_0(\xi) = \hat{V}_0(\xi), \quad \xi \in [0, \pi/\gamma], \quad V_0(\xi) = 0, \quad \xi > \pi/\gamma, \tag{D.12}$$

also satisfies $L_c[V_0] > 0$ on $(0, \pi/\gamma)$.

As a last step, we have to show that for small enough $\varepsilon > 0$ we have $Q_c[\varepsilon V_0] > 0$ on that same interval. Note that $\lambda < f'(0) - 1$ implies that $\lambda\varepsilon < f(\varepsilon) - \varepsilon$ for small enough $\varepsilon > 0$. Hence, we have $Q_c[\varepsilon V_0] > L_c[\varepsilon V_0] > 0$, on $(0, \pi/\gamma)$, which completes the proof.

Appendix E. The advection diffusion kernel for bounded domains. Movement is modeled by (4.1) on the interval $[-L/2, L/2]$ with initial condition $z(0, x) = \delta(x - y)$. The boundary conditions are

$$\left(z_x - \frac{v}{D} z\right)(t, -L/2) = 0, \qquad z(t, L/2) = 0. \tag{E.1}$$

We interpret these conditions as a stream where individuals cannot enter or leave at the upstream end and are washed out at the downstream end [36]. We nondimensionalize (4.1) by setting $X = x/L$, $T = \alpha t$, $Z = Lz$, which gives

$$Z_T = \frac{1}{\tilde{L}^2} Z_{XX} - \tilde{v} Z - Z, \quad Z(T, 1/2) = 0 = (Z_X - \tilde{L}^2 \tilde{v} Z)(T, -1/2), \tag{E.2}$$

where $\tilde{L}^2 = \alpha L/D$ and $\tilde{v} = v/(\alpha L)$. For convenience, we write the variables t, x in lower case letters again. We want to find the nondimensional kernel given by $\kappa(x, y) = \int_0^\infty \alpha Z(t, x) dt$. The function $W(t, x) = \exp(-L^2 v x/2) Z(t, x)$ satisfies

$$W_t = \frac{1}{L^2} W_{xx} - \beta W, \tag{E.3}$$

where $\beta = 1 + \frac{v^2 L^2}{4}$. Separating variables $W(t, x) = T(t)X(x)$, we get the two independent equations $T' = -(\lambda^2 + \beta)T$ and $X'' = -\lambda^2 L^2 X$ for some $\lambda^2 > 0$. The boundary conditions applied to the equations for X result in the defining condition

$$\lambda = -\frac{vL}{2} \tan(\lambda L). \tag{E.4}$$

We denote its infinitely many (symmetric) nonzero solutions by λ_n, $n = 1, 2, \ldots$. The corresponding family of orthogonal solutions is given by

$$\phi_n(x) = -\tan(\lambda_n L/2) \cos(\lambda_n L x) + \sin(\lambda_n L x) \tag{E.5}$$

with norm

$$\|\phi_n\|_2^2 = \frac{1}{2}(1 + \tan^2(\lambda_n L/2)).$$

The solution of (E.3) can hence be written as an infinite sum where each term is of the form $c_n e^{-(\lambda_n + \beta)t} \phi_n(x)$. To find expressions for the coefficients c_n we approximate the delta distribution by the top hat function,

$$\delta_m(x - y) = \begin{cases} 2m, & |x - y| \le 1/m, \\ 0, & \text{else.} \end{cases}$$

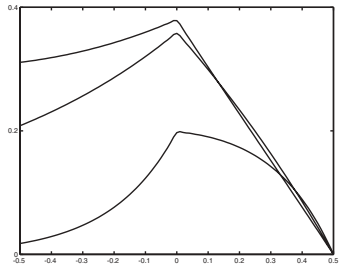


FIG. E.1. *The kernel with advection, no-flux boundary conditions at the left end, zero boundary conditions at the right end, release point in the middle, for* $v = 0.2$ *(top curve),* $v = 1$ *(middle), and* $v = 5$ *(bottom).*

Expanding the approximate initial condition

$$\delta_m(x-y)e^{-\frac{vL^2}{2}x} = \sum c_n\phi_n(x)$$

and using the intermediate value theorem gives

$$c_n = \frac{e^{-\frac{vL^2}{2}y}\phi_n(y)}{\|\phi_n\|_2^2}.$$

Hence, the nondimensionalized kernel is given by

$$k(x,y) = e^{-\frac{vL^2}{2}(y-x)} \sum_n \frac{1}{\lambda_n^2 + 1 + \frac{v^2L^2}{4}} \frac{2}{1+\tan^2(\frac{\lambda_n L}{2})} S(x)S(y), \tag{E.6}$$

where $S(x) = (\sin(\lambda_n Lx) - \tan(\lambda_n L/2)\cos(\lambda_n Lx))$. In Figure E.1 we plot this kernel for three different advection speeds.

Appendix F. Exact derivation of the critical domain length. We determine the critical domain length for the kernel (4.4) by computing the eigenvalue of the corresponding integral operator (C.1), extending earlier work [20, 40]. Scaling the space variable by L gives

$$\nu\phi(x) = \int_0^1 \tilde{\kappa}(x,y)\phi(y)dy, \tag{F.1}$$

where $\tilde{\kappa}$ is defined as κ with a_j, A replaced by $b_j = La_j$, $B = LA$. Differentiating (F.1) gives

$$\nu\phi'(x) = b_2\nu\phi(x) + (b_1 - b_2)\int_x^1 Be^{b_1(x-y)}\phi(y)dy. \tag{F.2}$$

Differentiating again, we obtain

$$\nu\phi''(x) = (b_2 - b_1)B\phi(x) + b_2^2\nu\phi(x) + (b_1^2 - b_2^2)\int_x^1 Be^{b_1(x-y)}\phi(y)dy. \tag{F.3}$$

Substituting (F.2) into (F.3), we get the regular Sturm–Liouville problem

$$\phi''(x) = -b_1|b_2|\left(\frac{1}{\nu} - 1\right)\phi(x) + (b_1 + b_2)\phi'(x), \quad \phi'(0) = b_1\phi(0), \quad \phi'(1) = b_2\phi(1). \tag{F.4}$$

We apply the transformation $\psi(x) = \exp(-\frac{b_1+b_2}{2}x)\phi(x)$ to (F.4) and substitute the original parameters back to obtain

$$\psi'' = -L^2\frac{(a_1-a_2)^2}{4}\left(\frac{4a_1|a_2|}{\nu(a_1-a_2)^2}-1\right)\psi, \tag{F.5}$$

together with the boundary conditions

$$\psi'(0) = L\frac{a_1-a_2}{2}\psi(0) \qquad \text{and} \qquad \psi'(1) = -L\frac{a_1-a_2}{2}\psi(1). \tag{F.6}$$

Equations (F.5) and (F.6) constitute a Sturm–Liouville problem, which one can solve for L as a function of ν [40], and the solution is given by formula (4.6).

Acknowledgment. The authors thank Roger Nisbet for insightful discussions and helpful comments.

REFERENCES

[1] J. Allan, *Stream Ecology: Structure and Function of Running Waters*, Chapman & Hall, London, 1995.
[2] D. Aronson, *The asymptotic speed of propagation of a simple epidemic*, in Nonlinear Diffusion, W. Fitzgibbon and H. Walker, eds., Res. Notes Math. 14, Pitman, London, 1977, pp. 1–23.
[3] D. Aronson and H. Weinberger, *Nonlinear diffusion in population genetics, combustion, and nerve pulse propagation*, in Partial Differential Equations and Related Topics, J. Goldstein, ed., Lecture Notes in Math. 446, Springer-Verlag, Berlin, 1975, pp. 5–49.
[4] R. Baker and P. Dunn, *New Directions in Biological Control*, Alan Liss, New York, 1990.
[5] M. Ballyk and H. Smith, *A model of microbial growth in a plug flow reactor with wall attachment*, Math. Biosci., 158 (1999), pp. 95–126.
[6] L. W. Botsford, A. Hastings, and S. D. Gaines, *Dependence of sustainability on the configuration of marine reserves and larval dispersal distance*, Ecology Lett., 4 (2001), pp. 144–150.
[7] S. R. Broadbent and D. G. Kendall, *The random walk of* Trichostrongylus retortaeformis, Biometrica, 9 (1953), pp. 460–466.
[8] R. S. Cantrell and C. Cosner, *Should a park be an island?*, SIAM J. Appl. Math., 53 (1993), pp. 219–252.
[9] R. S. Cantrell and C. Cosner, *On the effects of spatial heterogeneity on the persistence of interacting species*, J. Math. Biol., 37 (1998), pp. 103–145.
[10] D. Carrero, G. McDonald, E. Crawford, G. de Vries, and M. Hendzel, *Using FRAP and mathematical modeling to determine the in vivo kinetics of nuclear proteins*, Methods, 29 (2003), pp. 14–28.
[11] R. Gorenflo and F. Mainardi, *Approximation of Lévy-Feller diffusion by random walk*, Z. Anal. Anwendungen, 18 (1999), pp. 231–246.
[12] K. Hadeler, *Reaction transport systems*, in Mathematics Inspired by Biology, CIME Lectures 1997, Florence, V. Capasso and O. Diekmann, eds., Springer-Verlag, New York, 1998, pp. 95–150.
[13] K. Hadeler and M. Lewis, *Spatial dynamics of the diffusive logistic equation with sedentary component*, Canad. Appl. Math. Quart., 10 (2004), pp. 473–500.
[14] T. Hillen, *Transport equations with resting phase*, European J. Appl. Math., 14 (2003), pp. 613–636.
[15] E. Holmes, M. Lewis, J. Banks, and R. Veit, *Partial differential equations in ecology: Spatial interactions and population dynamics*, Ecology, 75 (1994), pp. 117–129.
[16] V. Hutson, S. Martinez, K. Kischaikow, and G. Vickers, *The evolution of dispersal*, J. Math. Biol., 46 (2003), pp. 483–517.
[17] H. Kierstead and L. B. Slobodkin, *The size of water masses containing plankton blooms*, J. Marine Res., 12 (1953), pp. 141–147.
[18] A. Kolmogorov, I. Petrovskii, and N. Piskunov, *A study of the equation of diffusion with increase in the quantity of matter, and its application to a biological problem*, Bjol. Moskovskovo Gos. Univ., 17 (1937), pp. 1–72.
[19] M. Kot, M. Lewis, and P. van den Driessche, *Dispersal data and the spread of invading organisms*, Ecology, 77 (1996), pp. 2027–2042.

[20] M. Kot and W. M. Schaffer, *Discrete-time growth-dispersal models*, Math. Biosci., 80 (1986), pp. 109–136.
[21] M. A. Krasnosel'skii, *Positive Solutions of Operator Equations*, Noordhoff, Groningen, The Netherlands, 1964.
[22] J. Lancaster and A. Hildrew, *Characterising instream flow refugia*, Canad. J. Fish. Aquat. Sci., 50 (1993), pp. 1663–1675.
[23] M. Lewis and G. Schmitz, *Biological invasion of an organism with separate mobile and stationary states: Modeling and analysis*, Forma, 11 (1996), pp. 1–25.
[24] D. Lockwood, A. Hastings, and L. Botsford, *The effects of dispersal patterns on marine reserve: Does the tail wag the dog?*, Theor. Popul. Biol., 61 (2002), pp. 297–309.
[25] F. Lutscher and M. A. Lewis, *Spatially-explicit matrix models. A mathematical analysis of stage-structured integrodifference equations*, J. Math. Biol., 48 (2004), pp. 293–324.
[26] J. Medlock and M. Kot, *Spreading diseases: Integro-differential equations new and old*, Math. Biosci., 184 (2003), pp. 201–222.
[27] K. Müller, *Investigations on the Organic Drift in North Swedish Streams*, Tech. report 34, Institute of Freshwater Research, Drottningholm, Sweden, 1954.
[28] K. Müller, *The colonization cycle of freshwater insects*, Oecologica, 53 (1982), pp. 202–207.
[29] J. Murray and R. Sperb, *Minimum domains for spatial patterns in a class of reaction diffusion equations*, J. Math. Biol., 18 (1983), pp. 169–184.
[30] M. G. Neubert, M. Kot, and M. A. Lewis, *Dispersal and pattern formation in a discrete-time predator-prey model*, Theor. Pop. Biol., 48 (1995), pp. 7–43.
[31] H. Othmer, S. Dunbar, and W. Alt, *Models of dispersal in biological systems*, J. Math. Biol., 26 (1988), pp. 263–298.
[32] E. Pachepsky, F. Lutscher, R. Nisbet, and M. A. Lewis, *Persistence, spread and the drift paradox*, Theor. Pop. Biol., 67 (2005), pp. 61–73.
[33] G. Skalski and J. Gilliam, *A diffusion-based theory of organism dispersal in heterogeneous populations*, Amer. Nat., 161 (2003), pp. 441–458.
[34] J. G. Skellam, *Random dispersal in theoretical populations*, Biometrika, 38 (1951), pp. 196–218.
[35] H. Smith and P. Waltman, *The Theory of the Chemostat*, Cambridge University Press, Cambridge, UK, 1995.
[36] D. Speirs and W. Gurney, *Population persistence in rivers and estuaries*, Ecology, 82 (2001), pp. 1219–1237.
[37] R. Syski, *Random Processes*, Marcel Dekker, New York, 1989.
[38] H. Thieme and X.-Q. Zhao, *Asymptotic spreads of speed and traveling waves for integral equations and delayed reaction-diffusion models*, J. Differential Equations, 195 (2003), pp. 430–470.
[39] P. Turchin, *Quantitative Analysis of Movement*, Sinauer, Sunderland, MS, 1998.
[40] R. W. Van Kirk and M. A. Lewis, *Integrodifference models for persistence in fragmented habitats*, Bull. Math. Biol., 59 (1997), pp. 107–137.
[41] R. Waters, *The drift of stream insects*, Ann. Rev. Entomol., 17 (1972), pp. 253–272.

SIAM J. APPL. MATH.
Vol. 65, No. 4, pp. 1328–1352

A MATHEMATICAL STUDY OF THE HEMATOPOIESIS PROCESS WITH APPLICATIONS TO CHRONIC MYELOGENOUS LEUKEMIA*

MOSTAFA ADIMY†, FABIEN CRAUSTE†, AND SHIGUI RUAN‡

Abstract. This paper is devoted to the analysis of a mathematical model of blood cell production in the bone marrow (hematopoiesis). The model is a system of two age-structured partial differential equations. Integrating these equations over the age, we obtain a system of two nonlinear differential equations with distributed time delay corresponding to the cell cycle duration. This system describes the evolution of the total cell populations. By constructing a Lyapunov functional, it is shown that the trivial equilibrium is globally asymptotically stable if it is the only equilibrium. It is also shown that the nontrivial equilibrium, the most biologically meaningful one, can become unstable via a Hopf bifurcation. Numerical simulations are carried out to illustrate the analytical results. The study may be helpful in understanding the connection between the relatively short cell cycle durations and the relatively long periods of peripheral cell oscillations in some periodic hematological diseases.

Key words. blood cells, hematopoiesis, differential equations, distributed delay, asymptotic stability, Lyapunov functional, Hopf bifurcation

AMS subject classifications. 34K20, 92C37, 34C23, 34D20, 34K99

DOI. 10.1137/040604698

1. Introduction.

Cellular population models have been investigated intensively since the 1960s (see, for example, Trucco [33, 34], Nooney [25], Rubinow [28], and Rubinow and Lebowitz [29]) and still interest a lot of researchers. This interest is greatly motivated, on one hand, by medical applications and, on the other hand, by the biological phenomena (such as oscillations, bifurcations, traveling waves, or chaos) observed in these models and, generally speaking, in the living world (Mackey and Glass [19], Mackey and Milton [20]).

Hematopoiesis is the process by which primitive stem cells proliferate and differentiate to produce mature blood cells. It is driven by highly coordinated patterns of gene expression under the influence of growth factors and hormones. The regulation of hematopoiesis is about the formation of blood cell elements in the body. White and red blood cells and platelets are produced in the bone marrow, from where they enter the blood stream. The principal factor stimulating red blood cell production is a hormone produced in the kidney, called erythropoietin. About 90% of the erythropoietin is secreted by renal tubular epithelial cells when blood is unable to deliver sufficient oxygen. A decrease in the level of oxygen in the blood leads to a release of a substance, which in turn causes an increase in the release of the blood elements from the marrow. There is feedback from the blood to the bone marrow. Abnormalities in the feedback are considered as major suspects in causing periodic hematological diseases, such as autoimmune hemolytic anemia (Bélair, Mackey, and Mahaffy [4] and Mahaffy, Bélair, and Mackey [23]), cyclical neutropenia (Haurie, Dale, and Mackey

*Received by the editors March 2, 2004; accepted for publication (in revised form) October 18, 2004; published electronically April 26, 2005.
http://www.siam.org/journals/siap/65-4/60469.html
†Laboratoire de Mathématiques Appliquées, FRE 2570, Université de Pau et des Pays de l'Adour, Avenue de l'université, 64000 Pau, France (mostafa.adimy@univ-pau.fr, fabien.crauste@univ-pau.fr).
‡Department of Mathematics, University of Miami, Coral Gables, FL 33124-4250 (ruan@math.miami.edu). The research of this author was partially supported by the National Science Foundation and the College of Arts and Sciences at the University of Miami.

[14]), and chronic myelogenous leukemia (Fowler and Mackey [12] and Pujo-Menjouet, Bernard, and Mackey [26]).

Cell biologists classify stem cells as proliferating cells and resting cells (also called G_0-cells) (see Mackey [16, 17]). Proliferating cells are committed to undergo mitosis a certain time after their entrance into the proliferating phase. Mackey supposed that this time of cytokinesis is constant, that is, it is the same for all cells. Most of committed stem cells are in the proliferating phase. The G_0-phase, whose existence is known due to the works of Burns and Tannock [8], is a quiescent stage in the cellular development. However, it is usually believed that 95% of pluripotent stem cells are in the resting phase. Resting cells can exit randomly to either enter into the proliferating phase or be irremediably lost. Proliferating cells can also be lost by apoptosis (programmed cell death).

The model of Mackey [16] has been numerically studied by Mackey and Rey [21] and Crabb, Losson, and Mackey [9]. Computer simulations showed strange behaviors of the stem cell population, such as oscillations and bifurcations. Recently, Pujo-Menjouet and Mackey [27] proved the existence of a Hopf bifurcation which causes periodic chronic myelogenous leukemia and showed the great dependence of the model on the parameters.

In this paper, based on the model of Mackey [16], we propose a more general model of hematopoiesis. We take into account the fact that a cell cycle has two phases, that is, stem cells in process are either in a resting phase or actively proliferating. However, we do not suppose that all cells divide at the same age, because this hypothesis is not biologically reasonable. For example, it is believed that pluripotent stem cells divide faster than committed stem cells, which are more mature cells. There is strong evidence (see Bradford et al. [7]) that indicate that the age of cytokinesis τ is distributed on an interval $[\underline{\tau}, \overline{\tau}]$ with $\underline{\tau} \geq 0$. Hence, we shall assume that τ is distributed with a density f supported on an interval $[\underline{\tau}, \overline{\tau}]$ with $0 \leq \underline{\tau} < \overline{\tau} < +\infty$. The resulting model is a system of two differential equations with distributed delay. A simpler model, dealing with the pluripotent stem cell population behavior, has been studied by Adimy, Crauste, and Ruan [1].

Some results about stability of differential equations with distributed delay can be mentioned. In [6], Boese studied the stability of a differential equation with gamma-distributed delay. Gamma distributions have the property to simplify the nature of the delay and this situation is close to the one with discrete delay. Anderson [2, 3] showed stability results linked to the different moments (especially the expectation and the variance) of the distribution. Kuang [15] also obtained general stability results for systems of delay differential equations. More recently, sufficient conditions for the stability of delay differential equations with distributed delay have been obtained by Bernard, Bélair, and Mackey [5]. They used some properties of the distribution to prove these results. However, in all these works, the authors focused on sufficient conditions for the stability, there is no necessary condition in these studies, and these results are not applicable directly to the model considered in this paper.

This paper is organized as follows. In section 2, we present the model and establish boundedness properties of the solutions. In section 3, we study the asymptotic stability of the equilibria. We give conditions for the trivial equilibrium to be globally asymptotically stable in section 3.1 and investigate the stability of the nontrivial equilibrium in section 3.2. In section 4, we show that a local Hopf bifurcation occurs in our model. In section 5, numerical simulations are performed to demonstrate that our results can be used to explain the long period oscillations observed in chronic

myelogenous leukemia.

2. The hematopoiesis process: Presentation of the model. Denote by $r(t,a)$ and $p(t,a)$ the population densities of resting an proliferating cells, respectively, which have spent a time $a \geq 0$ in their phase at time $t \geq 0$. Resting cells can either be lost randomly at a rate $\delta \geq 0$, which takes into account the cellular differentiation, or enter into the proliferating phase at a rate β. Proliferating cells can be lost by apoptosis (a programmed cell death) at a rate $\gamma \geq 0$ and, at mitosis, cells with age a divide in two daughter cells (which immediately enter the G_0-phase) with a rate $g(a)$.

The function $g : [0, \overline{\tau}) \to \mathbb{R}^+$ satisfies $g(a) = 0$ if $a < \underline{\tau}$ with $0 \leq \underline{\tau} < \overline{\tau} < +\infty$. Moreover, it is assumed to be piecewise continuous such that $\int_{\underline{\tau}}^{\overline{\tau}} g(a)da = +\infty$. The later assumption describes the fact that cells which did not die have to divide before they reach the maximal age $\overline{\tau}$.

The nature of the trigger signal for introduction in the proliferating phase is not clear. However, the work of Sachs [30] shows that we can reasonably think that it strongly depends on the entire resting cell population, that is, $\beta = \beta(x(t))$, with

$$x(t) = \int_0^{+\infty} r(t,a)da, \quad t \geq 0.$$

The function β is supposed to be continuous and positive. Furthermore, from a reasonable biological point of view, we assume that β is decreasing with $\lim_{x\to+\infty} \beta(x) = 0$. This describes the fact that the rate of reentry into the proliferating compartment is a decreasing function of the G_0-phase population.

Usually, it is believed that the function β is a monotone decreasing Hill function (see Mackey [16]), given by

$$\beta(x) = \beta_0 \frac{\theta^n}{\theta^n + x^n}, \quad x \geq 0, \tag{2.1}$$

with $\beta_0 > 0$, $\theta \geq 0$, and $n > 0$. β_0 is the maximal rate of reentry in the proliferating phase, θ is the number of resting cells at which β has its maximum rate of change with respect to the resting phase population, and n describes the sensitivity of the reintroduction rate with changes in the population.

The above parameters values are usually chosen (see Mackey [16]) to be

$$\delta = 0.05 \text{ day}^{-1}, \quad \gamma = 0.2 \text{ day}^{-1}, \quad \beta_0 = 1.77 \text{ day}^{-1}, \quad \text{and} \quad n = 3. \tag{2.2}$$

Although a usual value of θ is $\theta = 1.62 \times 10^8$ cells/kg, it can be normalized without loss of generality when one makes a qualitative analysis of the population.

Then $r(t,a)$ and $p(t,a)$ satisfy the system of partial differential equations

$$\frac{\partial r}{\partial t} + \frac{\partial r}{\partial a} = -\big(\delta + \beta(x(t))\big)r, \quad a > 0,\ t > 0, \tag{2.3}$$

$$\frac{\partial p}{\partial t} + \frac{\partial p}{\partial a} = -\big(\gamma + g(a)\big)p, \qquad 0 < a < \overline{\tau},\ t > 0, \tag{2.4}$$

with

$$r(0,a) = \nu(a),\ a \geq 0, \qquad p(0,a) = \Gamma(a),\ a \in [0, \overline{\tau}].$$

The functions $\nu = \nu(a)$ and $\Gamma = \Gamma(a)$ give the population densities of cells which have spent a time a in the resting and proliferating phase, respectively, at time $t = 0$, that is, the initial populations of cells with age a in each phase.

The boundary conditions of system (2.3)–(2.4), which describe the cellular flux between the two phases, are given by

$$\begin{cases} r(t,0) = 2\displaystyle\int_{\underline{\tau}}^{\overline{\tau}} g(\tau)p(t,\tau)d\tau, \\ \\ p(t,0) = \beta(x(t))x(t). \end{cases}$$

Moreover, we suppose that $\lim_{a\to+\infty} r(t,a) = 0$ and $\lim_{a\to\overline{\tau}} p(t,a) = 0$.

Let $y(t)$ denote the total population density of proliferating cells at time t; then

$$y(t) = \int_0^{\overline{\tau}} p(t,a)da, \qquad t \geq 0.$$

Thus, integrating (2.3) and (2.4) with respect to the age variable, we obtain

$$\frac{dx}{dt} = -\big(\delta + \beta(x(t))\big)x(t) + 2\int_{\underline{\tau}}^{\overline{\tau}} g(\tau)p(t,\tau)d\tau, \tag{2.5}$$

$$\frac{dy}{dt} = -\gamma y(t) + \beta(x(t))x(t) - \int_{\underline{\tau}}^{\overline{\tau}} g(\tau)p(t,\tau)d\tau. \tag{2.6}$$

We define a function G by

$$G(t,a) = \begin{cases} g(a)\exp\left(-\displaystyle\int_{a-t}^{a} g(s)ds\right) & \text{if } t < a, \\ \\ g(a)\exp\left(-\displaystyle\int_{0}^{a} g(s)ds\right) & \text{if } a < t. \end{cases}$$

Set

$$f(\tau) := g(\tau)\exp\left(-\int_0^{\tau} g(s)ds\right), \quad \tau > 0.$$

One can check that f is a density function, supported on $[\underline{\tau}, \overline{\tau}]$, and f represents the density of division of proliferating cells. In particular, $\int_{\underline{\tau}}^{\overline{\tau}} f(\tau)d\tau = 1$.

Using the method of characteristics to determine $p(t,a)$, we deduce, from (2.5)–(2.6), that the process of hematopoiesis is described by the following system:

$$
\left\{
\begin{array}{rcl}
\dfrac{dx}{dt} &=& -\big(\delta+\beta(x(t))\big)x(t) \\
&& + \left\{
\begin{array}{ll}
2e^{-\gamma t}\displaystyle\int_{\underline{\tau}}^{\overline{\tau}} G(t,\tau)\Gamma(\tau-t)d\tau, & 0\le t\le \underline{\tau}, \\
2\displaystyle\int_{\underline{\tau}}^{t} e^{-\gamma\tau} f(\tau)\beta(x(t-\tau))x(t-\tau)d\tau & \\
\qquad + 2e^{-\gamma t}\displaystyle\int_{t}^{\overline{\tau}} G(t,\tau)\Gamma(\tau-t)d\tau, & \underline{\tau}\le t\le \overline{\tau}, \\
2\displaystyle\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)\beta(x(t-\tau))x(t-\tau)d\tau, & \overline{\tau}\le t,
\end{array}
\right. \\
\dfrac{dy}{dt} &=& -\gamma y(t)+\beta(x(t))x(t) \\
&& - \left\{
\begin{array}{ll}
e^{-\gamma t}\displaystyle\int_{\underline{\tau}}^{\overline{\tau}} G(t,\tau)\Gamma(\tau-t)d\tau, & 0\le t\le \underline{\tau}, \\
\displaystyle\int_{\underline{\tau}}^{t} e^{-\gamma\tau} f(\tau)\beta(x(t-\tau))x(t-\tau)d\tau & \\
\qquad + e^{-\gamma t}\displaystyle\int_{t}^{\overline{\tau}} G(t,\tau)\Gamma(\tau-t)d\tau, & \underline{\tau}\le t\le \overline{\tau}, \\
\displaystyle\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)\beta(x(t-\tau))x(t-\tau)d\tau, & \overline{\tau}\le t.
\end{array}
\right.
\end{array}
\right. \tag{2.7}
$$

One can give a direct biological explanation of system (2.7).

In the equation for the resting cells $x(t)$, the first term in the right-hand side accounts for G_0-cell loss due to either mortality and cellular differentiation (δ) or introduction in the proliferating phase (β). The second term represents a cellular gain due to the movement of proliferating cells one generation earlier. It requires some explanation. First, we recall that all cells divide according to the density f, supported on $[\underline{\tau},\overline{\tau}]$. We shall call, in the following, new proliferating cells, the resting cells introduced in the proliferating phase at the considered time t. When $t\le\underline{\tau}$, no new proliferating cell is mature enough to divide, because cells cannot divide before they have spent time $\underline{\tau}$ in the proliferating phase. Therefore, the cellular gain can proceed only from cells initially in the proliferating phase. When $t\in[\underline{\tau},\overline{\tau}]$, the cellular increase is obtained by division of new proliferating cells and by division of the initial population. Finally, when $t\ge\overline{\tau}$, all initial proliferating cells have divided or died, and the cellular gain is obtained by division of new proliferating cells introduced one generation earlier. The factor 2 always accounts for the division of each cell into two daughter cells at mitosis. The term $e^{-\gamma t}$, with $t\in[0,\overline{\tau}]$, describes the attenuation of the population, in the proliferating phase, due to apoptosis.

In the equation for the proliferating cells $y(t)$, the first term in the right-hand side accounts for cellular loss by apoptosis and the second term is for cellular entry from the G_0-phase. The last term accounts for the flux of proliferating cells to the resting compartment.

We set $\mu:=\int_0^\infty \nu(a)da$. Then, initially, the populations in the two phases are given by

$$x(0)=\mu \qquad \text{and} \qquad y(0)=\int_0^{\overline{\tau}}\Gamma(a)da.$$

At this point, one can make a remark. Since resting cells are introduced in the proliferating phase with a rate β, then $\Gamma(0)$, which represents the population of cells introduced at time $t = 0$ in the cycle, must satisfy

$$\Gamma(0) = \beta(\mu)\mu.$$

Taking into account the inevitable loss of proliferating cells by apoptosis and by division, we suppose that $\Gamma(a)$ is given by

$$\Gamma(a) = \begin{cases} e^{-\gamma a}\beta(\mu)\mu & \text{if } a \in [0, \underline{\tau}), \\ e^{-\gamma a}\exp\left(-\displaystyle\int_{\underline{\tau}}^{a} g(s)ds\right)\beta(\mu)\mu & \text{if } a \in [\underline{\tau}, \overline{\tau}). \end{cases} \tag{2.8}$$

This simply describes that Γ satisfies (2.4) (see Webb [35, p. 8]). With (2.8) and integrating by parts, the initial conditions of system (2.7) become

$$x(0) = \mu, \qquad y(0) = \beta(\mu)\mu \int_{\underline{\tau}}^{\overline{\tau}} f(\tau)\left(\frac{1 - e^{-\gamma\tau}}{\gamma}\right)d\tau. \tag{2.9}$$

When $\gamma = 0$, we have

$$y(0) = \beta(\mu)\mu \int_{\underline{\tau}}^{\overline{\tau}} \tau f(\tau)d\tau.$$

Assume that the function $x \mapsto x\beta(x)$ is Lipschitz continuous. It is immediate to show by steps that, for all $\mu \geq 0$, the system (2.7) under condition (2.9) has a unique nonnegative continuous solution $(x(t), y(t))$ defined on $[0, +\infty)$.

One can notice that problem (2.7) reduces to a system of two delay differential equations, with initial conditions solutions of a system of ordinary differential equations. On $[0, \underline{\tau}]$, the first equation for $x(t)$ in system (2.7) reduces to the ordinary differential equation

$$\begin{cases} \dfrac{d\widetilde{\varphi}}{dt} = -\big(\delta + \beta(\widetilde{\varphi}(t))\big)\widetilde{\varphi}(t) + 2\beta(\mu)\mu \displaystyle\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau, \quad 0 \leq t \leq \underline{\tau}, \\ \widetilde{\varphi}(0) = \mu, \end{cases} \tag{2.10}$$

and, on $[\underline{\tau}, \overline{\tau}]$, the second equation reduces to the nonautonomous delay differential equation

$$\begin{cases} \dfrac{d\varphi}{dt} = -\big(\delta + \beta(\varphi(t))\big)\varphi(t) + 2\beta(\mu)\mu \displaystyle\int_{t}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau \\ \qquad\qquad + 2\displaystyle\int_{\underline{\tau}}^{t} e^{-\gamma\tau} f(\tau)\beta(\varphi(t-\tau))\varphi(t-\tau)d\tau, \qquad t \in [\underline{\tau}, \overline{\tau}], \\ \varphi(t) = \widetilde{\varphi}(t), \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad t \in [0, \underline{\tau}], \end{cases} \tag{2.11}$$

where $\widetilde{\varphi}(t)$ is the unique solution of (2.10) for the initial condition μ.

In the same way, the solution $y(t)$ of the second equation in (2.7), denoted $\psi(t)$, is given in terms of the unique solution $\widetilde{\varphi}(t)$ of (2.10), associated with μ, and the unique solution $\varphi(t)$ of (2.11), for $t \in [0, \overline{\tau}]$.

Then, system (2.7) can be written as an autonomous system of delay differential equations, for $t \geq \overline{\tau}$,

$$\frac{dx}{dt} = -\big(\delta + \beta(x(t))\big)x(t) + 2\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)\beta(x(t-\tau))x(t-\tau)d\tau, \tag{2.12a}$$

$$\frac{dy}{dt} = -\gamma y(t) + \beta(x(t))x(t) - \int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)\beta(x(t-\tau))x(t-\tau)d\tau, \tag{2.12b}$$

with, for $t \in [0, \overline{\tau}]$,

$$x(t) = \varphi(t), \quad y(t) = \psi(t). \tag{2.13}$$

The solutions of (2.12b) are given explicitly by

$$y(t) = \int_{\underline{\tau}}^{\overline{\tau}} f(\tau)\bigg(\int_{t-\tau}^{t} e^{-\gamma(t-s)}\beta(x(s))x(s)\ ds\bigg)d\tau \qquad \text{for} \quad t \geq \overline{\tau}. \tag{2.14}$$

One can notice that $y(t)$ no longer depends on the initial population $\Gamma(a)$ after one generation, that is, when $t \geq \overline{\tau}$. This can be explained as follows. Cells initially in the proliferating phase have divided or died after one generation; hence, new cells in the proliferating phase can come only from resting cells $x(t)$.

On the other hand, one may have already noticed that the solutions of (2.12a) do not depend on the solutions of (2.12b), whereas the converse is not true. The expression of $y(t)$ in (2.14) gives more precise information on the influence of the behavior of $x(t)$ on the stability of the solutions $y(t)$. These results are proved in the following lemma.

LEMMA 2.1. *Let $(x(t), y(t))$ be a solution of* (2.12). *If $\lim_{t\to+\infty} x(t)$ exists and equals $C \geq 0$, then*

$$\lim_{t\to+\infty} y(t) = \begin{cases} \beta(C)C\displaystyle\int_{\underline{\tau}}^{\overline{\tau}} f(\tau)\bigg(\frac{1-e^{-\gamma\tau}}{\gamma}\bigg)d\tau & \text{if } \gamma > 0, \\ \beta(C)C\displaystyle\int_{\underline{\tau}}^{\overline{\tau}} \tau f(\tau)d\tau & \text{if } \gamma = 0. \end{cases} \tag{2.15}$$

If $x(t)$ is P-periodic, then $y(t)$ is also P-periodic.

Proof. By using (2.14), we obtain that

$$y(t) = \int_{\underline{\tau}}^{\overline{\tau}} f(\tau)\bigg(\int_{0}^{\tau} e^{-\gamma s}\beta(x(t-s))x(t-s)\ ds\bigg)d\tau \qquad \text{for} \quad t \geq \overline{\tau}. \tag{2.16}$$

Hence,

$$\lim_{t\to+\infty} y(t) = \beta(C)C\int_{\underline{\tau}}^{\overline{\tau}} f(\tau)\bigg(\int_{0}^{\tau} e^{-\gamma s}\ ds\bigg)d\tau,$$

and (2.15) follows immediately.

When $x(t)$ is P-periodic, then using (2.16) it is obvious to see that $y(t)$ is also periodic with the same period. ∎

Lemma 2.1 shows the influence of (2.12a) on the stability of the entire system, since the stability of solutions of (2.12a) leads to stability of the solutions of (2.12b).

Before studying the stability of (2.12a), we prove a boundedness result for the solutions of this equation. The proof is based on the one given by Mackey and Rudnicki [22] for a differential equation with a discrete delay.

PROPOSITION 2.2. *Assume that* $\delta > 0$. *Then the solutions of* (2.12a) *are bounded.*

Proof. Assume that $\delta > 0$ and $2(\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau)\beta(0) \geq \delta$. Since β is decreasing and $\lim_{x\to+\infty} \beta(x) = 0$, there exists a unique $x_0 \geq 0$ such that

$$2\bigg(\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau\bigg)\beta(x_0) = \delta$$

and

$$2\bigg(\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau\bigg)\beta(x) \leq \delta \qquad \text{for} \quad x \geq x_0. \tag{2.17}$$

If $2(\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau)\beta(0) < \delta$, then (2.17) holds with $x_0 = 0$. Set

$$x_1 := 2\bigg(\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau\bigg)\frac{\beta(0)x_0}{\delta} \geq 0.$$

One can check that

$$2\bigg(\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau\bigg)\max_{0\leq y\leq x}\Big(\beta(y)y\Big) \leq \delta x \qquad \text{for} \quad x \geq x_1. \tag{2.18}$$

Indeed, let $y \in [0, x)$. If $y \leq x_0$, then

$$2\bigg(\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau\bigg)\beta(y)y \leq 2\bigg(\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau\bigg)\beta(0)x_0 = \delta x_1 \leq \delta x,$$

and, if $y > x_0$, then

$$2\bigg(\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau\bigg)\beta(y)y \leq \delta y \leq \delta x.$$

Hence, (2.18) holds.

Assume, by contradiction, that $\limsup_{t\to+\infty} x(t) = +\infty$, where $x(t)$ is a solution of (2.12a). Then, there exists $t_0 > \overline{\tau}$ such that

$$x(t) \leq x(t_0) \ \text{ for } t \in [t_0 - \overline{\tau}, t_0] \qquad \text{and} \qquad x(t_0) > x_1.$$

With (2.18), we obtain that

$$2\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)\beta(x(t_0-\tau))x(t_0-\tau)d\tau \leq \delta x(t_0).$$

This yields, with (2.12a), that

$$\frac{dx}{dt}(t_0) \leq -\beta(x(t_0)))x(t_0) < 0,$$

which gives a contradiction. Hence, $\limsup_{t\to+\infty} x(t) < +\infty$. ∎

When $\delta = 0$, the solutions of (2.12a) may not be bounded. We show, in the next proposition, that these solutions may explode under some conditions. However, one can notice, using (2.16), that the solutions of (2.12b) may still be stable in this case.

PROPOSITION 2.3. *Assume that* $\delta = 0$ *and*

$$\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau > \frac{1}{2}. \tag{2.19}$$

In addition, assume that there exists $\overline{x} \geq 0$ *such that the function* $x \mapsto x\beta(x)$ *is decreasing for* $x \geq \overline{x}$. *If* $\mu \geq \overline{x}$, *then the unique solution* $x(t)$ *of* (2.12a) *satisfies*

$$\lim_{t\to+\infty} x(t) = +\infty.$$

Proof. One can notice that, if $\lim_{t\to+\infty} x(t) = C$ exists, then (2.12a) leads to

$$\left(2\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau - 1\right)\beta(C)C = 0.$$

It follows that $C = 0$.

Let $\mu \geq \overline{x}$ be given. Consider the equation

$$\widetilde{\varphi}'(t) = 2\beta(\mu)\mu \int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau - \beta(\widetilde{\varphi}(t))\widetilde{\varphi}(t) \quad \text{for} \quad 0 \leq t \leq \underline{\tau} \tag{2.20}$$

with $\widetilde{\varphi}(0) = \mu$. Since the function $x \mapsto x\beta(x)$ is decreasing for $x \geq \overline{x}$, it is immediate that every solution $\widetilde{\varphi}(t)$ of (2.20) satisfies, for $t \in [0, \underline{\tau}]$,

$$\widetilde{\varphi}'(t) \geq \left(2\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau - 1\right)\beta(\mu)\mu > 0.$$

Consider now the problem

$$\begin{cases} \varphi'(t) = -\beta(\varphi(t))\varphi(t) + 2\beta(\mu)\mu \displaystyle\int_t^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau \\ \qquad\qquad + 2\displaystyle\int_{\underline{\tau}}^{t} e^{-\gamma\tau} f(\tau)\beta(\varphi(t-\tau))\varphi(t-\tau) d\tau, \quad t \in [\underline{\tau}, \overline{\tau}], \\ \varphi(t) = \widetilde{\varphi}(t), \qquad\qquad t \in [0, \underline{\tau}], \end{cases} \tag{2.21}$$

where $\widetilde{\varphi}(t)$ is the unique solution of (2.20) for the initial condition μ. Then,

$$\varphi'(\underline{\tau}) \geq \left(2\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau - 1\right)\beta(\mu)\mu > 0.$$

So, there exists $\varepsilon > 0$ such that $\underline{\tau} + \varepsilon \leq \overline{\tau}$ and $\varphi'(t) > 0$ for $t \in [\underline{\tau}, \underline{\tau} + \varepsilon)$. Since $\mu \leq \varphi(\underline{\tau}) \leq \varphi(\tau) \leq \varphi(\underline{\tau} + \varepsilon)$, for $\tau \in [\underline{\tau}, \underline{\tau} + \varepsilon]$, we have

$$\begin{aligned} \varphi'(\underline{\tau} + \varepsilon) &\geq \left(2\int_{\underline{\tau}+\varepsilon}^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau - 1\right)\beta(\varphi(\underline{\tau} + \varepsilon))\varphi(\underline{\tau} + \varepsilon) \\ &\quad + 2\left(\int_{\underline{\tau}}^{\underline{\tau}+\varepsilon} e^{-\gamma\tau} f(\tau) d\tau\right)\beta(\varphi(\underline{\tau} + \varepsilon))\varphi(\underline{\tau} + \varepsilon) \\ &\geq \left(2\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau - 1\right)\beta(\varphi(\underline{\tau} + \varepsilon))\varphi(\underline{\tau} + \varepsilon). \end{aligned}$$

Condition (2.19) leads to $\varphi'(\underline{\tau}+\varepsilon)>0$. Using a similar argument, we obtain that

$$\varphi'(t)>0 \qquad \text{for} \quad t\in[\underline{\tau},\overline{\tau}].$$

To conclude, consider the delay differential equation

$$x'(t)=2\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau}f(\tau)\beta(x(t-\tau))x(t-\tau)d\tau-\beta(x(t))x(t) \tag{2.22}$$

with an initial condition given on $[\underline{\tau},\overline{\tau}]$ by the solution $\varphi(t)$ of (2.21). Using the same reasoning as in the previous cases, we obtain that

$$x'(\overline{\tau})>0.$$

We thus deduce that

$$x'(t)>0 \qquad \text{for} \quad t\geq 0.$$

This completes the proof. □

The assumption on the function $x\mapsto x\beta(x)$ in Proposition 2.3 is satisfied for example when β is given by (2.1), with $n>1$. In this case, we can take $\overline{x}=\theta/(n-1)^{1/n}$.

We now turn our attention to the stability of (2.12). Problem (2.12) has at most two equilibria. The first, $E_0=(0,0)$, always exists: it corresponds to the extinction of the population. The second describes the expected equilibrium of the population; it is a nontrivial equilibrium $E^*=(x^*,y^*)$, where x^* is the unique solution of

$$\left(2\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau}f(\tau)d\tau-1\right)\beta(x^*)=\delta \tag{2.23}$$

and, from (2.7) and (2.9),

$$y^*=\begin{cases}\beta(x^*)x^*\displaystyle\int_{\underline{\tau}}^{\overline{\tau}} f(\tau)\left(\frac{1-e^{-\gamma\tau}}{\gamma}\right)d\tau & \text{if } \gamma>0,\\[2ex] \delta x^*\displaystyle\int_{\underline{\tau}}^{\overline{\tau}}\tau f(\tau)d\tau, & \text{if } \gamma=0.\end{cases} \tag{2.24}$$

Since β is a positive decreasing function and $\lim_{x\to+\infty}\beta(x)=0$, then the equilibrium E^* exists if and only if

$$0<\delta<\left(2\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau}f(\tau)d\tau-1\right)\beta(0). \tag{2.25}$$

We shall study in section 3 the stability of the two equilibria E_0 and E^*. From Lemma 2.1, we only need to focus on the behavior of the equilibria of (2.12a), that is, $x\equiv 0$ and $x\equiv x^*$, to obtain information on the behavior of the entire population.

3. Asymptotic stability. We first show that E_0 is globally asymptotically stable when it is the only equilibrium and that it becomes unstable when the nontrivial equilibrium E^* appears: a transcritical bifurcation occurs then. In a second part, we determine conditions for the nontrivial equilibrium E^* to be asymptotically stable.

3.1. Stability of the trivial equilibrium. In the next theorem, we give a necessary and sufficient condition for the trivial equilibrium of (2.12a) to be globally asymptotically stable using a Lyapunov functional. For a definition of and information about Lyapunov functionals for delay differential equations, see [13].

THEOREM 3.1. *The trivial equilibrium of the system* (2.12) *is globally asymptotically stable if*

$$\left(2\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau - 1\right)\beta(0) < \delta \tag{3.1}$$

and unstable if

$$\delta < \left(2\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau - 1\right)\beta(0). \tag{3.2}$$

Proof. We first assume that (3.1) holds. Denote by C^+ the set of continuous nonnegative functions on $[0,\overline{\tau}]$ and define the mapping $J : C^+ \to [0,+\infty)$ by

$$J(\varphi) = B(\varphi(\overline{\tau})) + \int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)\left(\int_{\overline{\tau}-\tau}^{\overline{\tau}} \Big(\beta\big(\varphi(\theta)\big)\varphi(\theta)\Big)^2 d\theta\right)d\tau$$

for all $\varphi \in C^+$, where

$$B(x) = \int_0^x \beta(s)s\ ds \qquad \text{for all } x \geq 0.$$

We set (see [13])

$$\dot{J}(\varphi) = \limsup_{t\to 0^+} \frac{J(x_t^\varphi) - J(\varphi)}{t} \qquad \text{for } \varphi \in C^+,$$

where x^φ is the unique solution of (2.12a) associated with the initial condition $\varphi \in C^+$ and $x_t^\varphi(\theta) = x^\varphi(t+\theta)$ for $\theta \in [0,\overline{\tau}]$. Then,

$$\begin{aligned}\dot{J}(\varphi) \;=\; & \frac{d\varphi}{dt}(\overline{\tau})\beta\big(\varphi(\overline{\tau})\big)\varphi(\overline{\tau}) \\ & + \int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)\big((\beta\big(\varphi(\overline{\tau})\big)\varphi(\overline{\tau}))^2 - (\beta\big(\varphi(\overline{\tau}-\tau)\big)\varphi(\overline{\tau}-\tau))^2\big)d\tau.\end{aligned} \tag{3.3}$$

Using (2.12a), we have

$$\frac{d\varphi}{dt}(\overline{\tau}) = -\Big(\delta + \beta\big(\varphi(\overline{\tau})\big)\Big)\varphi(\overline{\tau}) + 2\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)\beta\big(\varphi(\overline{\tau}-\tau)\big)\varphi(\overline{\tau}-\tau)d\tau.$$

Therefore, (3.3) becomes

$$\begin{aligned}\dot{J}(\varphi) &= -\Big(\delta + \beta\big(\varphi(\overline{\tau})\big)\Big)\beta\big(\varphi(\overline{\tau})\big)\varphi^2(\overline{\tau}) + \int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)\Big[\Big(\beta\big(\varphi(\overline{\tau})\big)\varphi(\overline{\tau})\Big)^2 \\ &\quad + 2\beta\big(\varphi(\overline{\tau})\big)\varphi(\overline{\tau})\beta\big(\varphi(\overline{\tau}-\tau)\big)\varphi(\overline{\tau}-\tau) - \Big(\beta\big(\varphi(\overline{\tau}-\tau)\big)\varphi(\overline{\tau}-\tau)\Big)^2\Big]d\tau \\ &= -\Big(\delta + \beta\big(\varphi(\overline{\tau})\big)\Big)\beta\big(\varphi(\overline{\tau})\big)\varphi^2(\overline{\tau}) + 2\Big(\beta\big(\varphi(\overline{\tau})\big)\varphi(\overline{\tau})\Big)^2\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau \\ &\quad - \int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)\Big[\beta\big(\varphi(\overline{\tau})\big)\varphi(\overline{\tau}) - \beta\big(\varphi(\overline{\tau}-\tau)\big)\varphi(\overline{\tau}-\tau)\Big]^2 d\tau.\end{aligned}$$

Hence,

$$\dot{J}(\varphi) \leq -u(\varphi(\overline{\tau})),$$

where the function u is defined, for $x \geq 0$, by

$$u(x) = r(x)\beta(x)x^2 \tag{3.4}$$

with

$$r(x) = \delta - \left(2\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau - 1\right)\beta(x).$$

Since β is decreasing, r is a monotone function. Moreover, (3.1) leads to $r(0) > 0$, and $\lim_{x\to\infty} r(x) = \delta \geq 0$. Therefore, r is positive on $[0, +\infty)$.

Consequently, the function u defined by (3.4) is nonnegative on $[0, +\infty)$ and $u(x) = 0$ if and only if $x = 0$. We deduce that every solution of (2.12a), with $\varphi \in C^+$, tends to zero as t tends to $+\infty$.

We suppose now that (3.2) holds. The linearization of (2.12a) around $x \equiv 0$ leads to the characteristic equation

$$\Delta_0(\lambda) := \lambda + \delta + \beta(0) - 2\beta(0)\int_{\underline{\tau}}^{\overline{\tau}} e^{-(\lambda+\gamma)\tau} f(\tau)d\tau = 0. \tag{3.5}$$

We consider Δ_0 as a real function. Since

$$\frac{d\Delta_0}{d\lambda} = 1 + 2\beta(0)\int_{\underline{\tau}}^{\overline{\tau}} \tau e^{-(\lambda+\gamma)\tau} f(\tau)d\tau > 0,$$

it follows that Δ_0 is an increasing function. Moreover, (3.5) yields

$$\lim_{\lambda\to-\infty} \Delta_0(\lambda) = -\infty, \qquad \lim_{\lambda\to+\infty} \Delta_0(\lambda) = +\infty,$$

and (3.2) implies that

$$\Delta_0(0) = \delta - \left(2\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau - 1\right)\beta(0) < 0.$$

Hence, $\Delta_0(\lambda)$ has a unique real root which is positive. Consequently, (3.5) has at least one characteristic root with positive real part. Therefore, the equilibrium $x \equiv 0$ of (2.12a) is not stable. This completes the proof. □

The inequality (3.1) is satisfied when δ or γ (the mortality rates) is large or when $\beta(0)$ is small. Biologically, these conditions correspond to a population which cannot survive, because the mortality rates are too large or, simply, because not enough cells are introduced in the proliferating phase and, then, the population renewal is not supplied.

Remark 1. One can notice that when

$$\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau < \frac{1}{2},$$

the trivial equilibrium E_0 is the only equilibrium of (2.12) and is globally asymptotically stable. When

$$\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau = \frac{1}{2},$$

then E_0 is globally asymptotically stable if $\delta > 0$. When the equality

$$\left(2\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau - 1\right)\beta(0) = \delta$$

holds, one can check that $\lambda = 0$ is a characteristic root of (3.5) and all other characteristic roots have negative real parts. Hence, we cannot conclude on the stability or instability of the trivial equilibrium E_0 of (2.12) without further analysis. However, this is not the subject of this paper.

3.2. Stability of the nontrivial equilibrium. We concentrate, in this section, on the equilibrium $E^* = (x^*, y^*)$ defined by (2.23)–(2.24). Hence, throughout this section, we assume that (2.25) holds, that is,

$$0 < \delta < \left(2\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau - 1\right)\beta(0).$$

Since $\delta > 0$ and $\beta(0) > 0$, (2.25) implies, in particular, that

$$\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau > \frac{1}{2}. \tag{3.6}$$

From Lemma 2.1, we only need to focus on the stability of the nontrivial equilibrium $x \equiv x^*$ of (2.12a). To that aim, we linearize (2.12a) around x^*. Denote by $\beta^* \in \mathbb{R}$ the quantity

$$\beta^* := \frac{d}{dx}\Big(x\beta(x)\Big)\Big|_{x=x^*} = \beta(x^*) + x^*\beta'(x^*) \tag{3.7}$$

and set $u(t) = x(t) - x^*$. The linearization of (2.12a) is given by

$$\frac{du}{dt} = -(\delta + \beta^*)u(t) + 2\beta^* \int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau) u(t-\tau) d\tau.$$

Then, the characteristic equation is

$$\Delta(\lambda) := \lambda + \delta + \beta^* - 2\beta^* \int_{\underline{\tau}}^{\overline{\tau}} e^{-(\lambda+\gamma)\tau} f(\tau) d\tau = 0. \tag{3.8}$$

One can notice that the function $x \mapsto x\beta(x)$ is usually not monotone. For example, if β is given by (2.1) with $n > 1$, the function $x \mapsto x\beta(x)$ is increasing for $x \le \theta/(n-1)^{1/n}$ and decreasing for $x > \theta/(n-1)^{1/n}$. In this case, β^* is nonnegative when x^* is close to zero and negative when x^* is large enough.

The following theorem deals with the asymptotic stability of E^*.

THEOREM 3.2. *Assume that* (2.25) *holds. If*

$$\beta^* \ge -\frac{\delta}{2\displaystyle\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau + 1}, \tag{3.9}$$

then E^* is locally asymptotically stable.

Proof. We first prove that the equilibrium $x \equiv x^*$ is locally asymptotically stable when $\beta^* \geq 0$. We consider the mapping $\Delta(\lambda)$, given by (3.8), as a real function of λ. Then $\Delta(\lambda)$ is continuously differentiable on $\mathbb{R}$ and its first derivative is given by

$$\frac{d\Delta}{d\lambda} = 1 + 2\beta^* \int_{\underline{\tau}}^{\overline{\tau}} \tau e^{-(\lambda+\gamma)\tau} f(\tau) d\tau > 0. \tag{3.10}$$

Hence, $\Delta(\lambda)$ is an increasing function of λ satisfying

$$\lim_{\lambda \to -\infty} \Delta(\lambda) = -\infty \qquad \text{and} \qquad \lim_{\lambda \to +\infty} \Delta(\lambda) = +\infty.$$

Then, there exists a unique $\lambda_0 \in \mathbb{R}$ such that $\Delta(\lambda_0) = 0$. Moreover, since

$$\Delta(0) = \delta - \left(2 \int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau - 1 \right) \beta^*,$$

we deduce, by using (2.23), (3.6), and (3.7), that

$$\Delta(0) = -\left(2 \int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau - 1 \right) x^* \beta'(x^*) > 0.$$

Consequently, $\lambda_0 < 0$.

Let $\lambda = \mu + i\omega$ be a characteristic root of (3.8) such that $\mu > \lambda_0$. Considering the real part of (3.8), we obtain that

$$\mu = -(\delta + \beta^*) + 2\beta^* \int_{\underline{\tau}}^{\overline{\tau}} e^{-(\mu+\gamma)\tau} f(\tau) \cos(\omega\tau) d\tau. \tag{3.11}$$

Using (3.8), with $\lambda = \lambda_0$, together with (3.11), we then obtain

$$\mu - \lambda_0 = 2\beta^* \int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau) \left[e^{-\mu\tau} \cos(\omega\tau) - e^{-\lambda_0\tau} \right] d\tau.$$

However,

$$e^{-\mu\tau} \cos(\omega\tau) - e^{-\lambda_0\tau} < 0$$

for all $\tau \in [\underline{\tau}, \overline{\tau}]$. So we obtain that $\mu - \lambda_0 < 0$, which leads to a contradiction. This implies that all characteristic roots of (3.8) have negative real part and the equilibrium $x \equiv x^*$ of (2.12a) is locally asymptotically stable.

Now, assume that $\beta^* < 0$ and

$$\beta^* > -\frac{\delta}{2 \displaystyle\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau + 1}. \tag{3.12}$$

Let $\lambda = \mu + i\omega$ be a characteristic root of (3.8) such that $\mu > 0$. Since

$$\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau) \left(e^{-\mu\tau} \cos(\omega\tau) + 1 \right) d\tau \geq 0,$$

we have

$$2\beta^* \int_{\underline{\tau}}^{\overline{\tau}} e^{-(\mu+\gamma)\tau} f(\tau)\cos(\omega\tau)d\tau \leq -2\beta^* \int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau.$$

So, (3.11) and (3.12) lead to

$$\mu \leq -(\delta+\beta^*) - 2\beta^* \int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau < 0,$$

a contradiction. Therefore, $\mu \leq 0$.

Suppose now that (3.8) has a purely imaginary characteristic root $i\omega$, with $\omega \in \mathbb{R}$. Then, (3.11) leads to

$$\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)\cos(\omega\tau)d\tau = \frac{\delta+\beta^*}{2\beta^*}.$$

However,

$$\left| \int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)\cos(\omega\tau)d\tau \right| \leq \int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau$$

and (3.12) yields

$$\frac{\delta+\beta^*}{2\beta^*} < -\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau.$$

Hence, (3.8) has no purely imaginary root. Consequently, all characteristic roots of (3.8) have negative real part and the nontrivial equilibrium $x \equiv x^*$ of (2.12a) is locally asymptotically stable.

Finally, assume that

$$\beta^* = -\frac{\delta}{2\displaystyle\int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)d\tau + 1}. \tag{3.13}$$

Consider a characteristic root $\lambda = \mu + i\omega$ of (3.8), which reduces, with (3.13), to

$$\lambda - 2\beta^* \int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)(1+e^{-\lambda\tau})d\tau = 0. \tag{3.14}$$

Suppose, by contradiction, that $\mu > 0$. By considering the real part of (3.14), we have

$$\mu = 2\beta^* \int_{\underline{\tau}}^{\overline{\tau}} e^{-\gamma\tau} f(\tau)(1+e^{-\mu\tau}\cos(\omega\tau))d\tau < 0.$$

We obtain a contradiction; therefore $\mu \leq 0$. If we suppose now that $\mu = 0$, then we easily obtain that

$$\cos(\omega\tau) = -1 \qquad \text{for all } \tau \in [\underline{\tau}, \overline{\tau}],$$

which is impossible. It follows that all characteristic roots of (3.8) have negative real parts when (3.13) holds and the equilibrium $x \equiv x^*$ is locally asymptotically stable.

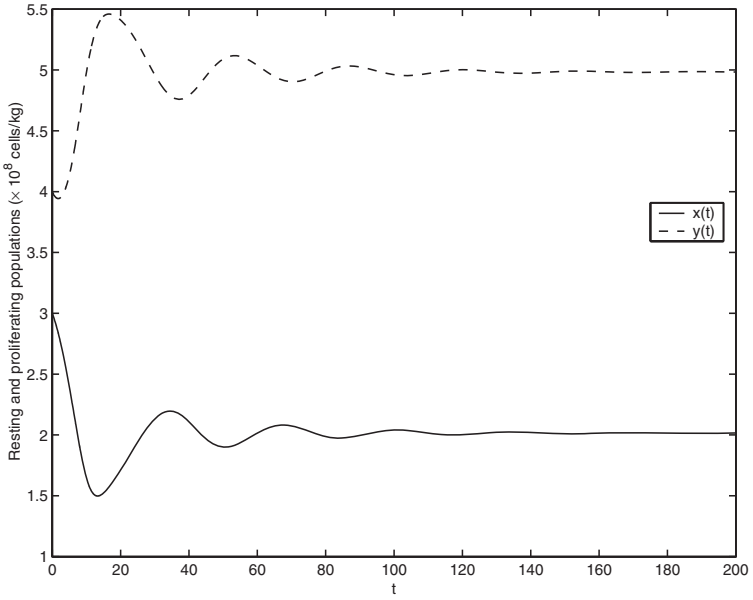


FIG. 3.1. *The solutions $x(t)$ (solid curve) and $y(t)$ (dashed curve) of system* (2.12) *are drawn for values of the parameters β_0, δ, and γ given by* (2.2), *$n = 2.42$, $\underline{\tau} = 0$, and $\overline{\tau} = 7$ days. In this case, the nontrivial equilibrium E^* is locally asymptotically stable, although the solutions oscillate transiently.*

From Lemma 2.1, we conclude that E^* is locally asymptotically stable when (3.9) holds. □

The asymptotic stability of E^* is shown in Figure 3.1. Values of the parameters are given by (2.2), except $n = 2.42$, $\underline{\tau} = 0$ and $\overline{\tau} = 7$ days. The function f is defined by

$$f(\tau) = \begin{cases} \dfrac{1}{\overline{\tau} - \underline{\tau}} & \text{if } \tau \in [\underline{\tau}, \overline{\tau}], \\ 0 & \text{otherwise.} \end{cases} \tag{3.15}$$

The MATLAB solver for delay differential equations, dde23 [32], is used to obtain Figure 3.1, as well as illustrations in sections 4 and 5.

When (3.9) does not hold, we have necessarily $\beta^* < 0$. In this case, we cannot obtain the stability of E^* for all values of β^*. In fact, in the next section we are going to show that the equilibrium E^* can be destabilized, in this case, via a Hopf bifurcation.

4. Hopf bifurcation and periodic solutions. In this section, we show that the equilibrium $x \equiv x^*$ of (2.12a) can become unstable when (3.9) does not hold anymore. Throughout this section, we assume that

$$\underline{\tau} = 0$$

and (2.25) holds, that is,

$$0 < \delta < \left(2 \int_0^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau - 1 \right) \beta(0).$$

From Proposition 2.2, the solutions of (2.12a) are bounded. Consequently, instability in (2.12a) occurs only via oscillatory solutions.

We assume that

$$\beta^* < -\frac{\delta}{2\int_0^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau + 1} := \widetilde{\delta}. \tag{4.1}$$

Otherwise, the nontrivial equilibrium $x \equiv x^*$ of (2.12a) is locally asymptotically stable (see Theorem 3.2).

If instability occurs for a particular value $\beta^* < \widetilde{\delta}$, a characteristic root of (3.8) must intersect the imaginary axis. Hence, we look for purely imaginary characteristic roots $i\omega$, $\omega \in \mathbb{R}$, of (3.8). If $i\omega$ is a characteristic root of (3.8), then ω is a solution of the system

$$\begin{cases} \delta + \beta^*(1 - 2C(\omega)) = 0, \\ \omega + 2\beta^* S(\omega) = 0, \end{cases} \tag{4.2}$$

where

$$C(\omega) := \int_0^{\overline{\tau}} e^{-\gamma\tau} f(\tau) \cos(\omega\tau) d\tau \quad \text{and} \quad S(\omega) := \int_0^{\overline{\tau}} e^{-\gamma\tau} f(\tau) \sin(\omega\tau) d\tau.$$

One can notice that $\omega = 0$ is not a solution of (4.2). Otherwise,

$$\delta = \left(2\int_0^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau - 1\right)\beta^* < 0,$$

which gives a contradiction. Moreover, if ω is a solution of (4.2), then $-i\omega$ is also a characteristic root. Thus, we look only for positive solutions ω.

LEMMA 4.1. *Assume that the function $\tau \mapsto e^{-\gamma\tau} f(\tau)$ is decreasing. Then, for each δ such that* (2.25) *is satisfied,* (4.2) *has at least one solution (β_c^*, ω_c) with $\beta_c^* < \widetilde{\delta}$ and $\omega_c > 0$. It follows that* (3.8) *has at least one pair of purely imaginary roots $\pm i\omega_c$ for $\beta^* = \beta_c^*$. Moreover, $\pm i\omega_c$ are simple characteristic roots of* (3.8). *Consider the branch of characteristic roots $\lambda(-\beta^*)$ such that $\lambda(-\beta_c^*) = i\omega_c$. Then*

$$\left.\frac{d\mathrm{Re}(\lambda)}{d(-\beta^*)}\right|_{\beta^*=\beta_c^*} > 0 \quad \textit{if and only if} \quad -\delta\left(\frac{S(\omega_c)}{\omega_c}\right)' > C'(\omega_c). \tag{4.3}$$

Proof. First, we show by induction that $S(\omega) > 0$ for $\omega > 0$. It is clear that $S(\omega) > 0$ if $\omega\overline{\tau} \in (0, \pi]$. Suppose that $\omega\overline{\tau} \in (\pi, 2\pi]$. Then

$$\begin{aligned} S(\omega) &= \frac{1}{\omega}\int_0^{\omega\overline{\tau}} e^{-\gamma\frac{\tau}{\omega}} f\left(\frac{\tau}{\omega}\right)\sin(\tau) d\tau \\ &= \frac{1}{\omega}\int_0^{\pi} e^{-\gamma\frac{\tau}{\omega}} f\left(\frac{\tau}{\omega}\right)\sin(\tau) d\tau + \frac{1}{\omega}\int_{\pi}^{\omega\overline{\tau}} e^{-\gamma\frac{\tau}{\omega}} f\left(\frac{\tau}{\omega}\right)\sin(\tau) d\tau. \end{aligned}$$

Since f is supported on the interval $[0, \overline{\tau}]$, it follows that

$$\int_{\omega\overline{\tau}}^{2\pi} e^{-\gamma\frac{\tau}{\omega}} f\left(\frac{\tau}{\omega}\right)\sin(\tau) d\tau = 0.$$

So, we obtain

$$S(\omega) = \frac{1}{\omega}\int_0^{\pi} e^{-\gamma\frac{\tau}{\omega}} f\Big(\frac{\tau}{\omega}\Big)\sin(\tau)d\tau + \frac{1}{\omega}\int_{\pi}^{2\pi} e^{-\gamma\frac{\tau}{\omega}} f\Big(\frac{\tau}{\omega}\Big)\sin(\tau)d\tau$$
$$= \frac{1}{\omega}\int_0^{\pi}\left(e^{-\gamma\frac{\tau}{\omega}} f\Big(\frac{\tau}{\omega}\Big) - e^{-\gamma\frac{\tau+\pi}{\omega}} f\Big(\frac{\tau+\pi}{\omega}\Big)\right)\sin(\tau)d\tau.$$

Since the function $\tau \mapsto e^{-\gamma\tau} f(\tau)$ is decreasing, we finally get $S(\omega) > 0$. Using a similar argument for $\omega\overline{\tau} \in (k\pi, (k+1)\pi]$, with $k \in \mathbb{N}$, $k \geq 2$, we deduce that $S(\omega) > 0$ for all $\omega > 0$.

Consider the equation

$$g(\omega) := \frac{\omega\big(1 - 2C(\omega)\big)}{2S(\omega)} = \delta, \qquad \omega > 0. \tag{4.4}$$

The function g is continuous with

$$\lim_{\omega\to 0} g(\omega) = \frac{1 - 2C(0)}{2\displaystyle\int_0^{\overline{\tau}} \tau e^{-\gamma\tau} f(\tau)d\tau} < 0 \tag{4.5}$$

because (2.25) leads to $1 - 2C(0) < 0$. Moreover, the Riemann–Lebesgue lemma implies that

$$\lim_{\omega\to+\infty} C(\omega) = \lim_{\omega\to+\infty} S(\omega) = 0.$$

This yields

$$\lim_{\omega\to+\infty} g(\omega) = +\infty.$$

We conclude that there exists a solution $\omega_c > 0$ of (4.4). Since $S(\omega_c) > 0$ and $g(\omega_c) = \delta > 0$, we obtain $1 - 2C(\omega_c) > 0$. Set

$$\beta_c^* = -\frac{\delta}{1 - 2C(\omega_c)} < 0. \tag{4.6}$$

Since $|C(\omega_c)| < C(0)$, it follows that

$$\beta_c^* < -\frac{\delta}{2C(0) + 1} = \widetilde{\delta}.$$

One can check that (β_c^*, ω_c) is a solution of (4.2). It follows that $\pm i\omega_c$ are characteristic roots of (3.8) for $\beta^* = \beta_c^*$.

Define a branch of characteristic roots $\lambda(-\beta^*)$ of (3.8) such that $\lambda(-\beta_c^*) = i\omega_c$. We use the parameter $-\beta^*$ because $\beta^* < \widetilde{\delta} < 0$.

Using (3.8), we obtain

$$\left[1 + 2\beta^*\int_0^{\overline{\tau}} \tau e^{-(\lambda+\gamma)\tau} f(\tau)d\tau\right]\frac{d\lambda}{d(-\beta^*)} = 1 - 2\int_0^{\overline{\tau}} e^{-(\lambda+\gamma)\tau} f(\tau)d\tau. \tag{4.7}$$

If we assume, by contradiction, that $i\omega_c$ is not a simple root of (3.8), then (4.7) leads to

$$C(\omega_c) = \frac{1}{2} \qquad \text{and} \qquad S(\omega_c) = 0.$$

Since $S(\omega_c) > 0$, we obtain a contradiction. Thus, $i\omega_c$ is a simple root of (3.8).

Moreover, using (4.7), we have

$$\left(\frac{d\lambda}{d(-\beta^*)}\right)^{-1} = \frac{1 + 2\beta^* \displaystyle\int_0^{\overline{\tau}} \tau e^{-(\lambda+\gamma)\tau} f(\tau) d\tau}{1 - 2\displaystyle\int_0^{\overline{\tau}} e^{-(\lambda+\gamma)\tau} f(\tau) d\tau}.$$

Since λ is a characteristic root of (3.8), we also have

$$1 - 2\int_0^{\overline{\tau}} e^{-(\lambda+\gamma)\tau} f(\tau) d\tau = -\frac{\lambda + \delta}{\beta^*}.$$

So, we deduce

$$\left(\frac{d\lambda}{d(-\beta^*)}\right)^{-1} = -\beta^* \frac{1 + 2\beta^* \displaystyle\int_0^{\overline{\tau}} \tau e^{-(\lambda+\gamma)\tau} f(\tau) d\tau}{\lambda + \delta}.$$

Then,

$$\begin{aligned}
\operatorname{sign}\left\{\frac{d\mathrm{Re}(\lambda)}{d(-\beta^*)}\right\}\Big|_{\beta^*=\beta_c^*} &= \operatorname{sign}\left\{\mathrm{Re}\left(\frac{d\lambda}{d(-\beta^*)}\right)^{-1}\right\}\Big|_{\beta^*=\beta_c^*} \\
&= \operatorname{sign}\left\{\mathrm{Re}\left(-\beta^* \frac{1 + 2\beta^* \displaystyle\int_0^{\overline{\tau}} \tau e^{-(\lambda+\gamma)\tau} f(\tau) d\tau}{\lambda + \delta}\right)\right\}\Big|_{\beta^*=\beta_c^*} \\
&= \operatorname{sign}\left\{ -\beta_c^* \frac{\delta(1 + 2\beta_c^* S'(\omega_c)) + 2\beta_c^* \omega_c C'(\omega_c)}{\delta^2 + \omega_c^2}\right\} \\
&= \operatorname{sign}\left\{\delta(1 + 2\beta_c^* S'(\omega_c)) + 2\beta_c^* \omega_c C'(\omega_c)\right\}.
\end{aligned}$$

From (4.6) and the fact that $1 - 2C(\omega_c) > 0$, this leads to

$$\begin{aligned}
\operatorname{sign}\left\{\frac{d\mathrm{Re}(\lambda)}{d(-\beta^*)}\right\}\Big|_{\beta^*=\beta_c^*} &= \operatorname{sign}\left\{1 - 2C(\omega_c) - 2\delta S'(\omega_c) - 2\omega_c C'(\omega_c)\right\} \\
&= \operatorname{sign}\left\{2\omega_c\left(- C'(\omega_c) - \delta\left(\frac{S(\omega_c)}{\omega_c}\right)'\right)\right\} \\
&= \operatorname{sign}\left\{ - C'(\omega_c) - \delta\left(\frac{S(\omega_c)}{\omega_c}\right)'\right\}.
\end{aligned}$$

This concludes the proof. □

Remark 2. Consider the function g defined by (4.4) and denote by α the quantity

$$\alpha := \left(2\int_0^{\overline{\tau}} e^{-\gamma\tau} f(\tau) d\tau - 1\right)\beta(0).$$

Define the sets

$$\Omega := \{\omega > 0;\ 0 < g(\omega) < \alpha \ \text{ and } \ g'(\omega) = 0\} \qquad \text{and} \qquad \Lambda := g(\Omega).$$

One can notice that Λ is finite (or empty). If $\delta \in (0,\alpha) \setminus \Lambda$, then

$$\left.\frac{d\mathrm{Re}(\lambda)}{d(-\beta^*)}\right|_{\beta^*=\beta_c^*} \neq 0.$$

Indeed, we have

$$g'(\omega) = -\frac{\omega}{S(\omega)}\left(g(\omega)\left(\frac{S(\omega)}{\omega}\right)' + C'(\omega)\right), \qquad \omega > 0.$$

Since $\delta \notin \Lambda$, we have $g'(\omega_c) \neq 0$. Moreover, $g(\omega_c) = \delta$. Thus

$$C'(\omega_c) \neq -\delta\left(\frac{S(\omega_c)}{\omega_c}\right)'.$$

We conclude by using (4.3).

Lemma 4.1, together with Remark 2, allows us to state and prove the following theorem.

THEOREM 4.2. *Assume that the function $\tau \mapsto e^{-\gamma\tau} f(\tau)$ is decreasing. Then, for each $\delta \notin \Lambda$ satisfying* (2.25)*, there exists $\beta_c^* < \widetilde{\delta}$ such that the equilibrium $x \equiv x^*$ is locally asymptotically stable when $\beta_c^* < \beta^* \leq \widetilde{\delta}$ and a Hopf bifurcation occurs at $x \equiv x^*$ when $\beta^* = \beta_c^*$.*

Proof. First, recall that $x \equiv x^*$ is locally asymptotically stable when $\beta^* = \widetilde{\delta}$ (see Theorem 3.2). We recall that, from the properties of the function g, (4.4) has a finite number of solutions (see Lemma 4.1). We set

$$\beta_c^* = -\frac{\delta}{1 - 2C(\omega_c^*)},$$

where ω_c^* is the smaller positive real such that

$$C(\omega_c^*) = \min\{C(\omega);\ \omega \text{ is a solution of (4.4)}\}.$$

Then, β_c^* is the maximum value of β^* (as defined in Lemma 4.1) which gives a solution of (4.2). From Lemma 4.1, (3.8) has no purely imaginary roots while $\beta_c^* < \beta^* \leq \widetilde{\delta}$. Consequently, Rouché's theorem [10, p. 248] leads to the local asymptotic stability of $x \equiv x^*$.

When $\beta^* = \beta_c^*$, (3.8) has a pair of purely imaginary roots $\pm i\omega_c$, $\omega_c > 0$ (see Lemma 4.1). Moreover, since $\delta \notin \Lambda$, Remark 2 implies that

$$\left.\frac{d\mathrm{Re}(\lambda)}{d(-\beta^*)}\right|_{\beta^*=\beta_c^*} \neq 0.$$

Assume, by contradiction, that

$$\frac{d\mathrm{Re}(\lambda)}{d(-\beta^*)} < 0$$

for $\beta^* > \beta_c^*$, β^* close to β_c^*. Then there exists a characteristic root $\lambda(-\beta^*)$ such that $\mathrm{Re}\lambda(-\beta^*) > 0$. This contradicts the fact that $x \equiv x^*$ is locally asymptotically stable when $\beta^* > \beta_c^*$. Thus, we obtain

$$\left.\frac{d\mathrm{Re}(\lambda)}{d(-\beta^*)}\right|_{\beta^*=\beta_c^*} > 0.$$

This implies the existence of a Hopf bifurcation at $x \equiv x^*$ for $\beta^* = \beta_c^*$. ∎

With the values of δ, γ and β_0 given by (2.2), and $\overline{\tau} = 7$ days, (2.12) has periodic solutions for $\beta_c^* = -0.3881$ with a period about 33 days. This value of β_c^* corresponds to $n = 2.53$ (see Figures 4.1 and 4.2). The function f is given by (3.15).

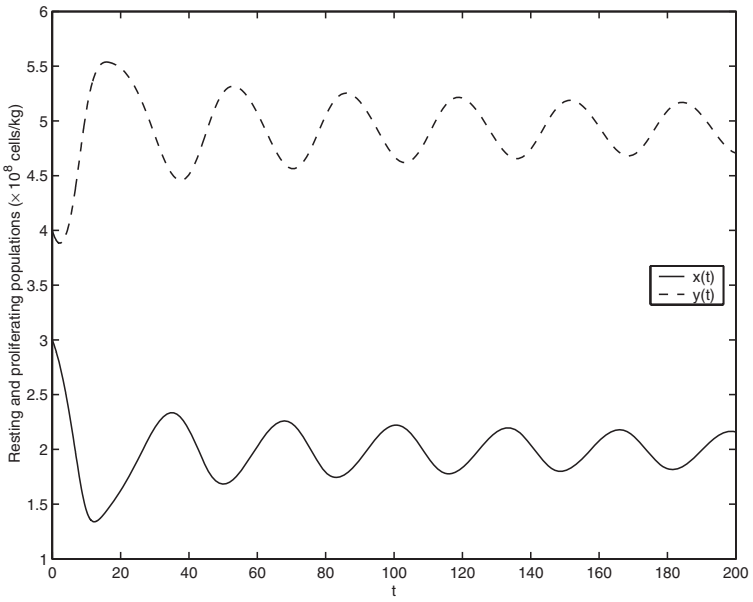


FIG. 4.1. *The solutions of system* (2.12), $x(t)$ (*solid curve*) *and* $y(t)$ (*dashed curve*), *are drawn when the Hopf bifurcation occurs. This corresponds to* $n = 2.53$ *with the other parameters given by* (2.2) *and* $\overline{\tau} = 7$ *days. Periodic solutions appear with period of the oscillations about* 33 *days.*

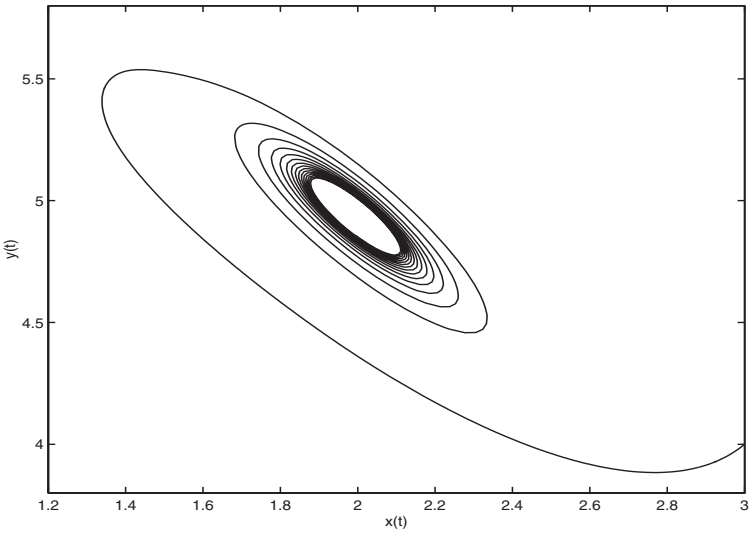


FIG. 4.2. *For the values used in Figure* 4.1, *the solutions are shown in the* (x, y)*-plane: the trajectories reach a limit cycle, surrounding the equilibrium.*

The bifurcation parameter was chosen to be β^* in this study, and the values of β^* depend strongly on the sensitivity n of the function $\beta(x)$, since all other parameters are fixed by (2.25). In this model, the sensitivity n plays a crucial role in the appearance of periodic solutions. Pujo-Menjouet and Mackey [27] already noticed the influence of this parameter on system (2.12) when the delay is constant (or equivalently, when f is a Dirac measure). The sensitivity n describes the way the rate of introduction in the proliferating phase reacts to changes in the resting phase population produced by external stimuli: a release of erythropoietin, for example, or the action of some growth factors.

Of course, the influence of other parameters (like mortality rates δ and γ, or the minimum and maximum delays $\underline{\tau}$ and $\overline{\tau}$) on the appearance of periodic solutions could be studied. However, since periodic hematological diseases—defined and described in section 5—are supposed to be due to hormonal control destabilization (see [11]), then the parameter n, among other parameters, seems to be appropriate to identify causes leading to periodic solutions in (2.12).

5. Discussion. Among the wide range of diseases affecting blood cells, periodic hematological diseases (Haurie, Dale, and Mackey [14]) are of main importance because of their intrinsic nature. These diseases are characterized by significant oscillations in the number of circulating cells, with periods ranging from weeks (19 to 21 days for cyclical neutropenia [14]) to months (30 to 100 days for chronic myelogenous leukemia [14]) and amplitudes varying from normal to low levels or normal to high levels, depending on the cells types [14]. Because of their dynamic character, periodic hematological diseases offer an opportunity to understand some of the regulating processes involved in the production of hematopoietic cells, which are still not well understood.

Some periodic hematological diseases involve only one type of blood cells, for example, red blood cells in periodic autoimmune hemolytic anemia (Bélair, Mackey, and Mahaffy [4]) or platelets in cyclical thrombocytopenia (Santillan et al. [31]). In these cases, periods of the oscillations are usually between two and four times the bone marrow production delay. However, other periodic hematological diseases, such as cyclical neutropenia (Haurie, Dale, and Mackey [14]) or chronic myelogenous leukemia (Fortin and Mackey [11]), show oscillations in all of the circulating blood cells, i.e., white cells, red blood cells, and platelets. These diseases involve oscillations with quite long periods (on the order of weeks to months). A destabilization of the pluripotential stem cell population (from which all of the mature blood cells types are derived) seems to be at the origin of these diseases.

We focus, in particular, on chronic myelogenous leukemia (CML), a cancer of the white cells, resulting from the malignant transformation of a single pluripotential stem cell in the bone marrow (Pujo-Menjouet, Bernard, and Mackey [26]). As described in Morley, Baikie, and Galton [24], oscillations can be observed in patients with CML, with the same period for white cells, red blood cells and platelets. This is called periodic chronic myelogenous leukemia (PCML). The period of the oscillations in PCML ranges from 30 to 100 days [14], [11] depending on patients. The difference between these periods and the average pluripotential cell cycle duration (between 1 and 4 days, as observed in mice [18]) is still not well understood.

Recently, to understand the dynamics of periodic chronic myelogenous leukemia, Pujo-Menjouet, Bernard, and Mackey [26] considered a model for the regulation of stem cell dynamics and investigated the influence of parameters in this stem cell model on the oscillations period when the model becomes unstable and starts to

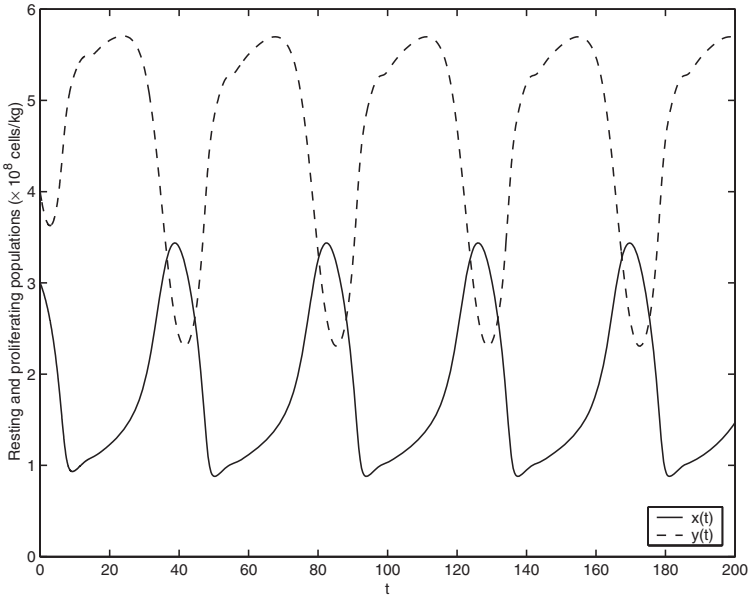


FIG. 5.1. *Solutions $x(t)$ (solid curve) and $y(t)$ (dashed curve) of system* (2.12) *oscillate with periods close to* 45 *days; the parameters are the same as in Figure* 4.1*, with $n = 3$. The amplitudes of the oscillations range from low values to normal values.*

oscillate. In this paper, taking into account the fact that a cell cycle has two phases, that is, stem cells in process are either in a resting phase or actively proliferating, and assuming that cells divide at different ages, we proposed a system of differential equations with distributed delay to model the dynamics of hematopoietic stem cells. By constructing a Lyapunov functional, we gave conditions for the trivial equilibrium to be globally asymptotically stable. Local stability and Hopf bifurcation of the nontrivial equilibrium were studied, the existence of a Hopf bifurcation leading to the appearance of periodic solutions in this model, with a period around 30 days at the bifurcation.

Numerical simulations show that periodic solutions occur after the bifurcation, with periods increasing as the bifurcation parameter (the sensitivity n) increases. In Figure 5.1, solutions oscillate around the equilibrium values with periods around 45 days. Moreover, amplitudes of the oscillations range from low values to normal values. The sensitivity is equal to $n = 3$; that is, the parameters are given by (2.2). This corresponds to values given by Mackey [16], values for which abnormal behavior (periodic) is usually observed in all circulating blood cells types.

When n continues to increase, longer oscillations periods are observed with amplitudes varying from low values to high values (see Figure 5.2). This situation characterizes periodic chronic myelogenous leukemia, with periods in the order of 2 months (70 days).

Moreover, the oscillations observed in Figures 5.1 and 5.2 look very much like relaxation oscillations. Experimental data from patients with PCML suggest that the shape of oscillations is of a relaxation oscillator type [11, 14]. Furthermore, Fowler and Mackey [12] showed that a model for hematopoiesis with a discrete delay may also exhibit relaxation oscillations. Therefore, it seems that not only periods and

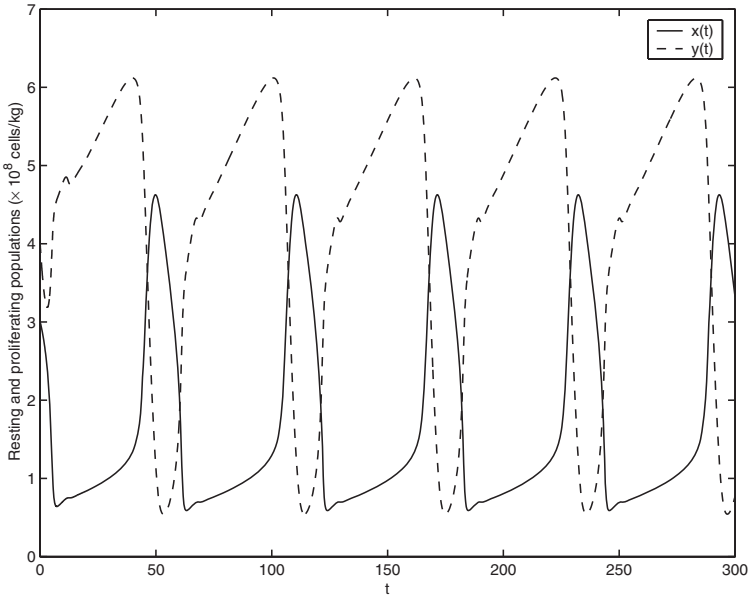


FIG. 5.2. *Solutions $x(t)$ (solid curve) and $y(t)$ (dashed curve) of system* (2.12) *oscillate with periods close to* 70 *days; the parameters are the same as in Figure* 4.1*, with $n = 4$. The amplitudes of the oscillations range from low values to high values.*

amplitudes of the oscillations correspond to the ones observed in PCML but also the shape of the oscillations.

Numerical simulations demonstrated that long period oscillations in the circulating cells are possible in our model even with short duration cell cycles. Thus, we are able to characterize some hematological diseases, especially those that exhibit a periodic behavior of all the circulating blood cells.

Acknowledgments. We are grateful to the two anonymous referees for their helpful comments and suggestions.

REFERENCES

[1] M. ADIMY, F. CRAUSTE, AND S. RUAN, *Stability and Hopf bifurcation in a mathematical model of pluripotent stem cell dynamics*, Nonlinear Anal. Real World Appl., to appear.

[2] R. F. V. ANDERSON, *Geometric and probabilistic stability criteria for delay systems*, Math. Biosci., 105 (1991), pp. 81–96.

[3] R. F. V. ANDERSON, *Intrinsic parameters and stability of differential-delay equations*, J. Math. Anal. Appl., 163 (1992), pp. 184–199.

[4] J. BÉLAIR, M. C. MACKEY, AND J. M. MAHAFFY, *Age-structured and two-delay models for erythropoiesis*, Math. Biosci., 128 (1995), pp. 317–346.

[5] S. BERNARD, J. BELAIR, AND M. C. MACKEY, *Sufficient conditions for stability of linear differential equations with distributed delay*, Discrete Contin. Dyn. Syst. Ser. B, 1 (2001), pp. 233–256.

[6] F. G. BOESE, *The stability chart for the linearized Cushing equation with a discrete delay and Gamma-distributed delays*, J. Math. Anal. Appl., 140 (1989), pp. 510–536.

[7] G. BRADFORD, B. WILLIAMS, R. ROSSI, AND I. BERTONCELLO, *Quiescence, cycling, and turnover in the primitive haematopoietic stem cell compartment*, Exper. Hematol., 25 (1997), pp. 445–453.

[8] F. J. BURNS AND I. F. TANNOCK, *On the existence of a G_0 phase in the cell cycle*, Cell. Tissue Kinet., 19 (1970), pp. 321–334.

[9] R. Crabb, J. Losson, and M. C. Mackey, *Dependence on initial conditions in non local PDE's and hereditary dynamical systems*, in Proc. Internat. Conf. Nonlinear Anal. 4, de Gruyter, Berlin, 1996, pp. 3125–3136.
[10] J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, New York, 1960.
[11] P. Fortin and M. C. Mackey, *Periodic chronic myelogenous leukemia: Spectral analysis of blood cell counts and etiological implications*, Brit. J. Haematol., 104 (1999), pp. 336–345.
[12] A. C. Fowler and M. C. Mackey, *Relaxation oscillations in a class of delay differential equations*, SIAM J. Appl. Math., 63 (2002), pp. 299–323.
[13] J. K. Hale, *Theory of Functional Differential Equations*, Springer–Verlag, New York, 1977.
[14] C. Haurie, D. C. Dale, and M. C. Mackey, *Cyclical neutropenia and other periodic hematological diseases: A review of mechanisms and mathematical models*, Blood, 92 (1998), pp. 2629–2640.
[15] Y. Kuang, *Nonoccurrence of stability switching in systems of differential equations with distributed delays*, Quart. Appl. Math., LII(3) (1994), pp. 569–578.
[16] M. C. Mackey, *Unified hypothesis of the origin of aplastic anaemia and periodic hematopoïesis*, Blood, 51 (1978), pp. 941–956.
[17] M. C. Mackey, *Dynamic hematological disorders of stem cell origin*, in Biophysical and Biochemical Information Transfer in Recognition, J. G. Vassileva-Popova and E. V. Jensen, eds., Plenum Press, New York, 1979, pp. 373–409.
[18] M. C. Mackey, *Cell kinetic status of haematopoietic stem cells*, Cell Prolif., 34 (2001), pp. 71–83.
[19] M. C. Mackey and L. Glass, *From Clocks to Chaos: The Rhythms of Life*, Princeton University Press, Princeton, NJ, 1988.
[20] M. C. Mackey and J. Milton, *Feedback, delays, and the origins of blood cell dynamics*, Commun. Theor. Biol., 1 (1990), pp. 299–327.
[21] M. C. Mackey and A. Rey, *Bifurcations and travelling waves in a delayed partial differential equation*, Chaos, 2 (1992), pp. 231–244.
[22] M. C. Mackey and R. Rudnicki, *Global stability in a delayed partial differential equation describing cellular replication*, J. Math. Biol., 33 (1994), pp. 89–109.
[23] J. M. Mahaffy, J. Bélair, and M. C. Mackey, *Hematopoietic model with moving boundary condition and state dependent delay*, J. Theor. Biol., 190 (1998), pp. 135–146.
[24] A. A. Morley, A. G. Baikie, and D. A. G. Galton, *Cyclic leukocytosis as evidence for retention of normal homeostatic control in chronic granulocytic leukaemia*, Lancet, 2 (1967), pp. 1320–1322.
[25] G. C. Nooney, *Age distributions in dividing populations*, Biophys. J., 7 (1967), pp. 69–76.
[26] L. Pujo-Menjouet, S. Bernard, and M. C. Mackey, *Long Period Oscillations in a G_0 Model of Hematopoietic Stem Cells*, SIAM J. Appl. Dynam. Systems, 4 (2005), pp. 312–332.
[27] L. Pujo-Menjouet and M. C. Mackey, *Contribution to the study of periodic chronic myelogenous leukemia*, C. R. Biologies, 327 (2004), pp. 235–244.
[28] S. I. Rubinow, *A maturity time representation for cell populations*, Biophys. J., 8 (1968), pp. 1055–1073.
[29] S. I. Rubinow and J. L. Lebowitz, *A mathematical model of neutrophil production and control in normal man*, J. Math. Biol., 1 (1975), pp. 187–225.
[30] L. Sachs, *The molecular control of hemopoiesis and leukemia*, C. R. Acad. Sci. Paris, 316 (1993), pp. 882–891.
[31] M. Santillan, J. Bélair, J. M. Mahaffy, and M. C. Mackey, *Regulation of platelet production: The normal response to perturbation and cyclical platelet disease*, J. Theor. Biol., 206 (2000), pp. 585–603.
[32] L. F. Shampine and S. Thompson, *Solving DDEs in* MATLAB, Appl. Numer. Math., 37 (2001), 441–458; also available online at http://www.radford.edu/thompson/webddes/.
[33] E. Trucco, *Mathematical models for cellular systems: The Von Foerster equation*, Parts I and II, Bull. Math. Biophys., 27 (1965), pp. 285–304; 449–470.
[34] E. Trucco, *Some remarks on changing populations*, J. Ferm. Technol., 44 (1966), pp. 218–226.
[35] G. F. Webb, *Theory of Nonlinear Age-Dependent Population Dynamics*, Monogr. Textbooks Pure Appl. Math. 89, Marcel Dekker, New York, 1985.

SIAM J. APPL. MATH.
Vol. 65, No. 4, pp. 1353–1368

ON THE SOUND IN UNBOUNDED AND DUCTED VORTEX FLOWS*

L. M. B. C. CAMPOS† AND P. G. T. A. SERRÃO†

Abstract. The propagation of sound is considered in a potential cylindrical vortex, with superimposed axial flow, by means of explicit analytical solutions. The sound waves are sinusoidal in time and in the axial and azimuthal directions; the convected wave equation leads to a radial dependence specified by an ordinary second-order differential equation, with two singularities, at the origin and at infinity. Both singularities are irregular, implying that the acoustic fields have an essential singularity. In the neighborhood of the vortex axis, the essential singularity of the acoustic field is specified by an exponential of the integrated Doppler shift; using the latter as a factor, the acoustic fields are specified by asymptotic expansions in ascending powers of the radius. In the neighborhood of the point at infinity, where the tangential mean flow velocity vanishes, the leading terms are outward or inward propagating cylindrical waves; these factors multiply asymptotic expansions in descending powers of the radius. The two pairs of solutions, around the vortex axis and the point at infinity, are valid in all space or overlapping regions, as far as the asymptotic expansions can be calculated. The case of an annular nozzle, with uniform axial flow, and potential swirl is used as an example; the eigenvalues are obtained for rigid wall boundary conditions and the corresponding eigenfunctions are plotted.

Key words. acoustics, ducts, vortex

AMS subject classifications. 76Q05, 35C20, 33A70, 41A60, 34A20

DOI. 10.1137/S0036139903427076

1. Introduction. The propagation of sound in swirling flows [9, 7, 8, 12, 5] is relevant to the acoustics of turbomachinery and has important engineering applications in propulsion and power generation, e.g., jet engines and power turbines. The best known case is the rigid body rotation, for which the angular velocity is constant. It is of some interest to consider less simple cases, e.g., with radially varying angular velocity, of which the potential vortex is the only case, for which an acoustic potential exists, satisfying the convected wave equation [10, 1, 2]. The present paper concerns the exact solution of the convected wave equation in a cylindrical duct in the presence of a uniform axial flow with a superimposed potential vortex.

A vortex has a core where the tangential velocity remains finite, e.g., it should be sufficient to match the present solution to a small core with rigid body rotation. Taking the limit of zero core radius leads to a vortex with an algebraic singularity for the mean flow velocity and an essential singularity for the acoustic field. The latter singularity specifies the leading term of the acoustic field, viz., a Doppler factor on the vortex axis; at infinity, where the mean flow velocity vanishes, the leading term is a cylindrical wave. In both cases, viz., the solution around the vortex axis or the solution at infinity, the exact solution beyond the leading term is an asymptotic expansion; since the vortex is unstable, its perturbation could not be expected to lead to a convergent series. The summation of the asymptotic expansion is accurate for low azimuthal orders, and it shows that the leading terms dominate the acoustic fields.

*Received by the editors May 2, 2003; accepted for publication (in revised form) September 15, 2004; published electronically April 26, 2005.
http://www.siam.org/journals/siap/65-4/42707.html
†Secção de Mecânica Aeroespacial, ISR, Instituto Superior Técnico, 1049-001 Lisboa, Portugal (lmbcampos.aero@mail.ist.utl.pt, pserrao@dem.ist.utl.pt).

The acoustic wave equation in cylindrical coordinates in a potential vortex is harmonic in time and in the circumferential and axial coordinates and thus leads to a linear ordinary differential equation in the radial direction (section 2.1), which has two irregular singularities. The singularity on the vortex axis can be represented by a Doppler effect associated with the mean flow velocity (section 2.2); once this factor is inserted the exact solution can be obtained as an ascending power series (section 2). The mean flow velocity tends to uniform at infinity, and thus the limit of the cylindrical wave specifies the singularity there (section 3.1); inserting this factor leads to an asymptotic expansion for the acoustic field at large radius (section 3). The combination of the two solutions specifies the acoustics of nozzles with axial flow and swirl (section 3.2), including the eigenvalues (section 4.1) and the radial eigenfunctions (section 4.2). These are discussed for several axial and tangential Mach numbers, dimensionless frequencies, and circumferential wavenumbers.

2. Acoustic fields in the neighborhood of the vortex axis. The convected wave equation, which describes sound propagation in a potential vortex (appendix), has an irregular singularity on the vortex axis (section 2.1), because the mean flow velocity is infinite there. This corresponds to an essential singularity for the acoustic potential, which is specified by the integrated Doppler shift (section 2.2); using this singularity as a factor, a pair of linearly independent solutions is obtained as a power series of the radius.

2.1. Convected wave equation for a vortex flow. The acoustic potential Φ satisfies the convected wave equation

$$\left\{ c^{-2} \left(\partial/\partial t + \overrightarrow{V} \cdot \nabla \right)^2 - \nabla^2 \right\} \Phi \left(\overrightarrow{x}, t \right) = 0 \tag{2.1}$$

in a nonuniform, incompressible flow for which the sound speed c is constant. A cylindrical vortex corresponds to a potential flow if the circulation per unit length Γ is a constant, corresponding to the tangential velocity,

$$V_\theta = \Gamma/(2\pi r) = \gamma/r, \tag{2.2a}$$

$$\gamma \equiv \Gamma/2\pi, \tag{2.2b}$$

where γ is the circulation per radian. If a uniform axial velocity U is added, the mean flow velocity

$$\overrightarrow{V} = U \overrightarrow{e_z} + (\gamma/r) \overrightarrow{e_\theta} \tag{2.3}$$

is still potential, and the convected wave equation (2.1) becomes

$$\begin{aligned} &\{c^{-2}[\partial/\partial t + U\, \partial/\partial z + (\gamma/r^2)\partial/\partial\theta]^2 - r^{-1} \left(\partial/\partial r\right) r \left(\partial/\partial r\right) \\ &- r^{-2}\partial^2/\partial\theta^2 - \partial^2/\partial z^2\}\Phi(r,\theta,z,t) = 0 \end{aligned} \tag{2.4}$$

in cylindrical coordinates.

Since the mean flow is steady and the velocity depends only on r, it is convenient to use Fourier decompositions in t, θ, z, i.e., the acoustic potential is represented

$$\Phi(r,\theta,z,t) = \sum_{m=-\infty}^{+\infty} e^{im\theta} \iint_{-\infty}^{+\infty} d\omega dk\, e^{i(kz-\omega t)} \Psi(r;m,k,\omega) \tag{2.5}$$

by (i) a Fourier series in the azimuthal direction θ, with integral wavenumber m; (ii) a Fourier integral in time t, with frequency ω; (iii) a Fourier integral in the axial direction z, with a continuous wavenumber k spectrum for an infinite vortex (or a Fourier series with wavenumber $2\pi l/L$ for a vortex of length L, with l integer). Substitution of (2.5) into (2.3) leads to an ordinary differential equation for the radial r dependence of the acoustic potential spectrum

(2.6)
$$r^2\Psi'' + r\Psi' + \left\{ \left[(\omega - k\,U)^2 / c^2 - k^2 \right] r^2 - 2m\gamma\,(\omega - k\,U)/c^2 - m^2 + (m\gamma/cr)^2 \right\} \Psi = 0,$$

where appear the transverse wavenumber (2.7a)

$$K^2 \equiv (\omega - k\,U)^2 / c^2 - k^2, \tag{2.7a}$$

$$q^2 \equiv m^2 + 2m\gamma\,(\omega - k\,U)/c^2 \tag{2.7b}$$

and the azimuthal constant (2.7b). The differential equation

$$r^2\Psi'' + r\Psi' + \left[K^2 r^2 - q^2 + (m\gamma/c)^2 / r^2 \right] \Psi = 0 \tag{2.8}$$

is similar to a Bessel equation, except for the last term in the square brackets, which involves the circulation.

Introducing the dimensionless variable incorporating the transverse wavenumber,

$$s \equiv Kr, \tag{2.9a}$$

$$\Psi(r; m, k, \omega) \equiv F(s), \tag{2.9b}$$

leads to the differential equation

$$s^2F'' + sF' + \left(s^2 - q^2 + a^2/s^2 \right) F = 0, \tag{2.10}$$

similar to the Bessel equation of order q, apart from the term involving the constant:

$$a \equiv m\gamma K/c. \tag{2.11}$$

This constant vanishes ($a = 0$) in the absence of the vortex $\gamma = 0$, in which case $q = m$ in (2.7b) is the azimuthal wavenumber, and the acoustic potential which is finite on axis is specified by a Bessel function:

$$\gamma = 0: \quad F(s) = J_m(s) = J_m(K\,r). \tag{2.12}$$

The other solution is a Neumann function $Y_m(s) \sim \log s$ which has a logarithmic singularity on the axis. These two solutions apply in the absence of a vortex $\gamma = 0$, i.e., for $a = 0$, when the origin $s = 0$ is a regular singularity of the differential equation, i.e., the case of cylindrical waves. They also apply in the presence of the vortex $\gamma \neq 0$, but only to the axisymmetric mode [9] $m = 0$, since then $a = 0$ in (2.11). For nonaxisymmetric modes $m \geq 1$, and in the presence of the vortex $\gamma \neq 0$, then $a \neq 0$ in (2.11), and the differential equation (2.10) has an irregular singularity at the origin $s = 0$, implying that the solution has an essential singularity. This corresponds to a singular phase for the acoustic field on the vortex axis, which is specified by the integrated Doppler shift, as shown next.

2.2. Singular Doppler shift on the vortex axis. The Doppler shift of a sound wave of wave vector $\overrightarrow{K}$ in a mean flow of velocity $\overrightarrow{V}$ is given by an integral along the ray path dl, viz.,

$$\phi = \int \left[\left(\overrightarrow{K}.\overrightarrow{V}\right)\Big/c\right] dl = (m/c)\int V_\theta dr = m\,\gamma/cr = m\,\gamma\,K/c\,s = a/s, \tag{2.13}$$

where the mean velocity is tangential (2.2a), and thus multiplies the azimuthal wavenumber m, and (2.9a), (2.11) were used. Thus the acoustic potential should be of the form

$$F(s) = e^{\Omega(s)}G(s), \tag{2.14}$$

where $\Omega(s) = i\,\phi = ia/s$ is the singular phase term, corresponding to the essential singularity in the solution, and $G(s)$ should be an ascending power series. Note that the differential equation (2.10) will not have an ascending power series solution, unless the factor $e^{\Omega(s)}$ is inserted. A solution of the type (2.14) is called a normal integral [6], and it will be shown next that a solution of this type exists, with essential singularity $e^{\Omega(s)}$ specified $\Omega(s) = i\,\phi$ by the integrated Doppler shift (2.13). To prove this, (2.14) is substituted in the differential equation (2.10), leading to

$$s^2\,G'' + s\,[1 + 2\Omega' s]\,G' + \left[s^2\left(\Omega'' + \Omega'^2\right) + s\,\Omega' + s^2 - q^2 + a^2/s^2\right] G = 0. \tag{2.15}$$

The origin $s = 0$ will be a regular singularity, and $G(s)$ will have solution as an ascending power series of Frobenius–Fuchs type if the coefficients in square brackets are analytic at $s = 0$,

$$X_1 \equiv 2\,\Omega'\,s = O(1) = s^2(\Omega'^2 + \Omega'') + s\,\Omega' + a^2/s^2 \equiv X_2. \tag{2.16}$$

Thus, if $\Omega(s)$ can be found such that both X_1 and X_2 in (2.16) are analytic functions of s at $s = 0$, then $s = 0$ is a regular singularity of (2.15), and $G(s)$ is an ascending power series, which substituted in (2.14) specifies the normal integral as solution of (2.10).

It turns out that it is not necessary to make both X_1 and X_2 analytical at $s = 0$, but it is sufficient to choose $\Omega(s)$ so as to eliminate the most singular term, viz., the double pole a^2/s^2, by taking this to cancel with $s^2\,\Omega'^2$, viz.,

$$\Omega'(s) = \pm i\,a/s^2; \tag{2.17}$$

this implies

$$\Omega(s) = \mp i\,a/s \tag{2.18}$$

that, as predicted before, the integrated Doppler shift (2.13) appears in the normal integral (2.14) in the form

$$F_\pm(s) = e^{\mp i\,a/s}\,G_\pm(s), \tag{2.19}$$

where $G_\pm(s)$ satisfies the differential equation

$$s^2\,G''_\pm + s(1 \pm 2i\,a/s)G'_\pm + (s^2 - q^2 \mp i\,a/s)G_\pm = 0, \tag{2.20}$$

obtained by substituting (2.18) in (2.15).

The origin $s = 0$ would be a regular singularity of (2.20) if the coefficients in parentheses were analytic at $s = 0$, in which case two ascending Frobenius–Fuchs power series would exist for each of $G_+$and G_-. In fact both coefficients in parentheses in (2.20) have simple poles, but this is an improvement over the double pole in the coefficient in parentheses in (2.10). The latter has no solutions in power series, as could be found by trying the Frobenius–Fuchs method. In contrast, substituting a Frobenius–Fuchs series

$$G_\pm(s) = \sum_{J=0}^{\infty} g_j^\pm \, s^{j+\sigma_\pm} \tag{2.21}$$

in (2.20), it will be shown that one solution exists; it could not have two, so it could have none or one, and the latter is the case. This follows from the recurrence formula for the coefficients,

$$\left[(j+\sigma_\pm)^2 - q^2\right] g_j^\pm + g_{j-2}^\pm = \mp i\, a\, (2j + 2\sigma_\pm + 1)\, g_{j+1}^\pm, \tag{2.22}$$

which could be identical to that for Bessel functions in the absence of the vortex $a = 0$, and otherwise $a \neq 0$ is triple instead of double. Setting $j = -1$ leads to the indicial equation

$$j = -1 : \quad (2\sigma_\pm - 1) g_0^\pm = 0, \tag{2.23}$$

which has one root, corresponding to

$$\sigma_\pm = 1/2 : \quad \left[(j+1/2)^2 - q^2\right] g_j^\pm + g_{j-2}^\pm = \mp 2i\, a\, (j+1)\, g_{j+1}^\pm, \tag{2.24}$$

as recurrence relation for the coefficients. Thus (2.24), (2.21) specify one solution for each of the two differential equations (2.20), yielding in total the two linearly independent solutions (2.19) needed for the second-order differential equation (2.10).

The two Frobenius–Fuchs series (2.21),

$$g_0^\pm \equiv 1 : \quad G_\pm(s) = \sum_{j=0}^{\infty} g_j^\pm \, s^{j+1/2}, \tag{2.25}$$

multiplied (2.19) by the Doppler shift (2.13) specify the normal integrals

$$F_\pm(s) = e^{\mp i\, a/s} s^{1/2} \sum_{j=0}^{\infty} g_j^\pm \, s^j, \tag{2.26}$$

corresponding to the acoustic potentials

$$\Psi_\pm(r; m, k, \omega) = e^{\mp i\, m\gamma/c\, r} \sum_{j=0}^{\infty} g_j^\pm \, (Kr)^{j+1/2}, \tag{2.27}$$

which are linearly independent. The total acoustic potential is a linear combination

$$r < \infty : \quad \Psi(r; m, k, \omega) = A_+ \Psi_+(r; m, k, \omega) + A_- \Psi_-(r; m, k, \omega), \tag{2.28}$$

where the arbitrary constants of integration $A_\pm$ incorporate the coefficients $g_0^\pm$, which can thus be set to unity in (2.25). The two arbitrary constants of integration $A_\pm$ are determined by two conditions, i.e., (i) the acoustic potential at a given radius $\Psi(r_0; m, k, \omega)$; (ii) a radiation condition at infinity $r = \infty$. Since the solution (2.28) holds only for finite r, to apply the radiation condition, the solution of (2.10) around the point at infinity must be obtained.

3. Asymptotic acoustic fields and radiation condition. Besides the origin, the other singularity of the convected wave equation in a potential vortex is the point at infinity; the latter is also an irregular singularity of the differential equation, implying an essential singularity for the acoustic field, which is simply a cylindrical wave, since the mean flow velocity vanishes at infinity. Using the asymptotic cylindrical wave as a factor, two solutions of the convected wave equation are obtained as descending power series of the radius (section 3.1). Since these solutions represent inward and outward propagating waves, they allow the application of the Sommerfeld radiation condition. The pair of asymptotic solutions (section 3.1) overlaps with the pair of solutions around the vortex axis (section 3.2), allowing the application of boundary conditions for a potential vortex plus axial flow in a cylindrical or annular nozzle (section 3.2).

3.1. Sound radiation in a cylindrical vortex flow. The point at infinity $s = \infty$ is mapped to the origin $\zeta = 0$ using the inversion as a change variable,

$$\zeta = 1/s, \tag{3.1a}$$

$$F(s) = H(\zeta), \tag{3.1b}$$

which transforms the differential equation (2.10) to

$$\zeta^2 H'' + \zeta H' + (1/\zeta^2 - q^2 + a^2\zeta^2)H = 0. \tag{3.2}$$

It is clear from the double pole in the coefficient in parentheses that the origin $\zeta = 0$ is an irregular singularity of the differential equation (3.2), and thus the point at infinity $s = \infty$ or $r = \infty$ is an irregular singularity of (2.10) or (2.8). In this case the search for a solution as a normal integral is facilitated by noting that the vortex does not produce a mean flow at infinity, i.e., a in (3.2) does not affect the singularity at $\zeta = 0$. This means that the acoustic potential at infinity consists of cylindrical waves propagating inward and outward, implying solutions of the form $\exp(\pm iKr) = \exp(\pm is) = \exp(\pm i/\zeta)$; this specifies the essential singularity in the normal integral,

$$H_\pm(\zeta) = e^{\pm i/\zeta} J_\pm(\zeta), \tag{3.3}$$

which substituted in (3.2) leads to

$$\zeta^2 J_\pm'' + (\zeta \mp 2i)\, J_\pm' + \left(i/\zeta - q^2 + a^2\zeta^2\right) J_\pm = 0. \tag{3.4}$$

Note that although the origin $\zeta = 0$ is not a regular singularity of the differential equation (3.4), because there is a simple pole in the second coefficient in parentheses, this is better than the double pole in the coefficient in parentheses in (3.2).

It can be checked by substituting a Frobenius–Fuchs series in (3.2) that no solution in ascending power series exists. Substituting in (3.4) a Frobenius–Fuchs series,

$$J_\pm(\zeta) = \sum_{l=0}^{\infty} j_l^\pm\, \zeta^{l+\vartheta_\pm}, \tag{3.5}$$

leads to the recurrence formula for the coefficients

$$\left[(l+\vartheta_\pm)^2 - q^2\right] j_l^\pm \mp (2l + 2\vartheta_\pm + 1)\, i\, j_{l+1}^\pm + a^2 j_{l-2}^\pm = 0 \tag{3.6}$$

and indicial equation

$$l = -1: \quad (2\vartheta_\pm - 1)\, j_0^\pm = 0, \tag{3.7}$$

which has one root:

$$\vartheta_\pm = 1/2: \quad \left[(l+1/2)^2 - q^2\right] j_l^\pm \mp 2(l+1) i\, j_{l+1}^\pm + a^2 j_{l-2}^\pm = 0. \tag{3.8}$$

This specifies (3.8), (3.5) one solution for each of the differential equations (3.4), and thus in total two solutions (3.3) for the differential equation (3.2). These solutions represent inward and outward propagating waves at infinity and thus allow the application of a radiation condition. By matching the solutions around the point at infinity with the solutions around the origin (2.27), other types of boundary conditions may be applied, e.g., for a vortex in a cylindrical duct.

Before proceeding to consider several kinds of boundary conditions, it is convenient to make explicit the two solutions at infinity, corresponding to outward and inward propagating waves. The power series (3.5),

$$j_0^\pm \equiv 1: \quad J_\pm(\zeta) = \sum_{l=0}^{\infty} j_l^\pm\, \zeta^{l+1/2}, \tag{3.9}$$

multiplied by the phase term (3.3),

$$H_\pm(\zeta) = e^{\pm i/\zeta} \sum_{l=0}^{\infty} j_l^\pm\, \zeta^{l+1/2}, \tag{3.10}$$

specify the acoustic potential around (3.1a), (3.1b) the point-at-infinity,

$$F^\pm(s) = e^{\pm is} \sum_{l=0}^{\infty} j_l^\pm\, s^{-l-1/2}, \tag{3.11}$$

which corresponds,

$$\Psi^\pm(r; m, k, \omega) = e^{\pm iKr} \sum_{l=0}^{\infty} j_l^\pm\, (Kr)^{-l-1/2}, \tag{3.12}$$

to, respectively, outward Ψ^+ and inward Ψ^- propagating waves.

3.2. Application to nozzles with axial flow and swirl. The total acoustic potential is a linear combination of the two,

$$r > 0: \quad \Psi(r; m, k, \omega) = B_+ \Psi^+(r; m, k, \omega) + B_- \Psi^-(r; m, k, \omega), \tag{3.13}$$

where the arbitrary constants of integration $B_\pm$ incorporate the coefficient $j_0^\pm$, which may be set to unity (3.9). The outward Ψ^+ and inward Ψ^- propagating acoustic potentials for all distances except for duct axis $r > 0$ are a linear combination of the acoustic potential $\Psi_\pm$ valid (2.27) at all finite distances $r < \infty$, in the region of overlap,

$$0 < r < \infty: \quad \Psi^\pm(r; m, k, \omega) = D_+^\pm \Psi_+(r; m, k, \omega) + D_-^\pm \Psi_-(r; m, k, \omega), \tag{3.14}$$

where the constants $D_{\pm}^{\pm}$ can be determined at any point. The radiation condition, specifying outward propagating waves Ψ^+ at infinity, requires selecting the first term of the right-hand side of (3.13), viz.,

$$r > 0: \quad \Psi(r;m,k,\omega) = B_+ \Psi^+(r;m,k,\omega), \tag{3.15a}$$

$$r < \infty: \quad \Psi(r;m,k,\omega) = B_+ [D_+^+ \Psi_+(r;m,k,\omega) + D_-^+ \Psi_-(r;m,k,\omega)]. \tag{3.15b}$$

The remaining constant of integration is determined,

$$0 < r_0 < \infty: \quad \Psi(r_0;m,k,\omega) = B_+ \Psi^+(r_0;m,k,\omega), \tag{3.16}$$

from the wave field at a given position. All these results concern acoustic propagation in an unbounded cylindrical vortex with axial flow.

It is also possible to confine the vortex into a cylindrical duct of radius R. In this case the problem concerns the acoustic modes in a cylindrical duct, with (2.3) uniform axial flow and a vortex on its axis with constant circulation Γ per unit length. The two solutions (2.27) around the duct vortex axis cover the whole duct. They are complex conjugates, as can be seen from (2.27), (2.24),

$$\Psi_+^* = \Psi_-: \quad \Psi_n(r;m,k,\omega) = A \operatorname{Re} \left\{ e^{\mp i\, m\gamma/c\, r} \sum_{j=0}^{\infty} g_j^{\pm} (Kr)^{j+1/2} \right\}, \tag{3.17}$$

and thus a real acoustic potential is obtained by taking the real part of either of them, which is equivalent to choosing $A_+ = A_- = 2A$ in (2.28). The eigenfrequencies are given by (2.7a), viz.,

$$\omega_n = KU + c\sqrt{K^2 + K_n^2}, \tag{3.18}$$

in terms of the radial wavenumbers K_n, which also specify the eigenfunctions (3.17). The wavenumbers K_n are determined by a boundary condition at the duct wall $r = R$. The simplest is a rigid wall boundary condition

$$0 = \partial\Psi/\partial r|_{r=R} = \operatorname{Re} \left\{ e^{\frac{im\gamma}{cR}} (KR)^{-1/2} K\, F(K) \right\}, \tag{3.19}$$

where

$$F_0 \prod_{n=0}^{\infty} (K - K_n) \equiv F(K) \equiv \sum_{j=0}^{\infty} g_j (KR)^j (j + 1/2 - im\gamma/cR); \tag{3.20}$$

thus the roots of (3.20) specify the eigenvalues K_n for the transverse wavenumber, substitution in (3.18) specifies the eigenfrequencies, and substitution in (3.17) specifies the eigenfunctions.

For an axial vortex and a uniform mean flow confined in an annular duct, the acoustic potential is a linear combination of the two solutions at infinity,

$$r_i < r < r_e: \quad \Psi_n(r;m,k,\omega) = B_+ \Psi_n^+(r;m,k,\omega) + B_- \Psi_n^-(r;m,k,\omega), \tag{3.21}$$

which cover the annular duct between inner radius r_i and outer radius r_e. In this case the problem concerns acoustic modes in an annular duct with potential swirling flow with eigenfunctions

$$\Psi_n^{\pm}(r;m,k,\omega) = e^{\pm iK_n r} \sum_{l=0}^{\infty} j_l^{\pm} (K_n r)^{-l-1/2}, \tag{3.22}$$

and radial eigenvalues K_n related to axial eigenvalues k_n in a dispersion relation,

$$K_n^2 = (\omega - k\,U)^2/c^2 - k_n^2, \tag{3.23}$$

where n represents the integer radial modal number.

The wavenumbers K_n are determined from the two rigid wall conditions

$$\partial\Psi/\partial r|_{r=r_i} = 0, \quad \partial\Psi/\partial r|_{r=r_e} = 0, \tag{3.24}$$

leading to

$$\begin{vmatrix} \Psi_n^{+\prime}(r_i; m, k_n, \omega) & \Psi_n^{-\prime}(r_i; m, k_n, \omega) \\ \Psi_n^{+\prime}(r_e; m, k_n, \omega) & \Psi_n^{-\prime}(r_e; m, k_n, \omega) \end{vmatrix} = 0. \tag{3.25}$$

4. Application to the acoustics of an annular nozzle with a potential swirl. The preceding results can be used to calculate the eigenvalues (section 4.1) and eigenfunctions (section 4.2) for sound in an annular nozzle containing an axially uniform mean flow, on which is superimposed a potential vortex swirl.

4.1. Effect of axial and swirling flow on radial wavenumbers. As an application an annular duct with inner radius $r_i = 4$ and outer radius $r_e = 6$ is considered. The dimensionless frequency $\varpi = \omega r_m/c$ is introduced, where $r_m = (r_i + r_e)/2$ is the mean radius. The calculations were performed for a dimensionless frequency $\varpi = 2.5$, an axial Mach number $M_z = 0.3$, and a tangential Mach number $M_\theta = 0.5$ specified at the mean radius r_m. The axial wavenumbers k_n as given by (3.23), (3.25) are represented in Figure 1 for the lowest integral wavenumber $m = 1, -1$. As m increases so does a in (2.11), and the asymptotic expansion (3.12), (3.13) deteriorates in accuracy. For the first circumferential mode $|m| = 1$ rotating in the same direction as the swirl of the mean flow $m = +1$, the axial wavenumbers are complex conjugate pairs,

$$k_n = \alpha_n \pm i\,\beta_n, \quad \exp(i\,k_n\,z) = \exp(i\,\alpha_n\,z)\exp(\mp\beta_n z), \tag{4.1}$$

with a small positive real part α_n almost independent of n representing propagation; in this case $m = +1$ of corotation the imaginary part increases with n, corresponding to a decay $+i\beta_n$ or instability $-i\beta_n$. In the case $m = -1$ of counterrotation, the first eigenvalue is real and of negative sign representing a wave of constant amplitude, propagating in the negative z-direction; the higher-order modes $n = 2, 3, 4$ are again complex conjugate pairs, with imaginary parts increasing with n, implying spatial decay or instability, as in the case of corotation. Also as in the case of corotation the real part α_n is almost independent of $n = 2, 3, \ldots$, but in contrast it has small negative (instead of positive) values $\alpha_n < 0$, implying propagation against the axial mean flow, in the negative z-direction. The larger propagation speed of corotating relative to counterrotating modes corresponds to the fast and slow acoustic-vortical modes in the axisymmetric case [9].

To each eigenvalue k_n, specified by a root of (3.25), corresponds an eigenfunction (3.21) with zero radial derivative at the wall. Each eigenfunction is a linear combination of inward and outward propagating waves (3.22) with amplitudes $B^\pm$. Since the rigid wall boundary conditions (3.24) only determine the ratios of amplitudes B^+/B^-, the choice $B^+ = 1$ is made for the plots in Figures 2 to 5. Each wave consists (3.12) of a sinusoidal radial oscillation multiplying an asymptotic expansion (3.9). Note that if the point at infinity were a regular singularity, the solution would

be a series expansion, whose convergence is assured by the Fuchs theorem. Since, in the present case, the point at infinity is an irregular singularity, the Fuchs theorem does not apply, and the Frobenius–Fuchs method leads to an asymptotic expansion (3.9), whose convergence or accuracy is not assured a priori. The asymptotic expansion was summed as long as the terms decreased; it was truncated when the sum of the absolute value of the last 10 terms did not exceed 10^{-7}. This criterion is quite demanding and is similar to that used for convergent series [3]. In the cases where this criterion was not met, the asymptotic expansion was summed up to the smallest term; if this term was less than 10^{-2} of the sum, the result was deemed acceptable. Otherwise, the result was discarded. Our choice of the inner and outer radius, mean flow, and swirl Mach numbers and dimensionless frequency coincides with [7, 8], whose authors have considered the same problem numerically; we have chosen to represent in Figure 1 the modes for $m = \pm 1$ and increasing $n = 1, 2, 3, 4, \ldots$; [7, 8] do not plot the waveforms in this case. For some of their other choices of dimensionless frequency and Mach number our asymptotic expansions do not meet the acceptance criteria. Thus an extensive comparison is not possible. The vortex flow $v_\theta = \gamma/r$ is unstable [11, 4], and therefore it could be expected that the solutions of the wave equation be asymptotic expansions rather than convergent series.

4.2. Radial eigenfunctions for acoustic modes in an annular nozzle. The eigenfunctions are plotted in Figures 2 and 3 for three radial modes. Mode $n = 1$ was excluded from Figure 2 plots by the accuracy criterion. All the eigenmodes in Figures 2 and 3 have zero slopes at the rigid walls; the number of zeros of the amplitude (left-hand side) increases with the order n, and the phases change sign at the zeros. In cases of both corotation (Figure 2) and counterrotation (Figure 3) the nth mode has $n - 1$ zeros. The zeros in the eigenfunctions (3.22) are due to the sinusoidal factor with unit amplitude, because the asymptotic expansions (3.9) are slowly varying monotonic functions in all cases shown in Figures 4 and 5. The asymptotic expansions for outward J_+ and inward J_- propagation in the case of corotation are shown in Figure 4: (i) in the case of outward propagation J_+ there is a decreasing amplitude for $n = 2, 3, 4$, and a phase of opposite sign (positive) for $n = 2, 4$ and $n = 3$ (negative); (ii) in the case of inward propagation J_- there is a small amplitude and phase, the latter being negative only for $n = 3$. In the case of counterrotation (Figure 5) the outward propagating asymptotic expansion J_+ has larger amplitude than the inward propagating asymptotic expansion J_- for $n = 2, 3$, with the exception of $n = 1$, when the amplitudes of $J_\pm$ are comparable. The phases are small and positive in all cases except for $n = 2, 3$ in outward propagation, when they are negative.

The phase corrections introduced by the asymptotic expansions (3.9) in the waveforms (3.22) are small relative to the sinusoidal factor; since the latter has unit modulus, the amplitude is determined by the modulos of the asymptotic expansions (3.9), which are dominant and all-important in this respect. The original eigenfunctions Ψ_n in (3.22) are plotted in Figures 2 and 3 as sets of six panels ($|\Psi_n|$ right-hand side, $\arg(\Psi_n)$ left-hand side) for integral wavenumber $m = 1, -1$. Removing the exponential leading term from the eigenfunctions $\Psi_n^\pm$ in (3.22) leaves the reduced eigenfunctions $J_\pm$ specified by the power series (3.8), (3.9) which are plotted in Figures 4 and 5 using the same reference values. Having shown how the eigenvalues and eigenfunctions can be obtained for the radial acoustic modes in an annular nozzle with uniform axial flow and potential vortex swirl, the main features of the method, the results are now reviewed and discussed.

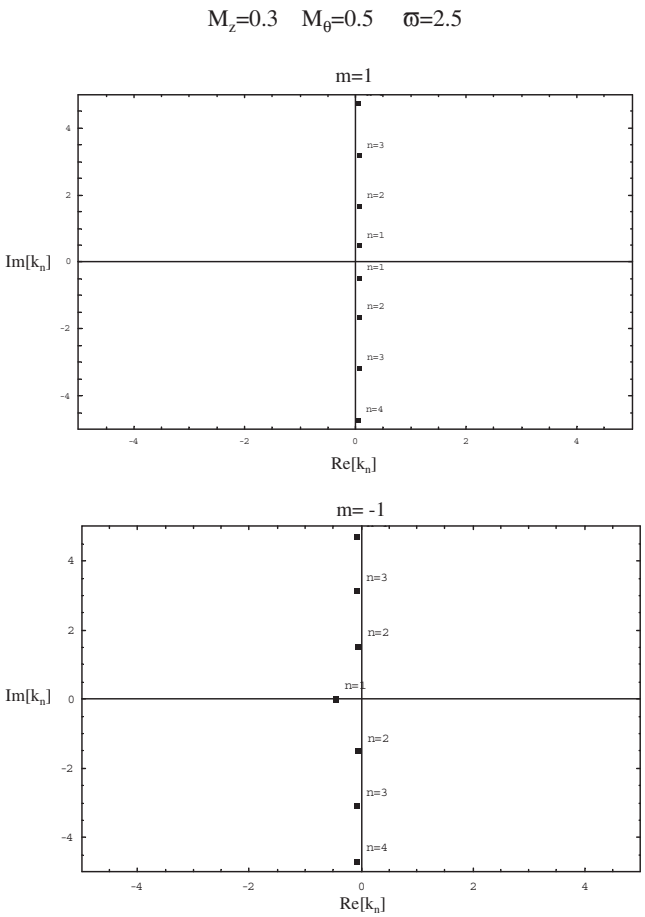


FIG. 1. *Axial eigenvalues for integral wavenumber* $m = 1$ *(top),* $m = -1$ *(bottom).*

5. Discussion. The problem considered is that of superposition of a line source of sound and a potential flow due to a coincident line vortex. It is well known that a line acoustic source generates a cylindrical wave, whose amplitude would be infinite on the axis; in reality dissipation limits the amplitude near the axis, but the singularity still determines the asymptotic decay of the amplitude of the cylindrical wave at large distances, like the inverse square root $1/\sqrt{r}$ of radial distance r. Likewise, a potential vortex would lead to an infinite velocity on the axis, which is excluded by matching a vortex core, e.g., in rigid body rotation; however the singularity on the vortex axis determines the asymptotic decay of the tangential velocity, at large distance from the vortex core, as the inverse of the radius $1/r$. The coincidence of a line-source sound and a line-vortex leads to an essential singularity because there are infinite phase oscillations near the axis, and thus the nondissipative solution is valid only in an annulus, whose inner radius is not too small. Despite this the nature of the essential singularities of the wave field, both at the origin and at infinity, still affects the acoustic field at all finite radial distances. The acoustic field can thus be represented at all finite radial distances, either by a linear combination of a pair of solutions $\Psi_\pm$ around the origin $r = 0$ or by a pair of solutions $\Psi^\pm$ around the point at infinity $r = \infty$.

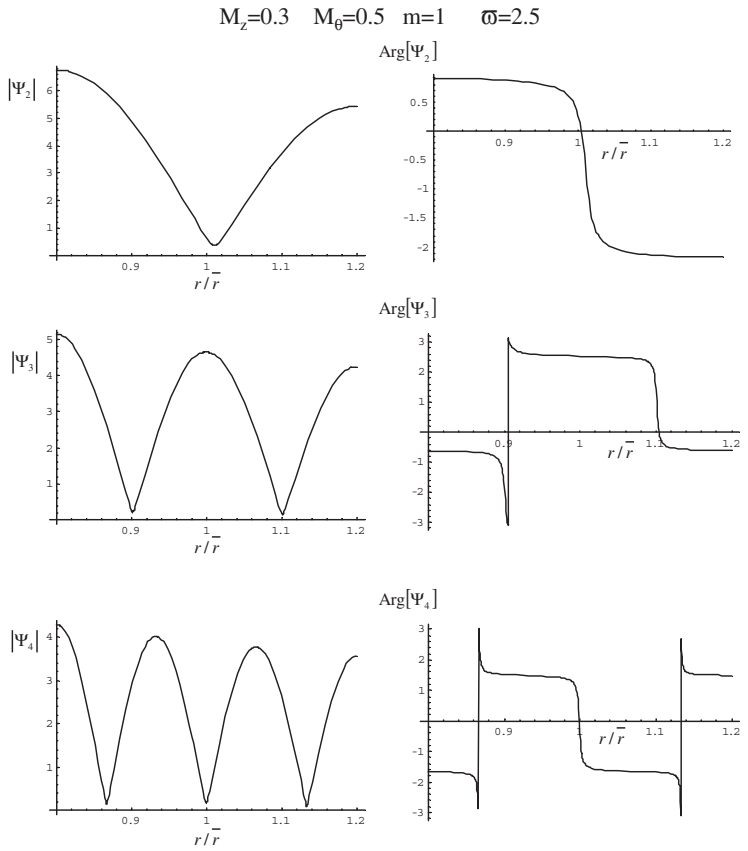


FIG. 2. *Amplitude (left) and phase (right) of the acoustic potential eigenfunctions* Ψ_n *radial modes* $n = 2$ *(top),* $n = 3$ *(middle),* $n = 4$ *(bottom) for integral azimuthal wavenumber* $m = 1$.

Since the wave equation is a differential equation of the second order, it can have only two linearly independent solutions; thus either function in the pair (Ψ_-, Ψ_+) and (Ψ^-, Ψ^+) is a linear combination of the other pair, e.g., $\Psi_- = C_-^- \Psi^- + C_-^+ \Psi^+$. The constant coefficients C_-^-, C_-^+ can be determined by matching the solutions at any two points in their common region of validity. Thus the consideration of the two essential singularities at the origin and infinity allows an exact analytical solution to be obtained, as an alternative to the numerical methods in the literature. The identification of the nature of the singularity is not a purely mathematical problem and is, in fact, guided by physical considerations: (i) the singularity at the origin is due to the integrated Doppler shift associated with the mean flow velocity; (ii) since at large distance the mean flow velocity is small the acoustic field must tend to a cylindrical wave. These simple considerations are used to obtain the exact analytical solutions of the problem of acoustic propagation in a potential vortex flow.

Appendix. Deduction of the convected wave equation for sound in a vortex with constant circulation. Consider a mean flow consisting of a uniform axial velocity U and a rotation with angular velocity $\Omega(r)$, so that the mean flow

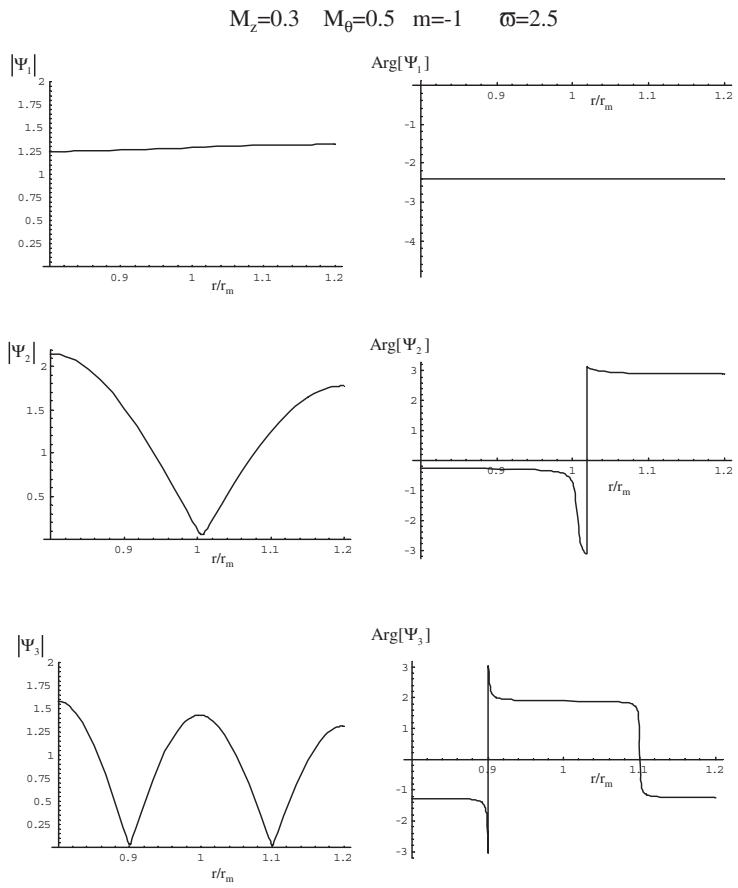


FIG. 3. *Amplitude (left) and phase (right) of the acoustic potential eigenfunctions* Ψ_n *radial modes* $n = 1$ *(top),* $n = 2$ *(middle),* $n = 3$ *(bottom) for integral azimuthal wavenumber* $m = -1$.

velocity is specified in cylindrical coordinates by

$$V_r = 0, \tag{A.1a}$$

$$V_z = U, \tag{A.1b}$$

$$V_\theta = r\,\Omega(r). \tag{A.1c}$$

The curl of the velocity or vorticity of the mean flow

$$\nabla \wedge \overrightarrow{V} = \overrightarrow{e_z}\frac{1}{r}\frac{\partial}{\partial r}\,(rV_\theta) = \overrightarrow{e_z}\frac{1}{r}\frac{d}{dr}\left[r^2\Omega(r)\right] \tag{A.2}$$

lies in the axial direction and vanishes only if the angular velocity varies like the inverse square of the radius (A.3a),

$$\nabla \wedge \overrightarrow{V} = 0: \quad \Omega(r) \sim r^{-2}, \tag{A.3a}$$

$$V_\theta \sim r^{-1}, \tag{A.3b}$$

which implies that the tangential velocity varies inversely with radius (A.3b), and thus the circulation Γ or γ is constant (2.2a), (2.2b). Note that from (A.1a), (A.1b),

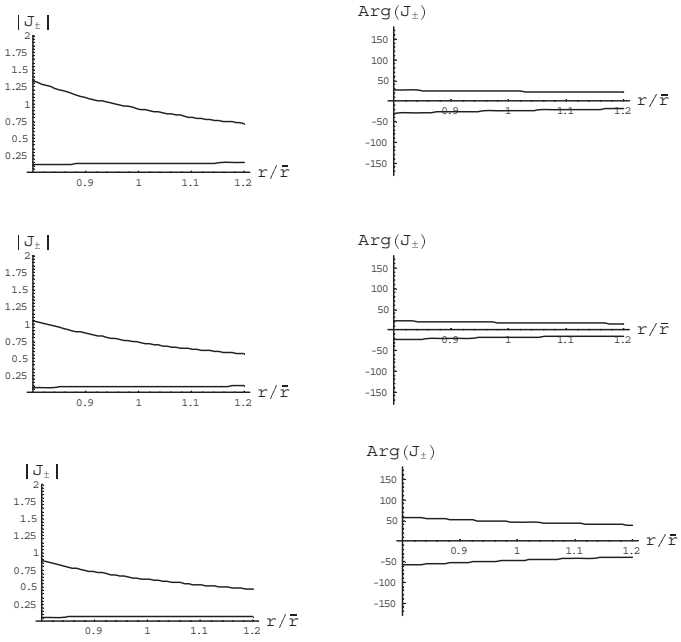


FIG. 4. *Power series outward J_+ and inward J_- for integral wavenumber $m = 1$. Radial mode $n = 2$ (top), $n = 3$ (middle), $n = 4$ (bottom).*

(A.1c) it follows that $\nabla . \overrightarrow{V} = 0$ for arbitrary uniform axial velocity U and angular velocity $\Omega(r)$, so that the mean flow is incompressible; it will be shown that in this case the propagation of sound is specified by the convected wave equation provided that the angular velocity $\Omega \sim r^{-2}$ corresponds to a potential vortex of any strength γ or Γ in (2.2a), (2.2b).

For a homentropic, inviscid flow, the linearized vorticity equation reads

$$\partial \left(\nabla \wedge \overrightarrow{v} \right) / \partial t - \nabla \wedge \left[\overrightarrow{V} \wedge \left(\nabla \wedge \overrightarrow{v} \right) \right] - \nabla \wedge \left[\overrightarrow{v} \wedge \left(\nabla \wedge \overrightarrow{V} \right) \right] = 0, \tag{A.4}$$

where $\overrightarrow{v}$ is the perturbation velocity. If the mean flow is irrotational,

$$\nabla \wedge \overrightarrow{V} = 0 : \quad \partial \left(\nabla \wedge \overrightarrow{v} \right) / \partial t + \nabla \wedge \left[\overrightarrow{V} \wedge \left(\nabla \wedge \overrightarrow{v} \right) \right] = 0, \tag{A.5}$$

a solution is that the velocity perturbation is irrotational (A.6),

$$\nabla \wedge \overrightarrow{v} = 0 \Rightarrow \overrightarrow{v} = \nabla \Phi, \tag{A.6}$$

and thus an acoustic potential exists in nondissipative conditions. Since both mean flow velocity and velocity perturbation are irrotational, the Bernoulli equation for an incompressible or compressible fluid reads (A.7a)

$$p = -\rho \frac{d\Phi}{dt}, \tag{A.7a}$$

$$d/dt \equiv \partial / \partial t + \overrightarrow{V} \cdot \nabla, \tag{A.7b}$$

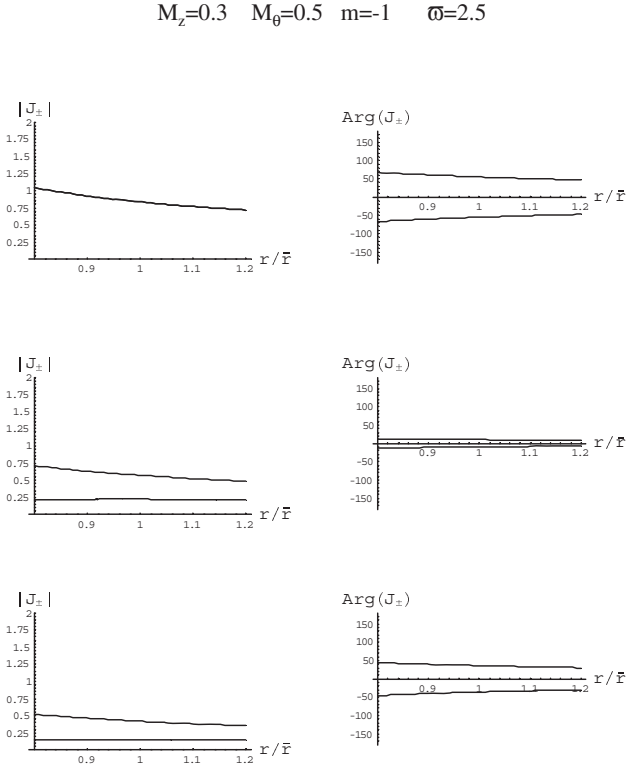


FIG. 5. *Power series outward J_+ and inward J_- for integral wavenumber $m = -1$. Radial mode $n = 1$ (top), $n = 2$ (middle), $n = 3$ (bottom).*

where p is the pressure perturbation, ρ is the constant mean flow density, and the linearized material derivative (A.7b) involves only the mean flow velocity.

The linearized equation of continuity is

$$d\rho'/dt + \rho\nabla \cdot \overrightarrow{v} = 0, \tag{A.8}$$

where the density perturbation is related to the pressure perturbation by the adiabatic condition,

$$dp/dt = c^2 d\rho'/dt, \tag{A.9}$$

where the sound speed is constant for an incompressible mean flow. Substitution of (A.9) in (A.8) yields

$$c^{-2} dp/dt + \rho\nabla \cdot \overrightarrow{v} = 0. \tag{A.10}$$

Substitution of (A.6), (A.7a) leads to the convected wave equation for the acoustic potential,

$$c^{-2} d^2\Phi/dt^2 - \nabla^2\Phi = 0, \tag{A.11}$$

in agreement with (2.3).

The swirl in the mean flow causes a centrifugal force, compensated by a radial pressure gradient,

$$dp_0/dr = \rho \left[V_\theta (r)\right]^2 / r = \rho\gamma^2/r^3; \tag{A.12}$$

thus the pressure in the mean flow is given by

$$p_0(r) = p_\infty + \rho\gamma^2/(2r^2), \tag{A.13}$$

where p_∞ is the pressure at infinity. Denoting by $\overline{\gamma}$ the ratio of specific heats, since the mass density is constant, the sound speed

$$\left[c\,(r)\right]^2 = \overline{\gamma} p(r)/\rho = c_\infty^2 + \overline{\gamma}\gamma^2/(2r^2) \tag{A.14}$$

is approximately constant, and equal to the sound speed at infinity, if

$$c_\infty^2 \equiv \overline{\gamma} p_\infty/\rho = \left[c\,(r)\right]^2 : \quad r^2 \gg \overline{\gamma}\gamma^2/(2c_\infty^2). \tag{A.15}$$

Thus the solution of the wave equation in the text, which assumed constant sound speed, is valid only at some distance from the axis. The singularity of the wave equation at the origin, although outside the physical region of interest, remains important, because it affects the wave field for finite radius, beyond (A.15).

Acknowledgment. The authors are grateful for the comments of the two referees, which helped improve the paper.

REFERENCES

[1] L. M. B. C. Campos, *On the emission of sound by an ionized inhomogeneity*, Proc. Roy. Soc. London Ser. A, 359 (1978), pp. 65–91.
[2] L. M. B. C. Campos, *On waves in gases.* I. *Acoustics of jets, turbulence, and ducts*, Rev. Modern Phys., 58 (1986), pp. 117–182.
[3] L. M. B. C. Campos and M. Kobayashi, *On the reflection and transmission of sound in a thick shear layer*, J. Fluid Mech., 420 (2000), pp. 1–24.
[4] S. Chandrasekhar, *Hydrodynamic and Hydromagnetic Stability*, Oxford University Press, London, 1961.
[5] A. Cooper and N. Peake, *Trapped acoustic modes in aeroengine intakes with swirling flow*, J. Fluid Mech., 419 (2000), pp. 151–175.
[6] A. Forsyth, *Theory of Differential Equations*, Cambridge University Press, Cambridge, UK, 1902–1904.
[7] V. Gobulev and H. Atassi, *Sound propagation in an annular duct with mean potential swirling flow*, J. Sound Vib., 198 (1996), pp. 601–606.
[8] V. Gobulev and H. Atassi, *Acoustic-vorticity waves in swirling flows*, J. Sound Vib., 209 (1998), pp. 203–222.
[9] M. Howe and J. Liu, *The generation of sound by vorticity waves in swirling flows*, J. Fluid Mech., 81 (1977), pp. 369–383.
[10] M. S. Howe, *Contributions to the theory of aerodynamic sound, with applications to excess jet noise and the theory of the flute*, J. Fluid Mech., 71 (1975), pp. 625–673.
[11] C. Lin, *Theory of Hydrodynamic Stability*, Cambridge University Press, Cambridge, UK, 1955.
[12] C. Tam and L. Auriault, *The wave modes in ducted swirling flows*, J. Fluid Mech., 419 (1998), pp. 151–175.

SIAM J. APPL. MATH.
Vol. 65, No. 4, pp. 1369–1387

ON A REGULARIZATION SCHEME FOR LINEAR OPERATORS IN DISTRIBUTION SPACES WITH AN APPLICATION TO THE SPHERICAL RADON TRANSFORM*

THOMAS SCHUSTER† AND ERIC TODD QUINTO‡

Abstract. This article provides a framework to regularize operator equations of the first kind where the underlying operator is linear and continuous between distribution spaces, the dual spaces of smooth functions. To regularize such a problem, the authors extend Louis' method of approximate inverse from Hilbert spaces to distribution spaces. The idea is to approximate the exact solution in the weak topology by a smooth function, where the smooth function is generated by a mollifier. The resulting regularization scheme consists of the evaluation of the given data at so-called reconstruction kernels which solve the dual operator equation with the mollifier as right-hand side. A nontrivial example of such an operator is given by the spherical Radon transform which maps a function to its mean values over spheres centered on a line or plane. This transform is one of the mathematical models in sonar and radar. After establishing the theory of the approximate inverse for distributions, we apply it to the spherical Radon transform. The article also contains numerical results.

Key words. spherical Radon transform, sonar, distribution, regularization, approximate inverse, mollifier, reconstruction kernel

AMS subject classifications. 44A12, 45A05, 46F12

DOI. 10.1137/S003613990343879X

1. Introduction.

We apply the method of approximate inverse to the problem of reconstructing a function from integrals over spheres. Applications of this mathematical problem include sonar when the source and detector are at the same point [15], thermoacoustic tomography for cancer detection [14], seismic testing [23], and radar. The article [5] provides an excellent introduction to synthetic aperture radar and the relation between spherical integrals and radar and sonar.

The approximate inverse was originally developed by Louis as a general method to regularize ill-posed operators on Hilbert spaces [17]. It has been applied to integral equations of the first kind [18] and tomography [27, 28]. However, the inversion formula for our problem is valid not on Hilbert spaces but on distributions. Therefore, we will generalize the approximate inverse to the setting of distributions. It is hoped this generalization will be useful for other inverse problems for which the ambient spaces are not Hilbert spaces.

In seismology or sonar the acoustic wave equation is

$$n^2(x)u_{tt} = \Delta u + \delta(t)\delta(x - a_0), \text{ where } a_0 \in A,$$

and A is a small section of the surface of the earth. After linearization, the determination of $n^2(x)$ from back-scattered data is equivalent to recovering n^2 from integrals

*Received by the editors December 15, 2003; accepted for publication (in revised form) September 14, 2004; published electronically April 26, 2005. This material is based on work supported by the National Science Foundation under NSF grant DMS 0200788 and support from Tufts University FRAC.

http://www.siam.org/journals/siap/65-4/43879.html

†Tufts University, Department of Mathematics. Permanent address: Institut für Angewandte Mathematik, Universität des Saaleandes, 66041, Saarbrücken, Germany (thomas.schuster@num.uni-sb.de). The research of this author was partially supported by a Feodor Lynen Fellowship of the Alexander von Humboldt Foundation under V-3.FLF-DEU/1073550.

‡Tufts University, Department of Mathematics, 503 Boston Avenue, Medford, MA 02155 (todd.quinto@tufts.edu).

over spheres with centers on A [15]. Knowing n^2 or at least an approximation to n^2 can show boundaries of objects in the water. This linearized model is reasonable from a practical standpoint when the speed of sound in the ambient water is fairly constant. This would occur in water of depth less than 100 feet with fairly constant temperature [3]. Since the speed of sound is constant in shallow water with constant temperature, a pulse travels from a point source, a, making a spherical wavefront. The sound that is reflected back to the source at time t gives the amount reflected back from the sphere centered at a and radius $t/2$ times the speed of sound (assuming no multiple reflections). See [12] for practical information about sonar.

The mathematical problem can be described as trying to recover a function by its integrals over all spheres centered on a given line (in $\mathbb{R}^2$), plane (in $\mathbb{R}^3$), or hyperplane (in $\mathbb{R}^n$).

We first discuss the inversion methods that have been implemented numerically and then the pure mathematical results behind them. Denisjuk has an inversion method based on a transformation that changes the spherical transform into a limited data line transform [9]. He has implemented his method with good results. Klein [13] has developed and numerically tested a promising inversion method based on the ideas of Andersson discussed below. Beltukov proposed a numerical inversion method using a discrete SVD for the sonar transform [4]. He showed that the singular values are fairly flat and then drop off precipitously, which reflects the ill-posedness of the problem.

Our numerical reconstructions are given in section 6 and they show the potential of our method.

Many authors have proven injectivity and inversion methods for this transform. Courant and Hilbert [7, p. 699] proved injectivity for functions that are even about the hyperplane. Fawcett [10] and Andersson [2] provide inversion formulas in $\mathbb{R}^n$. Norton provides an inversion method for the circular transform if the center set is a circle in the plane [22] and if the center set is a line [21], and [23] gives three-dimensional results. Ranges and inversion formulas on a subspace of Schwartz functions are given in [20].

Finch, Patch, and Rakesh [11] develop an explicit inversion formula for recovering a function from spherical integrals when the center set is the boundary of a bounded, connected, open set in $\mathbb{R}^n$. Ramm proves injectivity and inversion theorems in [26]. Fairly general uniqueness theorems are given in [1].

Louis and Quinto [19] develop the microlocal analysis of the transform when A is a real-analytic surface (e.g., an open subset of a hyperplane), and they prove the local transform is injective under fairly general hypotheses. They characterize singularities (jumps, etc.) of the object that are stably visible from the data. Palamodov [24] and Denisjuk [8] continue this microlocal analysis when S is a hyperplane, providing instability results, inversion methods, and range theorems. Beltukov has proven an inversion method for the transform on hyperbolic space.

Section 2 contains the extension of the method of approximate inverse to distribution spaces. In particular, we define what we mean by a mollifier in the distributional sense. In section 3, we apply this concept to the inverse problem of inverting the spherical Radon transform. Section 4 deals with the design of a mollifier for this problem. The computation of the corresponding reconstruction kernel is outlined in section 5. Section 6 provides a couple of numerical tests using synthetic Radon data, and the proof that our functions satisfy the conditions to be mollifiers is in the appendix.

2. Approximate inverse in distribution spaces. In this section we extend the method of approximate inverse as introduced by Louis and Maass [18] and Louis [16, 17] to distribution spaces.

To this end let $\Omega_1 \subset \mathbb{K}^n$, $\Omega_2 \subset \mathbb{K}^m$ be open sets, $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$, and $V \subset \mathcal{C}^\infty(\Omega_1)$, $W \subset \mathcal{C}^\infty(\Omega_2)$ be subspaces which are closed in their own topology. We denote the dual spaces (continuous linear functionals) for V and W by V', W', respectively. Furthermore we assume $\mathsf{A} : V' \to W'$ to be a linear mapping which is one-to-one. The inverse problem under consideration is as follows. Given a $g \in W'$ lying in the range $\mathsf{A}(V')$ of A, find $f \in V'$ such that

$$\mathsf{A}f = g. \tag{2.1}$$

The concept of approximate inverse involves so called *mollifiers*. The aim is to calculate convolutions of them with the sought solution f rather than to calculate f itself. To extend this concept to distribution spaces V', W' we first define what we mean by a mollifier.

DEFINITION 2.1. *For $\gamma > 0$ let $e_\gamma(\cdot, y) \in V''$ for all $y \in \Omega_1$ such that*

$$\langle \varphi, e_\gamma(\cdot, y)\rangle_{V'\times V''} \in V' \quad \text{for all } \varphi \in V'. \tag{2.2}$$

We call e_γ a mollifier *if and only if*

$$\langle\langle \varphi, e_\gamma(\cdot, y)\rangle_{V'\times V''}, \beta\rangle_{V'\times V} \to \langle \varphi, \beta\rangle_{V'\times V} \tag{2.3}$$

as $\gamma \to 0$ for all $\beta \in V$.

Let $V_1 \subset V'$ and let $V_2 \subset V$. Then, e_γ is a (V_1, V_2)-mollifier if and only if (2.2) *holds for all $\varphi \in V_1$ and* (2.3) *holds for all $\varphi \in V_1$ and $\beta \in V_2$.*

In Definition 2.1 we denote the double dual of V by V'', and $\langle\cdot,\cdot\rangle_{V'\times V}$, $\langle\cdot,\cdot\rangle_{V'\times V''}$ are the corresponding dual pairings.

If e_γ is a mollifier in the sense of Definition 2.3, then for $f \in V'$,

$$f_\gamma(y) := \langle f, e_\gamma(\cdot, y)\rangle_{V'\times V''}, \quad y \in \Omega_1, \tag{2.4}$$

is a distribution in V' which converges to f in the (weak) topology of V'. Because $V \subset V''$, e_γ can be chosen from V. Thus, f_γ is a kind of smooth version of f. If e_γ is a (V_1, V_2)-mollifier, then (2.4) holds for all $f \in V_1$ and convergence holds when tested against all $\beta \in V_2$.

To obtain f_γ from Af we consider the adjoint operator of A. Since $\mathsf{A} : V' \to W'$ is linear, continuous, and one-to-one, it has a linear and continuous adjoint $\mathsf{A}^* : W'' \to V''$ with dense range. Suppose that for each $y \in \Omega_1$ we have an element $\Psi_\gamma(y) \in W''$ satisfying

$$\mathsf{A}^*\Psi_\gamma(y) = e_\gamma(\cdot, y). \tag{2.5}$$

Then, f_γ can be expressed as

$$\begin{aligned} f_\gamma(y) &= \langle f, e_\gamma(\cdot, y)\rangle_{V'\times V''} = \langle f, \mathsf{A}^*\Psi_\gamma(y)\rangle_{V'\times V''} \\ &= \langle \mathsf{A}f, \Psi_\gamma(y)\rangle_{W'\times W''} = \langle g, \Psi_\gamma(y)\rangle_{W'\times W''}, \end{aligned}$$

where $g = \mathsf{A}f$ are the given data. The mapping $S_\gamma : W' \to V'$ defined by

$$S_\gamma g = \langle g, \Psi_\gamma(y)\rangle_{W'\times W''} \tag{2.6}$$

is called the *approximate inverse* of A; the element $\Psi_\gamma(y)$ is the *reconstruction kernel* corresponding to e_γ. Thus, the approximate inverse consists of evaluations of dual pairings of the given data g and the reconstruction kernels $\Psi_\gamma(y)$.

Three main features of the approximate inverse are as follows:

- The reconstruction kernels $\Psi_\gamma(y)$ can be precomputed before the measurement process starts.
- Equation (2.5) is independent of the data g and hence not influenced by noise.
- Invariance properties of A^* help to improve the efficiency of the method, if (2.5) has to be solved only for one single $y \in \Omega_1$. We will demonstrate this in section 3.

REMARK 2.2. *In general, it does not follow that choosing a mollifier e_γ from V results in a reconstruction kernel $\Psi_\gamma(y) \in W$. The key is that* (2.5) *must have a solution in W. If $\mathsf{A}^*(W) \cap V$ is dense in V, then this is more likely. This density condition will happen if the adjoint A^* maps W to $V \subset V''$.*

In practical situations we have only finitely many measurement data available rather than a distribution g. For this reason investigating the semidiscrete operator equation

$$\mathsf{A}_N f = g_N, \tag{2.7}$$

where $\mathsf{A}_N = \Phi_N \mathsf{A}$, $g_N = \Phi_N g \in \mathbb{K}^N$, may fit better to that situation. Here, the observation operator $\Phi_N \in W''$ can be, e.g., point evaluations, if $\mathsf{A}(V')$ consists of continuous, not necessarily integrable, functions. But following the outlines of Rieder and Schuster [27, 28] we formulate the approximate inverse of (2.7) by

$$S_{\gamma,N} g_N(y) = \langle g_N, G_N \, \Phi_N \Psi_\gamma(y) \rangle_{\mathbb{K}^N}, \tag{2.8}$$

where $\Psi_\gamma(y)$ is a reconstruction kernel for (2.1) and $G_N \in \mathbb{K}^{N \times N}$ is a matrix containing the weights of a numerical integration rule which is applied to get the discrete version (2.8) of the dual pairing $\langle \cdot, \cdot \rangle_{W' \times W''}$. Thus, we continue in this article to focus on the continuous problem.

REMARK 2.3. *Compared to the concept of approximate inverse in Hilbert spaces as established by Louis* [16]*, Definition* 2.1 *applies to more general spaces and requires less restrictive assumptions on an element e_γ to be a mollifier. The L^2-theory requires convergence of $f_\gamma(y) = \langle f, e_\gamma(\cdot, y) \rangle \to f(y)$ in L^2 as $\gamma \to 0$, but this distributional setup requires only weak convergence. We should point out that our theory is meant for distribution spaces and does not directly subsume the L^2- or H^s-theory since these Hilbert spaces are not closed subspaces of distribution spaces, the topologies are too different, and their standard duals are not their duals as distribution spaces. It should also be pointed out that this generalization to distributions is necessary for the spherical transform since the transform does not map L^2 into L^2 and the inversion formula we use applies to distributions.*

3. Approximate inverse meets the spherical Radon transform. In this section we apply the method of approximate inverse established in section 2 to the spherical Radon transform. We use the mathematical setup of Andersson's article [2] and formulate some of his main results first.

We start with some notation. Throughout the paper a scalar product $\langle \cdot, \cdot \rangle$ or norm $\| \cdot \|$ without subscript always means the Euclidean scalar product or norm, respectively. We denote the space of all rapidly decreasing, smooth functions by $\mathcal{S}(\mathbb{R}^n)$ and give this space the usual seminorms [29, section 7.3]. This topology turns $\mathcal{S}(\mathbb{R}^n)$ into a Fréchet space. The Fourier transform $\mathsf{F} : \mathcal{S}(\mathbb{R}^n) \to \mathcal{S}(\mathbb{R}^n)$ and its inverse are given by

$$\mathsf{F}f(\xi) = \hat{f}(\xi) = \int_{\mathbb{R}^n} f(x) \, \mathrm{e}^{-\imath \langle \xi, x \rangle} \, dx, \qquad \mathsf{F}^{-1} f(x) = (2\pi)^{-n} \int_{\mathbb{R}^n} f(\xi) \, \mathrm{e}^{\imath \langle x, \xi \rangle} \, d\xi.$$

The dual space $\mathcal{S}'(\mathbb{R}^n)$ of $\mathcal{S}(\mathbb{R}^n)$ is called the set of *tempered distributions*. Each distribution $\varphi \in \mathcal{S}'(\mathbb{R}^n)$ is of finite order [29] and can be written as the derivative of a continuous function of polynomial growth [6].

The Fourier transform gives isomorphisms on $\mathcal{S}(\mathbb{R}^n)$ and on $\mathcal{S}'(\mathbb{R}^n)$. Finally, we often write a vector $x \in \mathbb{R}^{n+1}$ in the form $x = (x', x_{n+1})^\top$, where $x' = (x_1, \dots, x_n)^\top \in \mathbb{R}^n$ contains the first n components of x and x_{n+1} is the last component. We will drop the $\top$ when this correspondence is clear.

The *spherical Radon transform* R assigns a function $f \in \mathcal{S}(\mathbb{R}^{n+1})$ its mean values over all spheres centered about $(z, 0) \in \mathbb{R}^{n+1}$, $z \in \mathbb{R}^n$ with radius $r \geq 0$:

$$\mathsf{R}f(z, r) = \frac{1}{\omega_n} \int_{S^n} f(z + r\,\xi, r\,\eta)\, dS_n(\xi, \eta) = g(z, r). \tag{3.1}$$

Here, ω_n is the area of the n-dimensional sphere $S^n = \{(\xi, \eta) \in \mathbb{R}^{n+1} : \xi \in \mathbb{R}^n, \eta \in \mathbb{R}, \|\xi\|^2 + \eta^2 = 1\}$ and dS_n is the surface measure on S^n.

Obviously $\mathsf{R}f = 0$ holds true for every $f \in \mathcal{S}(\mathbb{R}^{n+1})$ that is odd in the last variable: $f(x', -x_{n+1}) = -f(x', x_{n+1})$. Courant and Hilbert [7] proved that the kernel of R consists exactly of all such functions. This suggests restricting R to the subspace of even functions in the last variable,

$$\mathcal{S}_{\mathrm{e}} := \mathcal{S}_{\mathrm{e}}(\mathbb{R}^{n+1}) = \{f \in \mathcal{S}(\mathbb{R}^{n+1}) : f(x', -x_{n+1}) = f(x', x_{n+1})\}.$$

Unfortunately, even if $f \in \mathcal{S}_{\mathrm{e}}(\mathbb{R}^{n+1})$, the image $\mathsf{R}f$ does not have to be in $L^2(\mathbb{R}^{n+1})$. In fact, if f is the characteristic function of a circle, then $\mathsf{R}f$ has infinite support and does not decrease at infinity. Furthermore, one can show (e.g., using ideas in [19, 24]) that R^{-1} is not continuous in any range of Sobolev norms, at least with data for bounded centers or radii (see Remark 2.3).

Identifying the radius r in (3.1) with the norm $\|w\|$ of a vector $w \in \mathbb{R}^{n+1}$, we introduce the following subspace of $\mathcal{S}(\mathbb{R}^{2n+1})$:

$$\begin{aligned}\mathcal{S}_{\mathrm{r}} &:= \mathcal{S}_{\mathrm{r}}(\mathbb{R}^n \times \mathbb{R}^{n+1}) \\ &= \{f \in \mathcal{S}(\mathbb{R}^{2n+1}) : f(z, w) = \check{f}(z, \|w\|) \text{ for a function } \check{f} \in \mathcal{S}_{\mathrm{e}}(\mathbb{R}^{n+1})\}.\end{aligned}$$

Thus, $\mathcal{S}_{\mathrm{r}}(\mathbb{R}^n \times \mathbb{R}^{n+1})$ consists of the functions in $\mathcal{S}(\mathbb{R}^{2n+1})$ which are radially symmetric in the last $n+1$ variables. We will often view functions in $\mathcal{S}_{\mathrm{r}}(\mathbb{R}^n \times \mathbb{R}^{n+1})$ as functions on $\mathbb{R}^n \times \mathbb{R}$ where we write $f(z, r) = f(z, w)$ with $r = \|w\|$, but when we take the Fourier transform, it will be the Fourier transform on $\mathbb{R}^{2n+1}$.

As mentioned before, we cannot expect that $\mathsf{R}f \in \mathcal{S}_{\mathrm{r}}(\mathbb{R}^n \times \mathbb{R}^{n+1})$ even when $f \in \mathcal{S}_{\mathrm{e}}(\mathbb{R}^{n+1})$. But it is easy to show that $\mathsf{R}f \in \mathcal{S}'_{\mathrm{r}}(\mathbb{R}^n \times \mathbb{R}^{n+1})$, the dual space of $\mathcal{S}_{\mathrm{r}}(\mathbb{R}^n \times \mathbb{R}^{n+1})$. By a density argument we may extend R to domain $\mathcal{S}_{\mathrm{e}}(\mathbb{R}^{n+1})'$. The following theorem summarizes some properties of R considered as mapping between $\mathcal{S}'_{\mathrm{e}}$ and $\mathcal{S}'_{\mathrm{r}}$. The proofs are in [2] or [13].

THEOREM 3.1 (see [2, Theorem 2.1 and Proposition 2.2]). *The spherical Radon transform $\mathsf{R} : \mathcal{S}'_{\mathrm{e}} \to \mathcal{S}'_{\mathrm{r}}$ is a linear, continuous operator which is one-to-one and has range*

$$\mathsf{R}(\mathcal{S}'_{\mathrm{e}}) = \mathcal{S}'_{r,cone} := \left\{ g \in \mathcal{S}'_{\mathrm{r}} : \operatorname{supp} \hat{g} \subset \{(\sigma, \rho) \in \mathbb{R}^n \times [0, \infty) : \rho \geq \|\sigma\|\} \right\} \subset \mathcal{S}'_{\mathrm{r}}. \tag{3.2}$$

If the Fourier transform of $f \in \mathcal{S}'_{\mathrm{e}}$ is equal to an integrable function $\hat{f}(\sigma, \omega)$, then the inversion formula

$$\hat{f}(\sigma, \omega) = c_n\, |\omega|\, (\|\sigma\|^2 + \omega^2)^{(n-1)/2}\, \hat{g}(\sigma, \sqrt{\|\sigma\|^2 + \omega^2}) \tag{3.3}$$

is valid with $c_n = \omega_n / (2\,(2\pi)^n)$ and $g = \mathsf{R}f$.

The adjoint operator $\mathsf{R}^* : \mathcal{S}_{\mathrm{r}} \to \mathcal{S}_{\mathrm{e}}$ *has dense range and is given by*

$$\mathsf{R}^* g(x', x_{n+1}) = \int_{\mathbb{R}^n} g\left(z, \sqrt{\|z - x'\|^2 + x_{n+1}^2}\right) dz; \tag{3.4}$$

its Fourier transform is

$$\mathsf{F}\,\mathsf{R}^* g(\sigma, \rho) = \hat{g}(\sigma, \sqrt{\|\sigma\|^2 + \rho^2}). \tag{3.5}$$

Note that the right-hand side of (3.3) is the Fourier transform of the function g in $\mathbb{R}^{2n+1}$ that is radial in the last $n+1$ variables. The reason to consider R as a map into $\mathcal{S}'_{\mathrm{r}}(\mathbb{R}^n \times \mathbb{R}^{n+1})$ rather than $\mathcal{S}'_{\mathrm{e}}(\mathbb{R}^{n+1})$ is that the relationship between the Fourier transform and spherical transform is easier in these spaces. The constant c_n in (3.3) differs from the corresponding constant in Andersson's article by a factor of $(2\pi)^{-n}$. This inaccuracy was found by Klein [13].

In order to apply the approximate inverse (section 2) to solve the inverse problem of finding a distribution $f \in \mathcal{S}'_{\mathrm{e}}$ satisfying

$$\mathsf{R} f = g \tag{3.6}$$

for a given $g \in \mathcal{S}'_{\mathrm{r}}$ in the range of R, we identify $V = \mathcal{S}_{\mathrm{e}}$, $W = \mathcal{S}_{\mathrm{r}}$, and $\mathsf{A} = \mathsf{R}$. Note that due to Theorem 3.1, R^* maps $\mathcal{S}_{\mathrm{r}}$ into $\mathcal{S}_{\mathrm{e}}$ and we have the situation mentioned in Remark 2.2 and may choose a mollifier $e_\gamma(\cdot, y) \in \mathcal{S}_{\mathrm{e}}$ for every $y \in \mathbb{R}^{n+1}$. Once having a mollifier e_γ at hand, the following extension lemma, whose proof also can be found in [2], helps us to find a solution of the equation

$$\mathsf{R}^* \Psi_\gamma(y) = e_\gamma(\cdot, y), \tag{3.7}$$

which is our reconstruction kernel; see (2.5).

Lemma 3.2 (see [2, Extension Lemma 2.4 and Corollary 2.5]). *There exists a continuous linear mapping* $\mathsf{E} : \mathcal{S}_{\mathrm{e}} \to \mathcal{S}_{\mathrm{r}}$ *such that*

$$\mathsf{R}^*\,\mathsf{E} = id_{\mathcal{S}_{\mathrm{e}}}. \tag{3.8}$$

For $\rho \geq \|\sigma\|$ *the mapping* E *satisfies*

$$\mathsf{F}\,\mathsf{E} f(\sigma, \rho) = \hat{f}(\sigma, \sqrt{\rho^2 - \|\sigma\|^2}). \tag{3.9}$$

If $e_\gamma(\cdot, y) \in \mathcal{S}_{\mathrm{e}}$ is a mollifier in the sense of Definition 2.1, then the reconstruction kernel $\Psi_\gamma(y)$ belonging to e_γ is given by

$$\Psi_\gamma(y) = \mathsf{E} e_\gamma(\cdot, y). \tag{3.10}$$

With the help of (3.8) we easily see that $\Psi_\gamma(y)$ from (3.10) is a solution of (3.7).

From (3.5), it is clear that any continuous E that satisfies (3.9) will satisfy (3.8). We will choose E so that for a mollifier $e_\gamma(\cdot, y)$ in $\mathcal{S}_{\mathrm{e}}$, $\mathsf{E} e_\gamma(\cdot, y)$ is in $\mathcal{S}_{\mathrm{r}}$.

So far we know how to get the reconstruction kernel once we have chosen a mollifier. Theorem 4.1 will provide general criteria that will allow us to construct mollifiers, and with the help of Lemma 3.2 we know how to find a corresponding solution of (3.7). But it would be very time-consuming if we had to solve (3.7) for all reconstruction points y. To this end we prove an invariance property of R^*, Lemma 3.3, which allows us to solve (3.7) only once and to generate *all* reconstruction kernels by applying the invariance to that *one* solution.

For a given $M > 1$, we denote

$$\mathcal{H}^M = \mathcal{H}^M(\mathbb{R}^{n+1}) = \{y = (y', y_{n+1}) \in \mathbb{R}^{n+1} : 1/M < |y_{n+1}|\},$$

$$\mathcal{H}^{M,M} = \mathcal{H}^{M,M}(\mathbb{R}^{n+1}) = \{y = (y', y_{n+1}) \in \mathbb{R}^{n+1} : 1/M < |y_{n+1}| < M\}. \tag{3.11}$$

Furthermore, if $U \subset \mathbb{R}^{n+1}$ is open, we define

$$\begin{aligned}
\mathcal{S}_{\mathrm{e}}(U) &= \{f \in \mathcal{S}_{\mathrm{e}}(\mathbb{R}^{n+1}) : \operatorname{supp} f \subset U\},\\
\mathcal{S}'_{\mathrm{e}}(U) &= \{f \in \mathcal{S}'_{\mathrm{e}}(\mathbb{R}^{n+1}) : \operatorname{supp} f \subset U\},\\
\mathcal{E}'_{\mathrm{e}}(U) &= \{f \in \mathcal{S}'_{\mathrm{e}}(\mathbb{R}^{n+1}) : \operatorname{supp} f \subset U \text{ is compact}\}.
\end{aligned}$$

Note that, in general, $\mathcal{S}'_{\mathrm{e}}(U)$ is a proper subspace of the dual space of $\mathcal{S}_{\mathrm{e}}(U)$.

We define mappings $\mathsf{S}^y_{\mathrm{e}} : \mathcal{S}_{\mathrm{e}} \to \mathcal{S}_{\mathrm{e}}$ and $\mathsf{S}^y_{\mathrm{r}} : \mathcal{S}_{\mathrm{r}} \to \mathcal{S}_{\mathrm{r}}$ by

$$\mathsf{S}^y_{\mathrm{e}} f(x) = \begin{cases} |y_{n+1}|^{-n-1} f\Big(\frac{x'-y'}{|y_{n+1}|}, \frac{x_{n+1}}{|y_{n+1}|}\Big), & y \in \mathcal{H}^M(\mathbb{R}^{n+1}),\\ 0, & y \notin \mathcal{H}^M(\mathbb{R}^{n+1}), \end{cases} \tag{3.12}$$

$$\mathsf{S}^y_{\mathrm{r}} g(z, r) = \begin{cases} |y_{n+1}|^{-2n-1} g\Big(\frac{z-y'}{|y_{n+1}|}, \frac{r}{|y_{n+1}|}\Big), & y \in \mathcal{H}^M(\mathbb{R}^{n+1}),\\ 0, & y \notin \mathcal{H}^M(\mathbb{R}^{n+1}). \end{cases} \tag{3.13}$$

Because $\mathsf{S}^y_{\mathrm{e}}$ and $\mathsf{S}^y_{\mathrm{r}}$ are compositions of dilations and translations, they are linear and continuous mappings on $\mathcal{S}_{\mathrm{e}}$ and $\mathcal{S}_{\mathrm{r}}$, respectively. Moreover, both operators intertwine with the adjoint R^*. It is also clear that $\mathsf{S}^y_{\mathrm{e}} f$ and $\mathsf{S}^y_{\mathrm{r}} g$ can be discontinuous in y for $y_{n+1} = \pm 1/M$.

LEMMA 3.3. *Let $\mathsf{S}^y_{\mathrm{e}} : \mathcal{S}_{\mathrm{e}} \to \mathcal{S}_{\mathrm{e}}$ and $\mathsf{S}^y_{\mathrm{r}} : \mathcal{S}_{\mathrm{r}} \to \mathcal{S}_{\mathrm{r}}$ be defined as in* (3.12) *and* (3.13)*, respectively. Then,*

$$\mathsf{S}^y_{\mathrm{e}}\, \mathsf{R}^* = \mathsf{R}^*\, \mathsf{S}^y_{\mathrm{r}}. \tag{3.14}$$

Proof. Let $y \in \mathcal{H}^M(\mathbb{R}^{n+1})$. Using representation (3.4) together with the definitions (3.12) and (3.13) gives

$$\begin{aligned}
\mathsf{R}^*\mathsf{S}^y_{\mathrm{r}} g(x', x_{n+1}) &= |y_{n+1}|^{-2n-1} \int_{\mathbb{R}^n} g\Big(\frac{z-y'}{|y_{n+1}|}, |y_{n+1}|^{-1} \sqrt{\|z-x'\|^2 + x_{n+1}^2}\Big)\, dz\\
&= |y_{n+1}|^{-n-1} \int_{\mathbb{R}^n} g\Big(z, \sqrt{\|z - |y_{n+1}|^{-1}(x'-y')\|^2 + |y_{n+1}|^{-2} x_{n+1}^2}\Big)\, dz\\
&= \mathsf{S}^y_{\mathrm{e}}\, \mathsf{R}^* g(x', x_{n+1})
\end{aligned}$$

for all $g \in \mathcal{S}_{\mathrm{r}}$. For $y \notin \mathcal{H}^M(\mathbb{R}^{n+1})$ assertion (3.14) follows immediately, since both sides are equal to zero. ∎

Lemma 3.3 tells us that under certain conditions we may restrict ourselves to solving (3.7) only for *one single* $y \in \mathbb{R}^{n+1}$.

COROLLARY 3.4. *For each $\gamma > 0$ let $\bar{e}_\gamma \in \mathcal{S}_{\mathrm{e}}(\mathbb{R}^{n+1})$ and $e_\gamma(\cdot, y) \in \mathcal{S}_{\mathrm{e}}$ be defined by $\mathsf{S}^y_{\mathrm{e}}$:*

$$e_\gamma(x, y) = \mathsf{S}^y_{\mathrm{e}} \bar{e}_\gamma(x). \tag{3.15}$$

Assume e_γ is a mollifier. Then, we get all corresponding reconstruction kernels by solving

$$\mathsf{R}^* \bar{\Psi}_\gamma = \bar{e}_\gamma \tag{3.16}$$

and setting

$$\Psi_\gamma(y) = \Psi_\gamma(y; z, r) = \mathsf{S}_{\mathrm{r}}^y \bar{\Psi}_\gamma(z, r). \tag{3.17}$$

If e_γ is an $(\mathcal{E}'_{\mathrm{e}}(\mathcal{H}^{M,M}), \mathcal{S}_{\mathrm{e}}(\mathcal{H}^{M,M}))$-mollifier, then

$$S_\gamma \mathsf{R} f := \langle Rf, \Psi_\gamma \rangle_{\mathcal{S}'_{\mathrm{r}} \times \mathcal{S}_{\mathrm{r}}} \to f$$

for $f \in \mathcal{E}'_{\mathrm{e}}(\mathcal{H}^{M,M})$. This means that

$$\langle \langle Rf, \Psi_\gamma \rangle_{\mathcal{S}'_{\mathrm{r}} \times \mathcal{S}_{\mathrm{r}}}, \beta \rangle_{\mathcal{E}'_{\mathrm{e}}(\mathcal{H}^{M,M}) \times \mathcal{S}_{\mathrm{e}}(\mathcal{H}^{M,M})} \to \langle f, \beta \rangle_{\mathcal{E}'_{\mathrm{e}}(\mathcal{H}^{M,M}) \times \mathcal{S}_{\mathrm{e}}(\mathcal{H}^{M,M})}$$

for all $\beta \in \mathcal{S}_{\mathrm{e}}(\mathcal{H}^{M,M})$.

We will construct a general class of $\bar{e}_\gamma$ in section 4 and show that the resulting e_γ satisfy the definition. We now prove the corollary.

Proof. Taking into account (3.17) and (3.14), statement (3.16) is a consequence of

$$e_\gamma(x, y) = \mathsf{S}_{\mathrm{e}}^y \bar{e}_\gamma(x) = \mathsf{S}_{\mathrm{e}}^y \mathsf{R}^* \bar{\Psi}_\gamma(x) = \mathsf{R}^* \mathsf{S}_{\mathrm{r}}^y \bar{\Psi}_\gamma(x) = \mathsf{R}^* \{\Psi_\gamma(y)\}(x). \qquad \square$$

Considering (3.8) a solution of (3.16) is given by $\bar{\Psi}_\gamma = \mathsf{E}\bar{e}_\gamma$.

REMARK 3.5. *Putting*

$$f_\gamma(y) = \langle f, \mathsf{S}_{\mathrm{e}}^y \bar{e}_\gamma \rangle_{\mathcal{S}'_{\mathrm{e}} \times \mathcal{S}_{\mathrm{e}}}$$

it becomes clear from (3.12) *that* $\operatorname{supp} f_\gamma \subset \mathcal{H}^M(\mathbb{R}^{n+1})$. *Thus, using the invariance $\mathsf{S}_{\mathrm{e}}^y$ to generate mollifiers, we can only recover objects $f \in \mathcal{S}'_{\mathrm{e}}$ with support in $\mathcal{H}^M(\mathbb{R}^{n+1})$. But this is not a restriction in applications, e.g., in sonar or radar, since the support of any object to be reconstructed is always a positive distance from the line $y_{n+1} = 0$. For technical reasons, our mollifiers satisfy the convergence assumption* (2.3) *for bounded $|y_{n+1}|$, so we will reconstruct f_γ only on $\mathcal{H}^M$ or $\mathcal{H}^{M,M}$. This is not a serious practical restriction since M can be chosen arbitrarily large. Therefore, we will construct $(\mathcal{E}'_{\mathrm{e}}(\mathcal{H}^{M,M}), \mathcal{S}_{\mathrm{e}}(\mathcal{H}^{M,M}))$-mollifiers.*

To use the method of approximate inverse for inverting R, we

- choose a mollifier e_γ fulfilling the conditions of Theorem 4.1 defined by $\mathsf{S}_{\mathrm{e}}^y$: $e_\gamma(x, y) = \mathsf{S}_{\mathrm{e}}^y \bar{e}_\gamma(x)$ and calculate $\bar{\Psi}_\gamma = \mathsf{E}\bar{e}_\gamma$;
- compute the approximate inverse of R as

$$S_\gamma g(y) = \langle g, \mathsf{S}_{\mathrm{r}}^y \bar{\Psi}_\gamma \rangle_{\mathcal{S}'_{\mathrm{r}} \times \mathcal{S}_{\mathrm{r}}}, \tag{3.18}$$

  where $g = \mathsf{R}f$ are the given data.

Considering (3.9), we have only an explicit representation for $\mathsf{F}\,\mathsf{E}\bar{e}_\gamma$ when $\rho \geq \|\sigma\|$. We want to obtain $\bar{\Psi}_\gamma$ rather than its Fourier transform because a discrete Fourier transform would extend the data, which are given in applications only on a bounded domain, periodically and could cause large artifacts. Furthermore even in the two-dimensional case ($n = 1$) we would have to compute a three-dimensional Fourier transform of the data. Therefore, we need an explicit representation of $\mathsf{F}\,\mathsf{E}\bar{e}_\gamma$ for *all* $\rho \geq 0$ and $\sigma \in \mathbb{R}^n$. (Andersson uses an extension method from Stein [30] which is fairly arbitrary and not explicit for calculations.) We will present an idea in section 4 that will circumvent these difficulties.

4. Design of a mollifier for R. Due to Corollary 3.4 we let the mollifier e_γ be defined $e_\gamma(x,y) = \mathsf{S}_{\mathrm{e}}^y \bar{e}_\gamma(x)$ as in (3.12).

Since we will need the Fourier transform of $\bar{e}_\gamma$ to compute the reconstruction kernel (see (3.9)) it is appropriate to choose $\bar{e}_\gamma$ as a tensor product

$$\bar{e}_\gamma(x) = e_\gamma^1(x') \otimes e_\gamma^2(x_{n+1}), \tag{4.1}$$

where $e_\gamma^1 \in \mathcal{S}(\mathbb{R}^n)$, $e_\gamma^2 \in \mathcal{S}(\mathbb{R})$, e_γ^2 even. Defining $e_\gamma(x,y)$ as in (3.15), (4.1) it is obvious that $e_\gamma(\cdot, y) \in \mathcal{S}_{\mathrm{e}}(\mathbb{R}^{n+1})$ for all $y \in \mathbb{R}^{n+1}$.

In view of (3.9) and Theorem 4.1 below we want e_γ and $\bar{e}_\gamma$ to have the following properties:

1. $\int_{\mathbb{R}^n} e_\gamma^1(z)\,dz = 1 = \int_{\mathbb{R}} e_\gamma^2(t)\,dt$.
2. $\mathsf{F} e_\gamma^1$ is easy to calculate.
3. $\mathsf{F} e_\gamma^2(\sqrt{\xi})$ has a nice extension for $\xi < 0$.

By "nice" in 3, we mean that the extension is explicitly known since we do not want to apply an extension lemma [30] like Andersson did it in his article [2]. Moreover we need an explicit expression for that extension to calculate the corresponding reconstruction kernel.

Now we get more explicit with our choices for e_γ^1 and e_γ^2. We define

$$e_\gamma^1(x') = \gamma^{-n}\, e^1(x'/\gamma) \quad \text{for } e^1(x') \in \mathcal{S}(\mathbb{R}^n), \quad \int_{\mathbb{R}^n} e^1(z)\,dz = 1. \tag{4.2}$$

We have to be careful with respect to the choice of e_γ^2. Let $F \in \mathcal{S}_e(\mathbb{R})$ have mean value 1. To guarantee the mollifier property, because of the dilation by y_{n+1} in $\mathsf{S}_{\mathrm{e}}^y$ (see (3.12) and (3.15)), we define

$$e_\gamma^2(q) = \frac{1}{2\,\gamma}\left\{F\left(\frac{q+1}{\gamma}\right) + F\left(\frac{q-1}{\gamma}\right)\right\} \quad \text{for } F \in \mathcal{S}_{\mathrm{e}}(\mathbb{R}), \quad \int_{\mathbb{R}} F(t)\,dt = 1. \tag{4.3}$$

We will show that property 3 is fulfilled when we define F as in (4.5) below.

The following key theorem asserts that these properties guarantee e_γ is a mollifier. The proof will be given in the appendix.

THEOREM 4.1. *Let $M > 1$ and let functions e_γ^1 and e_γ^2 be given by* (4.2) *and* (4.3). *Then, e_γ defined by* (3.15) *and* (4.1) *is an $(\mathcal{E}_{\mathrm{e}}'(\mathcal{H}^{M,M}), \mathcal{S}_{\mathrm{e}}(\mathcal{H}^{M,M}))$-mollifier.*

We will now construct specific functions e_γ^1 and e_γ^2 that we will use in our algorithm. We define

$$e_\gamma^1(x') = \gamma^{-n} e^1(x'/\gamma), \quad e^1(x') = (2\,\pi)^{-n/2}\,\exp(-\|x'\|^2/2), \quad x' \in \mathbb{R}^n, \tag{4.4}$$

which obviously is a function in $\mathcal{S}(\mathbb{R}^n)$ with mean value 1, since $\int_{\mathbb{R}^n} e_\gamma^1(x')\,dx' = \hat{e}_\gamma^1(0) = 1$.

We have to be more careful in the choice of e_γ^2. The desirable extension property 3 for e_γ^2 is fulfilled if there exists a function $g \in \mathcal{S}(\mathbb{R})$ satisfying

$$\mathsf{F} e_\gamma^2(\sqrt{\xi}) = g(\xi^2). \tag{4.5}$$

The function

$$F(q) := 2\,\mathsf{F}^{-1}\{\exp(-|\xi|^4)\}(2\,q) \tag{4.6}$$

satisfies (4.5) with $g(\xi) = \exp(-|\xi|^2)$. So, F is an even function in $\mathcal{S}(\mathbb{R})$ with mean value equal to 1. We define e_γ^2 using (4.3) and the specific function (4.6).

REMARK 4.2. *Since the inverse Fourier transform of* $\exp(-|\xi|^4)$ *does not decrease as rapidly as* $\exp(-|\xi|^2)$ *near* $\xi = 0$*, we introduced the dilation factor* 2 *in* (4.6) *to make the decay behavior the same in both variables (see also Figure* 1*).*

COROLLARY 4.3. *Let* $M > 1$*. The function* $e_\gamma = e^1_\gamma \otimes e^2_\gamma$ *defined using* (4.4) *and* (4.3) *with* F *defined by* (4.6) *satisfies the assumptions of Theorem* 4.1 *and therefore is an* $(\mathcal{E}'_{\mathrm{e}}(\mathcal{H}^{M,M}), \mathcal{S}_{\mathrm{e}}(\mathcal{H}^{M,M}))$*-mollifier.*

Proof. All we need to do is observe that our specific e^1 and F satisfy $\int_{\mathbb{R}^n} e^1(z)\,dz = 1 = \int_{\mathbb{R}} F(t)\,dt$ and that e_γ is constructed according to Theorem 4.1. ∎

Figure 1 displays $\bar{e}_\gamma$ in the case of $n = 1$, $\gamma = 0.06$. It has its peak in $(0, 1)$.

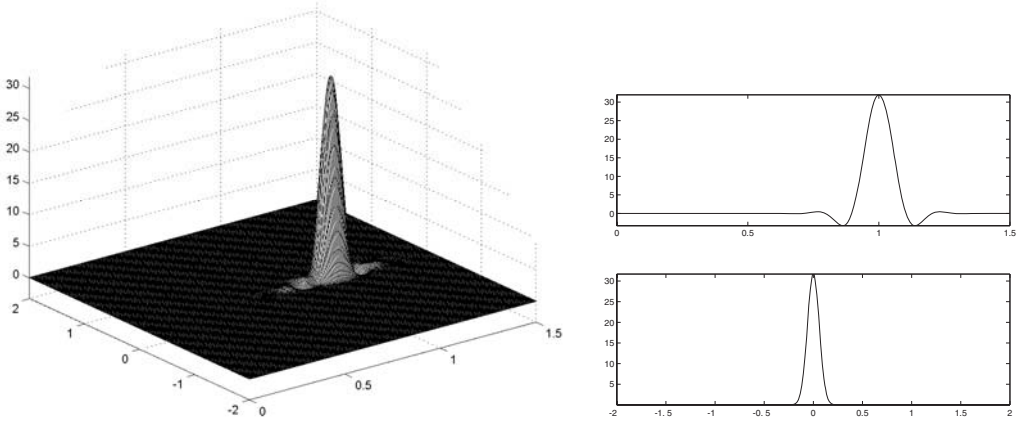


FIG. 1. *Plot of* $e_\gamma(x_1, x_2)$ *in the two-dimensional case* ($n = 1$) *for* $\gamma = 0.06$ *(left picture). On the right-hand side is the graph of* e^1_γ *(bottom) and* e^2_γ *(top). The width of the peak is about* 0.5 *units in each case (note the different scales), which is achieved by the dilation in* (4.6)*.*

5. Computation of the reconstruction kernel $\bar{\Psi}_\gamma$. Throughout this section we assume $\bar{e}_\gamma$ to be given as in (4.1), (4.2), (4.3), (4.4), and (4.6) and $e_\gamma(x, y) = \mathsf{S}^y_{\mathrm{e}}\bar{e}_\gamma(x)$. Our aim is to compute $\bar{\Psi}_\gamma = \mathsf{E}\bar{e}_\gamma$.

From Lemma 3.2 we know that

$$\mathsf{F}\bar{\Psi}_\gamma(\sigma, \rho) = \mathsf{F}\,\mathsf{E}\bar{e}_\gamma = \mathsf{F}\bar{e}_\gamma(\sigma, \sqrt{\rho^2 - \|\sigma\|^2}) \quad \text{if } \rho \geq \|\sigma\|, \tag{5.1}$$

where $\rho \geq 0$, $\sigma \in \mathbb{R}^n$. Thus, we have to compute the Fourier transform of $\bar{e}_\gamma$ at first.

LEMMA 5.1. *We have that*

$$\mathsf{F}\bar{e}_\gamma(\sigma, \rho) = \hat{e}^1_\gamma(\sigma)\,\hat{e}^2_\gamma(\rho) = \cos(\rho)\,\mathrm{e}^{-\gamma^2\|\sigma\|^2/2}\,\mathrm{e}^{-\gamma^4\rho^4/16}, \tag{5.2}$$

where $\sigma \in \mathbb{R}^n$*,* $\rho \in \mathbb{R}$*.*

Proof. The proof follows from a straightforward calculation using the definition of $\bar{e}_\gamma$. ∎

So far by Lemma 5.1 we have the representation

$$\mathsf{F}\bar{\Psi}_\gamma(\sigma, \rho) = \cos(\sqrt{\rho^2 - \|\sigma\|^2})\,\mathrm{e}^{-\gamma^2\|\sigma\|^2/2}\,\mathrm{e}^{-\gamma^4(\rho^2 - \|\sigma\|^2)^2/16} \quad \text{if } \rho \geq \|\sigma\|. \tag{5.3}$$

To get $\bar{\Psi}_\gamma$ for *all* $\rho \geq 0$ and $\sigma \in \mathbb{R}^n$, we have to find an extension of $\cos\sqrt{\xi}$ for $\xi < 0$ that turns (5.3) into a function in $\mathcal{S}_{\mathrm{r}}$. The natural extension involves $\cosh\sqrt{-\xi}$ for $\xi < 0$. As noted in section 3, we can extend $\bar{\Psi}_\gamma$ arbitrarily, and for computational

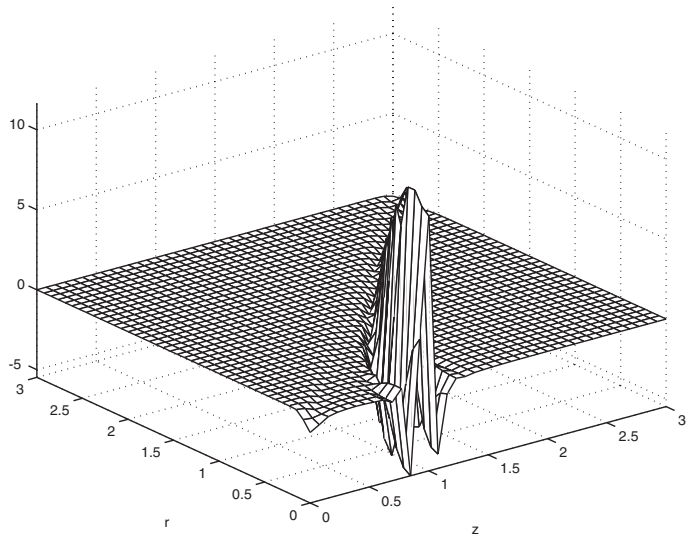


FIG. 2. *The reconstruction kernel* $\bar{\Psi}_\gamma$ *given as in* (5.6) *for* $\gamma = 0.06$ *and* $n = 1$. *The integrals have been computed using numerical integration.*

reasons, we will cut this function off away from $\xi = 0$. Let $\chi \in \mathcal{C}^\infty(\mathbb{R})$ be zero on $(-\infty, -1]$ and 1 on $[0, \infty)$ and let

$$G(\xi) = \begin{cases} \cos\sqrt{\xi}, & \xi \geq 0, \\ \chi(\xi)\,\cosh(\sqrt{|\xi|}), & \xi < 0. \end{cases} \tag{5.4}$$

The Fourier transform $\mathsf{F}\bar{\Psi}_\gamma$ is given by

$$\mathsf{F}\bar{\Psi}_\gamma(\sigma, \rho) = G(\rho^2 - \|\sigma\|^2)\,\mathrm{e}^{-\gamma^2\,\|\sigma\|^2/2}\,\mathrm{e}^{-\gamma^4\,(\rho^2 - \|\sigma\|^2)^2/16} \tag{5.5}$$

and we get $\bar{\Psi}_\gamma$ by applying the inverse Fourier transform.

LEMMA 5.2. *Let* $\bar{e}_\gamma$ *be given as in* (4.1), (4.2), (4.3), (4.4), *and* (4.6). *Then, a solution of* $\mathsf{R}^*\bar{\Psi}_\gamma = \bar{e}_\gamma$ *is represented by*

$$\begin{aligned} \bar{\Psi}_\gamma(z, r) = 2^n\,(2\,\pi)^{-\frac{3}{2}\,n - \frac{1}{2}} \int_{\mathbb{R}^n_+} \int_0^\infty \Big\{ & G(\rho^2 - \|\sigma\|^2)\,\mathrm{e}^{-\gamma^2\,(\frac{\|\sigma\|^2}{2} + \frac{\gamma^2}{16}\,(\rho^2 - \|\sigma\|^2)^2)} \\ & \cdot \rho^{(n+1)/2}\,\mathcal{J}_{(n-1)/2}(\rho\,r)\,\cos(\langle \sigma, z\rangle) \Big\}\,d\rho\,d\sigma. \end{aligned} \tag{5.6}$$

Here, $\mathbb{R}^n_+ = \{x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n : x_j \geq 0\}$, $\mathcal{J}_\nu$ *is the Bessel function of first kind of order* ν, *and* G *is given as in* (5.4).

Proof. The proof follows by a simple application of an inverse Fourier transform of dimension $2n + 1$ to (5.5) in which one uses Lemma 5.1, spherical coordinates, and the identity

$$\int_{S^n} \mathrm{e}^{\imath\,\rho\,r\,\langle \omega, \theta\rangle}\,d\omega = (2\,\pi)^{(n+1)/2}\,(\rho\,r)^{(1-n)/2}\,\mathcal{J}_{(n-1)/2}(\rho\,r),$$

which can be found, e.g., in [10]. ∎

Figure 2 displays a picture of $\bar{\Psi}_\gamma$ for $\gamma = 0.06$ and $n = 1$ corresponding to the two-dimensional case. The integrals in (5.6) have been computed using numerical integration, where the integrals were cut off when the absolute value of the integrand was less then 10^{-12}. The reconstruction kernel in Figure 2 belongs to the mollifier shown in Figure 1 and has its absolute maximum point in $(0, 1)$, just as the mollifier $\bar{e}_\gamma$.

6. Implementation and numerical results. We now have all the ingredients to implement the approximate inverse for the spherical Radon transform. We present results for the two-dimensional case ($n = 1$). The reconstruction kernel $\bar{\Psi}_\gamma$ (5.6) belonging to the mollifier (4.1), (4.2), (4.3) has the representation

(6.1)
$$\bar{\Psi}_\gamma(z,r) = \frac{2}{(2\,\pi)^2}\left\{\int_0^\infty\!\!\int_0^\infty \tau\,\mathcal{J}_0(\sqrt{\tau^2+\sigma^2}\,r)\,\cos\tau\,\mathrm{e}^{-\gamma^2\,(\frac{\sigma^2}{2}+\frac{\gamma^2\,\tau^4}{16})}\,\cos(\sigma\,z)\,d\tau\,d\sigma \right.$$
$$\left. + \int_0^\infty\!\!\int_0^\sigma \tau\,\mathcal{J}_0(\sqrt{\sigma^2-\tau^2}\,r)\,\chi(-\tau^2)\,\cosh\tau\,\mathrm{e}^{-\gamma^2\,(\frac{\sigma^2}{2}+\frac{\gamma^2\,\tau^4}{16})}\,\cos(\sigma\,z)\,d\tau\,d\sigma\right\},$$

where we used the substitutions $\rho = \sqrt{\tau^2+\sigma^2}$ and $\rho = \sqrt{\sigma^2-\tau^2}$, respectively.

Throughout this section we suppose that f has compact support in $\mathcal{H}^{M,M}(\mathbb{R}^2)$ for a certain $M > 1$. The method of approximate inverse used to solve the problem $\mathsf{R}f = g$ for $n = 1$ has the form $S_\gamma \mathsf{R}f(y) = \langle \mathsf{R}f, \mathsf{S}_{\mathrm{r}}^y \bar{\Psi}_\gamma\rangle_{\mathcal{S}'_{\mathrm{r}}\times\mathcal{S}_{\mathrm{r}}}$.

We now adjust the algorithm to practical situations where only finitely many data on a bounded domain are available. Assume that equally spaced centers $z_k \in [\lambda, \Lambda]$, $\lambda < \Lambda$, $k = 0, \ldots, P$, and equally spaced radii $r_m \in [0, R]$, $R > 0$, $m = 0, \ldots, Q$, are given, so we have $N = (P+1)\,(Q+1)$ spherical averages of f at hand. More explicitly, instead of $\mathsf{R}f$ itself we have only the vector $\phi_N\,\mathsf{R}f \in \mathbb{R}^N$ as data, where $\phi_N : \mathcal{C}(\mathbb{R}\times[0,\infty)) \to \mathbb{R}^N$ are the point evaluations

$$(\phi_N v)_{k,m} = v(z_k, r_m), \quad 0 \le k \le P, \quad 0 \le m \le Q.$$

Remark 6.1. *The observation operator ϕ_N, which contains all information about the measurement geometry, is well defined only if the function to be evaluated is continuous. Since $\mathsf{R}f \in \mathcal{S}'_{\mathrm{r}}$ we have to postulate that $\mathsf{R}f$ is a continuous, but not necessarily integrable, function in order to apply ϕ_N properly. Thus, we assume $\mathsf{R}f \in \mathcal{C}(\mathbb{R}\times[0,\infty))$ which is not a large restriction since R smooths of order $n/2$ in Sobolev scales.*

To recover f from $\phi_N \mathsf{R}f$ we apply the trapezoidal sum corresponding to the nodes $\{z_k\}, \{r_m\}$ and obtain

$$\begin{aligned} S_{\gamma,N}\,\phi_N\,\mathsf{R}f(y) &= \langle \phi_N\,\mathsf{R}f, Q_N\,\phi_N\,\mathsf{S}_{\mathrm{r}}^y\bar{\Psi}_\gamma\rangle_{\mathbb{R}^N} \\ &= \frac{2\,\pi}{|y_2|^3}\,h_z\,h_r \sum_{k=0}^{P}\sum_{m=0}^{Q} r_m\,\bar{\Psi}_\gamma\!\left(\frac{z_k - y_1}{|y_2|}, \frac{r_m}{|y_2|}\right)\mathsf{R}f(z_k, r_m) \end{aligned} \tag{6.2}$$

for $y \in \mathcal{H}^M(\mathbb{R}^2)$, $Q_N = h_z\,h_r\,I_{N,N}$ (compare (2.8)).

Formula (6.2) was applied to get the reconstructions in Figures 3 and 4.

As mentioned in section 5 we compute $\bar{\Psi}_\gamma$ by applying numerical integration to (6.1) choosing convenient integration boundaries. Moreover we determine $\bar{\Psi}_\gamma(z,r)$ on the square $[0, 15]^2$ on an equidistant mesh grid consisting of 128×128 grid points. Since the kernel is rapidly decreasing, the absolute value of $\bar{\Psi}_\gamma$ outside the square

$[0,15]^2$ is rather small, so we can extend the kernel by 0 there. Using the symmetry $\bar{\Psi}_\gamma(z,r) = \bar{\Psi}_\gamma(-z,r)$ and linear interpolation we get $\bar{\Psi}_\gamma(z,r)$ for every $z \in \mathbb{R}$, $r \geq 0$.

To check the performance of the above algorithm we implemented it to reconstruct several objects. All reconstructions were computed for $(y_1, y_2) \in [0,7] \times [1,8]$ using an equidistant mesh grid with 64×64 grid points. The objects are assumed to have their support in $\mathcal{H}^1(\mathbb{R}^2)$. The data are given on equally spaced points with $\lambda = -36$, $\Lambda = 36$, $P = 384$, $R = 50$, and $Q = 256$. *Note that in all pictures the y_2-axis is the horizontal one, whereas the y_1-axis (the sonar sources, circle centers) is the vertical one.*

First, we recovered the characteristic function of a circle centered at $(4,4)$ with radius 1 and density 2. Figure 3 shows the original circle as well as the approximate inverse $S_{\gamma,N}\phi_N \mathsf{R}f$. We used the reconstruction kernel (6.1) with $\gamma = 0.06$ which was precomputed for $(z,r) \in [0,15]^2$ using 128×128 equally distributed grid points.

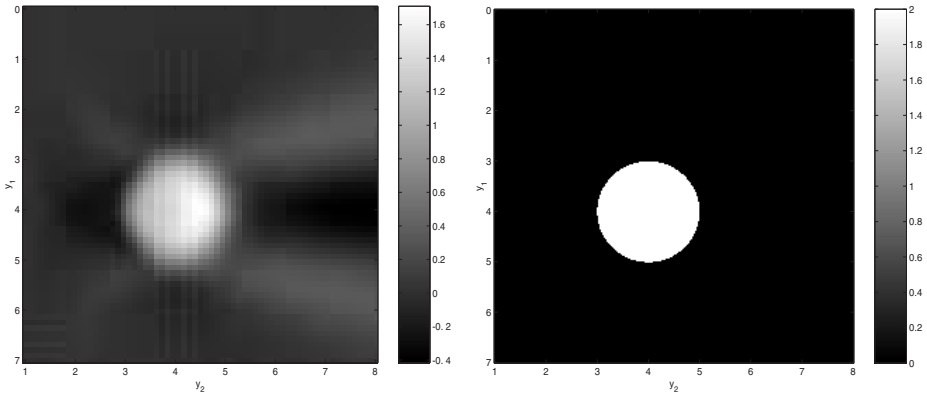


FIG. 3. *Reconstruction of the characteristic function of a circle (left) and original object function (right),* $\gamma = 0.06$.

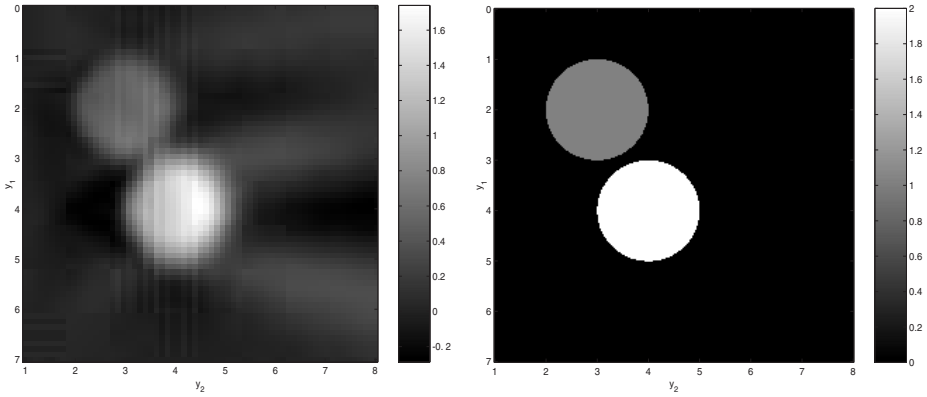


FIG. 4. *Reconstruction of two circles f_1 and f_2 (left) and original object function (right),* $\gamma = 0.06$.

Second, we applied the algorithm to the sum of the function in Figure 3 and the characteristic function of a disk centered at $(2,3)$ and of radius 1. The reconstruction as well as the original object can be seen in Figure 4; the parameters are the same as

in Figure 3.

These tests show that the method of approximate inverse works fine, and the reconstructions are comparable to those in [9]. Some blurring in the reconstructions is probably caused by the numerical calculation of the reconstruction kernel and truncation error. However, some ill-posedness is inherent in the problem.

REMARK 6.2. *Some of the fuzzy reconstruction boundaries in Figures* 3 *and* 4 *are intrinsic to the problem. As shown in* [19, 24]*, the object boundaries that are most difficult to reconstruct are those not tangent to circles in the data set. This means that horizontal boundaries in Figures* 3 *and* 4 *will be intrinsically hardest to reconstruct since the set of circle centers is the vertical axis. Since more-or-less vertical boundaries are tangent to spheres in the data set, the microlocal analysis predicts they will be easiest to reconstruct. This is analogous to limited angle X-ray tomography in which some boundaries are "invisible" in the data* [25].

7. Conclusions. In this paper we extended the method of approximate inverse, a regularization scheme for operators between Hilbert spaces, to distribution spaces. We applied the method to the inversion problem of the spherical Radon transform which appears in sonar as well as in radar. This algorithm allows one to solve inverse problems for linear operators which are not bounded mappings between Hilbert or Banach spaces.

We presented a representation for a reconstruction kernel $\bar{\Psi}_\gamma$ in arbitrary dimensions (5.6). Unfortunately, in the three-dimensional case ($n = 2$) numerical integration to get $\bar{\Psi}_\gamma$ is too time consuming and we are working on other ways to get the reconstruction kernel. In this case a modified inversion formula presented by Klein [13] might be useful. This inversion formula could also be helpful to obtain an *analytic* expression for the reconstruction kernel $\bar{\Psi}_\gamma$, which would also increase the accuracy of the reconstructed solution. This and stability and error analysis (as for Hilbert space in [27]) will be part of future research.

Appendix A. Proof of Theorem 4.1. Let $M > 1$. We recall the general construction of e_γ given in section 4. Let $e_\gamma(x, y) = \mathsf{S}_{\mathrm{e}}^y \bar{e}_\gamma(x)$, where

$$\bar{e}_\gamma(x) = \mathrm{e}_\gamma^1(x') \otimes \mathrm{e}_\gamma^2(x_{n+1}), \tag{A.1}$$

$$e_\gamma^1(x') = \gamma^{-n} e^1(x'/\gamma), \quad \int_{\mathbb{R}^n} e^1(x')\,dx' = 1, \quad e^1 \in \mathcal{S}(\mathbb{R}^n), \tag{A.2}$$

$$e_\gamma^2(q) = \frac{1}{2\,\gamma}\left\{F\left(\frac{q+1}{\gamma}\right) + F\left(\frac{q-1}{\gamma}\right)\right\} \quad \text{for } F \in \mathcal{S}_{\mathrm{e}}(\mathbb{R}), \quad \int_{\mathbb{R}} F(t)\,dt = 1. \tag{A.3}$$

We will use several steps to show that e_γ is an $(\mathcal{E}'_{\mathrm{e}}(\mathcal{H}^{M,M}), \mathcal{S}_{\mathrm{e}}(\mathcal{H}^{M,M}))$-mollifier. First, we will prove (2.2) using Lemma A.1. Then, we will prove a distributional Fubini's theorem, Lemma A.2, and finally, we will prove the convergence result (2.3) which concludes the proof of Theorem 4.1.

LEMMA A.1. *Let $\gamma > 0$ be fixed, e_γ be defined by* (A.1)–(A.3)*, and $\varphi \in \mathcal{S}'_{\mathrm{e}}(\mathbb{R}^{n+1})$. Then, the function $\langle \varphi, e_\gamma(\cdot, y)\rangle_{\mathcal{S}'_{\mathrm{e}} \times \mathcal{S}_{\mathrm{e}}}$ is a continuous function of polynomial growth for $y \in \mathcal{H}^M$ and is 0 for $y \notin \mathcal{H}^M$. Therefore $\langle \varphi, e_\gamma(\cdot, y)\rangle_{\mathcal{S}'_{\mathrm{e}}(\mathbb{R}^{n+1}) \times \mathcal{S}_{\mathrm{e}}(\mathbb{R}^{n+1})} \in \mathcal{S}'_{\mathrm{e}}(\mathbb{R}^{n+1})$.*

Proof. First, using the definition of e_γ, one proves the map $y \mapsto e_\gamma(\cdot, y)$ is a continuous map from $\mathcal{H}^M$ to $\mathcal{S}_{\mathrm{e}}(\mathbb{R}^{n+1})$. Therefore, $\langle \varphi, e_\gamma(\cdot, y)\rangle_{\mathcal{S}'_{\mathrm{e}} \times \mathcal{S}_{\mathrm{e}}}$ is continuous for $y \in \mathcal{H}^M$ and is 0 if not.

We simplify the problem by reducing the calculation to integrals of functions. By [6] there exists a multi-index $\alpha \in \mathbb{N}_0^{n+1}$ and a continuous function P_φ of polynomial

growth such that

$$\varphi = D^\alpha P_\varphi, \tag{A.4}$$

where $\alpha = (\alpha', \alpha_{n+1})$ and $D^\alpha = \partial_{x_1}^{\alpha_1} \cdots \partial_{x_{n+1}}^{\alpha_{n+1}}$.

For $y \in \mathcal{H}^M$, we obtain

(A.5)

$$\begin{aligned}\varphi_\gamma(y) &:= \langle \varphi, e_\gamma(\cdot, y)\rangle_{\mathcal{S}_e' \times \mathcal{S}_e} = (-1)^{|\alpha|} \int_{\mathbb{R}^{n+1}} P_\varphi(x)\, D_x^\alpha e_\gamma(x, y)\, dx \\ &= \frac{1}{2(-\gamma\,|y_{n+1}|)^{|\alpha|}} \int_{\mathbb{R}^n} \int_{\mathbb{R}} \Big[P_\varphi(\gamma\,|y_{n+1}|\, z' + y', \gamma\,|y_{n+1}|\, z_{n+1} + |y_{n+1}|) \\ &\quad + P_\varphi(\gamma\,|y_{n+1}|\, z' + y', \gamma\,|y_{n+1}|\, z_{n+1} - |y_{n+1}|) \Big] D^{\alpha'} e^1(z')\, D^{\alpha_{n+1}} F(z_{n+1})\, dz_{n+1}\, dz',\end{aligned}$$

where we used the substitutions $z' = (x' - y')/(\gamma\,|y_{n+1}|)$ and $z_{n+1} = (x_{n+1}/|y_{n+1}| \pm 1)/\gamma$, as well as the symmetry of F. Since P_φ is polynomially increasing, there exists a constant $C_\varphi > 0$ and a $\kappa > 0$ such that

$$|P_\varphi(x)| \le C_\varphi\,(1 + \|x\|^2)^\kappa \quad \text{as } \|x\| \to \infty, \quad x \in \mathbb{R}^{n+1}. \tag{A.6}$$

Using (A.6) and some simple estimates, we show

$$\Big| P_\varphi\big(\gamma\,|y_{n+1}|\, z' + y', \gamma\,|y_{n+1}|\, z_{n+1} \pm |y_{n+1}|\big) \Big| \le C_\varphi\, 2^\kappa\, (1 + \gamma^2\,|y_{n+1}|^2\,\|z\|^2)^\kappa\, (1 + \|y\|^2)^\kappa.$$

This allows us to estimate (A.5) as

$$|\varphi_\gamma(y)| \le C_\varphi\, 2^\kappa\, q_\gamma\, (\gamma\,|y_{n+1}|)^{-|\alpha|}\, (1 + \|y\|^2)^\kappa, \qquad y \in \mathcal{H}^M,$$

with $q_\gamma := \int_{\mathbb{R}^n} \int_{\mathbb{R}} (1 + \gamma^2\,|y_{n+1}|^2\,\|z\|^2)^\kappa\, D^{\alpha'} e^1(z')\, D^{\alpha_{n+1}} F(z_{n+1})\, dz_{n+1}\, dz' < \infty$, which finishes the proof. ∎

Our next task is to prove a distributional Fubini's theorem that will allow us to examine the pairing $\langle e_\gamma(x, \cdot), \beta\rangle$ to show the convergence result (2.3) in Definition 2.1.

LEMMA A.2 (distributional Fubini's theorem). *Let $\gamma > 0$ be fixed and e_γ be defined by* (A.1)–(A.3). *Further assume that $\varphi \in \mathcal{S}_e'(\mathbb{R}^{n+1})$ and $\beta \in \mathcal{S}_e(\mathcal{H}^M)$. Then,*

$$\begin{aligned}&\big\langle \langle \varphi, e_\gamma(\cdot, y)\rangle_{\mathcal{S}_e'(\mathbb{R}^{n+1}) \times \mathcal{S}_e(\mathbb{R}^{n+1})}, \beta \big\rangle_{\mathcal{S}_e'(\mathbb{R}^{n+1}) \times \mathcal{S}_e(\mathcal{H}^M)} \\ &\qquad = \big\langle \varphi, \langle e_\gamma(x, \cdot), \beta\rangle_{\mathcal{S}_e'(\mathbb{R}^{n+1}) \times \mathcal{S}_e(\mathcal{H}^M)} \big\rangle_{\mathcal{S}_e'(\mathbb{R}^{n+1}) \times \mathcal{S}_e(\mathbb{R}^{n+1})}.\end{aligned} \tag{A.7}$$

Furthermore,

$$\beta_\gamma(x) := \langle e_\gamma(x, \cdot), \beta\rangle_{\mathcal{S}_e'(\mathbb{R}^{n+1}) \times \mathcal{S}_e(\mathcal{H}^M)} \in \mathcal{S}_e(\mathbb{R}^{n+1})\,. \tag{A.8}$$

Note that here, β_γ is a function of x, and in section 2, f_γ is a function of y.

Proof. We reduce this to a Fubini theorem for functions. Since $\varphi = D^\alpha P_\varphi$ for a function P_φ with polynomial growth by (A.4), we can again use (A.5) to write

(A.9)

$$\langle \varphi_\gamma, \beta\rangle_{\mathcal{S}_e'(\mathbb{R}^{n+1}) \times \mathcal{S}_e(\mathcal{H}^M)} = \int_{\mathcal{H}^M} \int_{\mathbb{R}^n} \int_{\mathbb{R}} I_\varphi^\gamma(y', y_{n+1}, x', x_{n+1})\, dx_{n+1}\, dx'\, dy_{n+1}\, dy',$$

where

$$I^{\gamma}_{\varphi}(y', y_{n+1}, x', x_{n+1}) := \frac{(-1)^{|\alpha|}}{2} \left(\gamma\,|y_{n+1}|\right)^{-n-1-|\alpha|} \beta(y', y_{n+1})\, P_{\varphi}(x', x_{n+1})$$
$$\cdot\, (D^{\alpha'} e^1)\left(\frac{x'-y'}{\gamma\,|y_{n+1}|}\right) \left\{ (D^{\alpha_{n+1}} F)\left(\frac{x_{n+1}-|y_{n+1}|}{\gamma\,|y_{n+1}|}\right) + (D^{\alpha_{n+1}} F)\left(\frac{x_{n+1}+|y_{n+1}|}{\gamma\,|y_{n+1}|}\right)\right\}.$$

Using (A.6), $y \in \mathcal{H}^M$, the fact that F, β, and e^1 are in $\mathcal{S}_{\mathrm{e}}$, and some basic inequalities (e.g., $(1+\|a-b\|^2)^{-q} \le 2^q\,(1+\|b\|^2)^q\,(1+\|a\|^2)^{-q}$, $a, b \in \mathbb{R}^n$, $q \in \mathbb{N}$), we may estimate

$$|I^{\gamma}_{\varphi}(y', y_{n+1}, x', x_{n+1})| \le (C_{\varphi}/2)\ (1+\|x\|^2)^{\kappa}\,(\gamma/M)^{-n-1-|\alpha|}\,|\beta(y', y_{n+1})|$$
$$\cdot \left(1+\frac{\|x'-y'\|^2}{\gamma^2\, y_{n+1}^2}\right)^{-q_1} \left\{ \left(1+\frac{(x_{n+1}-|y_{n+1}|)^2}{\gamma^2\, y_{n+1}^2}\right)^{-q_2} + \left(1+\frac{(x_{n+1}+|y_{n+1}|)^2}{\gamma^2\, y_{n+1}^2}\right)^{-q_2}\right\}$$
$$\le \frac{C_{\varphi}}{2}\,\frac{(1+\|x\|^2)^{\kappa}}{(1+\|y\|^2)^{q_3}} \left(\frac{\gamma}{M}\right)^{-n-1-|\alpha|} \frac{(1+\|y'\|^2)^{q_1}}{(1+\|x'\|^2)^{q_1}}\,\frac{(1+|y_{n+1}|^2)^{q_2}}{(1+|x_{n+1}|^2)^{q_2}}\,[2(1+\gamma^2\, y_{n+1}^2)]^{q_1+q_2}$$

for arbitrary $q_1, q_2, q_3 \in \mathbb{N}$.

We see for sufficiently large q_1, q_2, q_3 that the integrand in (A.9) is bounded by an integrable function in $(x, y) \in \mathbb{R}^{n+1} \times \mathcal{H}^M$.

This allows us to switch the order of integration in (A.9). Since the integral in this switched version is smooth with uniformly integrable derivatives in $y \in \mathcal{H}^M$ for x in any compact set, we can pull the D^{α} out of the inner integral. Finally, we use the definition of derivative on $\mathcal{S}'_{\mathrm{e}}$ to prove (A.7).

To show (A.8), we let $\alpha \in \mathbb{N}_0^{n+1}$ be an arbitrary multi-index. We will prove that $D^{\alpha}\beta_{\gamma}$ decreases rapidly. We bring the D^{α} inside the integral for β_{γ} and use estimates as above, and we find a constant $\tilde{c}_{\gamma} > 0$ such that

$$|D^{\alpha}\beta_{\gamma}(y)| \le \tilde{c}_{\gamma}\,(1+\|y'\|^2)^{-q_1}\,(1+y_{n+1}^2)^{-q_2}, \qquad (y', y_{n+1}) \in \mathbb{R}^{n+1},$$

for arbitrary numbers $q_1, q_2 \in \mathbb{N}$ since $\beta \in \mathcal{S}_{\mathrm{e}}(\mathcal{H}^M)$ and γ is fixed. Now, using similar arguments as for the bound on $|I^{\gamma}_{\varphi}|$, we prove assertion (A.8). ∎

The final key is the following important convergence result.

LEMMA A.3. *Let e_{γ} be defined by* (A.1)–(A.3). *Let $\beta \in \mathcal{S}_{\mathrm{e}}(\mathcal{H}^{M,M})$ and $\alpha \in \mathbb{N}_0^{n+1}$ be a multi-index. Assume that β_{γ} is defined by* (A.8). *Then, $D^{\alpha}\beta_{\gamma} \to D^{\alpha}\beta(x)$ pointwise in $\mathcal{H}^{M,M}$, and $D^{\alpha}\beta_{\gamma}$ is uniformly bounded in $(x, \gamma) \in \mathcal{H}^{M,M} \times (0, 1)$.*

Proof. We first use the symmetry of F to write

(A.10)
$$\beta_{\gamma}(x) = \int_{\mathcal{H}^{M,M}} \frac{1}{(\gamma\,|y_{n+1}|)^{n+1}}\, e^1\left(\frac{x'-y'}{\gamma\,|y_{n+1}|}\right) F\left(\left(\frac{x_{n+1}}{y_{n+1}}-1\right)\Big/\gamma\right)\beta(y)\,dy_{n+1}\,dy'.$$

We assume $(x, y) \in \mathcal{H}^{M,M} \times \mathcal{H}^{M,M}$ and then we use the change of variables

$$z' = (x'-y')/(|y_{n+1}|\gamma), \qquad z_{n+1} = \left(\frac{x_{n+1}}{y_{n+1}}-1\right)\Big/\gamma, \tag{A.11}$$

and we have the following simple but important estimate:

$$\frac{1}{M^2} < \frac{1}{M|x_{n+1}|} < \frac{1}{|\gamma z_{n+1}+1|} < \frac{M}{|x_{n+1}|} < M^2. \tag{A.12}$$

Then, the integral in (A.10) becomes

$$\beta_\gamma(x) = \int_{\mathbb{R}^n}\int_{1/|\gamma z_{n+1}+1|<M^2} e^1(z')F(z_{n+1}) \cdot \beta\left(x' - \frac{\gamma|x_{n+1}|}{|\gamma z_{n+1}+1|}z', \frac{x_{n+1}}{\gamma z_{n+1}+1}\right)\frac{1}{|\gamma z_{n+1}+1|}dz'dz_{n+1}, \tag{A.13}$$

where the limits of integration in (A.13) are determined because $1/M < |y_{n+1}| < M$ and $\operatorname{supp}\beta \subset \mathbb{R}^n \times [1/M, M]$.

In order to subtract $\beta(x)$ within the integral (A.13), we define an auxiliary function that simplifies the calculation,

$$b_\gamma(x) = \beta(x)\int_{1/|\gamma z_{n+1}+1|<M^2} e^1(z')\,F(z_{n+1})\,\frac{1}{|\gamma z_{n+1}+1|}\,dz_{n+1}.$$

We must calculate $D^\alpha[\beta_\gamma - b_\gamma]$ and show this difference goes to zero as $\gamma \to 0$. To do this, we take the derivative inside the integral:

$$D^\alpha[\beta_\gamma(x) - b_\gamma(x)] = \int_{\mathbb{R}^n}\int_{1/|\gamma z_{n+1}+1|<M^2} e^1(z')\,F(z_{n+1}) \cdot D_x^\alpha\left\{\beta\left(x' - \frac{\gamma|x_{n+1}|}{|\gamma z_{n+1}+1|}z', \frac{x_{n+1}}{\gamma z_{n+1}+1}\right) - \beta(x)\right\}\frac{1}{|\gamma z_{n+1}+1|}\,dz'\,dz_{n+1}. \tag{A.14}$$

To show that (A.14) converges to zero, we must do two things:

1. We need to show for each $x \in \mathcal{H}^{M,M}$ that the integrand in (A.14) is bounded by an integrable function uniformly in $\gamma \in (0,1)$.
2. We need to show $D^\alpha\beta_\gamma$ is bounded by an integrable function, uniformly in $\gamma \in (0,1)$.

To show 1, we need to examine the derived integrand. The $D_{x'}^{\alpha'}$ terms are evaluated on β in both terms of (A.14) and they do not cause a problem, so we will evaluate them first. This gives an expression

$$D_x^\alpha\left\{\beta\left(x' - \frac{\gamma\,|x_{n+1}|}{|\gamma\,z_{n+1}+1|}\,z', \frac{x_{n+1}}{\gamma\,z_{n+1}+1}\right) - \beta(x)\right\} = D_{x_{n+1}}^{\alpha_{n+1}}(D_{x'}^{\alpha'}\beta)\left(x' - \frac{\gamma\,|x_{n+1}|}{|\gamma\,z_{n+1}+1|}\,z', \frac{x_{n+1}}{\gamma z_{n+1}+1}\right) - D_x^\alpha\beta(x). \tag{A.15}$$

However, because x_{n+1} appears in both coordinates of the first β in (A.15), some of the derivatives in $D_{x_{n+1}}^{\alpha_{n+1}}$ fall on the first coordinate. We will let $\delta' = (\delta_1, \ldots, \delta_n)$ denote a multi-index in $\mathbb{N}_0^n$. An explicit calculation shows that the integrand in (A.14) can be written for $x_{n+1} > 1/M > 0$ as a sum of terms in which some derivatives in x_{n+1} fall on the first coordinates of β and then the term in which all derivatives fall on the last coordinate, the integrand in (A.14) becomes

(A.16)

$$\frac{e^1(z')\,F(z_{n+1})}{|\gamma\,z_{n+1}+1|}\left[\sum_{0<|\delta'|\le\alpha_{n+1}}\left(\frac{\gamma^{|\delta'|}\,(-z)^{\delta'}}{|\gamma\,z_{n+1}+1|^{|\delta'|}\,(\gamma\,z_{n+1}+1)^{\alpha_{n+1}-|\delta'|}} \cdot \left(\partial_{x_{n+1}}^{\alpha_{n+1}-|\delta'|}\,D_{x'}^{\delta'+\alpha'}\beta\right)\left(x' - \frac{\gamma\,|x_{n+1}|}{|\gamma\,z_{n+1}+1|}z', \frac{x_{n+1}}{\gamma\,z_{n+1}+1}\right)\right) + \left\{(\gamma\,z_{n+1}+1)^{-\alpha_{n+1}}\,(D^\alpha\beta)\left(x' - \frac{\gamma\,|x_{n+1}|}{|\gamma\,z_{n+1}+1|}z', \frac{x_{n+1}}{\gamma\,z_{n+1}+1}\right) - D^\alpha\beta(x)\right\}\right].$$

A similar formula is obtained for $x_{n+1} < -1/M < 0$.

Because $1/|\gamma z_{n+1} + 1| \leq M^2$, it can be seen from (A.16) that the integrand of (A.14) can be bounded by an integrable function uniformly in $\gamma \in (0,1)$. Hence, an application of Lebesgue's dominated convergence theorem shows $D^\alpha(\beta_\gamma - b_\gamma) \to 0$ pointwise for $x \in \mathcal{H}^{M,M}$. This is valid for two reasons: the sum in (A.16) is a factor of γ times a bounded function, and the difference in braces goes to zero as $\gamma \to 0$. Since $(b_\gamma - \beta) \to 0$ in $\mathcal{S}_{\mathrm{e}}(\mathcal{H}^{M,M})$ we thus have $D^\alpha(\beta_\gamma - \beta) \to 0$ pointwise in $\mathcal{H}^{M,M}$.

A similar boundedness argument shows that $D^\alpha \beta_\gamma$ is bounded by an integrable function uniformly in $\gamma \in (0,1)$. ∎

At last, we finish the proof of Theorem 4.1. Recall that in the statement of this theorem, φ has compact support in $\mathcal{H}^{M,M}$ and $\varphi = D^\alpha P_\varphi$ for a function P_φ of polynomial growth (A.4). Thus, there are compactly supported functions $\psi_1(x')$ and $\psi_2(x_{n+1})$ such that ψ_2 is one on $[-M, -1/M] \cup [1/M, M]$ and supported in $[-2M, -1/2M] \cup [1/2M, 2M]$ and $\psi(x) = \psi_1(x')\psi_2(x_{n+1})$ is one on a neighborhood of $\operatorname{supp}\varphi$. Then, $\varphi = \psi\, D^\alpha P_\varphi$.

By Lemma A.2,

$$\begin{aligned} \langle \varphi_\gamma, \beta \rangle_{\mathcal{S}'_{\mathrm{e}}(\mathbb{R}^{n+1}) \times \mathcal{S}_{\mathrm{e}}(\mathcal{H}^{M,M})} &= \langle \varphi, \beta_\gamma \rangle_{\mathcal{S}'_{\mathrm{e}}(\mathcal{H}^{M,M}) \times \mathcal{S}_{\mathrm{e}}(\mathbb{R}^{n+1})} \\ &= (-1)^{|\alpha|} \int_{\mathcal{H}^{M,M}} P_\varphi(x)\, D^\alpha \big\{ \psi(x)\, \beta_\gamma(x) \big\}\, dx. \end{aligned} \tag{A.17}$$

By the product rule for derivatives and the convergence result Lemma A.3, we see that the derivative in (A.17) converges pointwise on any compact set in x, and it is uniformly bounded. Therefore, we can use Lebesgue's dominated convergence theorem again to finish the proof of Theorem 4.1.

Acknowledgments. The authors are indebted to Jens Klein, Aleksei Beltukov, and Alfred Louis for very useful conversations. The authors are grateful to the referees and editors, in particular Adel Faridani, for thoughtful comments that improved the article.

REFERENCES

[1] M. Agranovsky and E. T. Quinto, *Injectivity sets for Radon transform over circles and complete systems of radial functions*, J. Funct. Anal., 139 (1996), pp. 383–414.
[2] L.-E. Andersson, *On the determination of a function from spherical averages*, SIAM J. Math. Anal., 19 (1988), pp. 214–232.
[3] R. Barakat, *Private communication*, 1997.
[4] A. Beltukov, *Sonar Transforms*, Ph. D. thesis, Tufts University, Medford, MA, 2004.
[5] M. Cheney, *Tomography problems arising in Synthetic Aperture Radar*, in Radon Transforms and Tomography, Contemp. Math. 278, AMS, Providence, RI, 2001, pp. 15–28.
[6] F. Constantinescu, *Distributions and Their Applications in Physics*, Teubner, Leipzig, 1974.
[7] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. II, Wiley-Interscience, New York, 1962.
[8] A. Denisjuk, *Integral geometry on the family of semi-spheres*, Fract. Calc. Appl. Anal., 2 (1999), pp. 31–46.
[9] A. Denisjuk, *On Numerical Reconstruction of a Function by Its Arc Means with Incomplete Data*, Tech. report, Brest State University, Belarus, 1999.
[10] J. Fawcett, *Inversion of n-dimensional spherical averages*, SIAM J. Appl. Math., 45 (1985), pp. 336–341.
[11] D. Finch, S. K. Patch, and Rakesh, *Determining a function from its mean values over a family of spheres*, SIAM J. Math. Anal., 35 (2004), pp. 1213–1240.
[12] F. Jensen, W. Kuperman, M. Porter, and H. Schmidt, *Computational Ocean Acoustics*, AIP Press, Springer, New York, 2000.

[13] J. Klein, *Inverting the spherical Radon transform for physically meaningful functions*, preprint, Westfälische Wilhelms-Universität, Institut für Numerische und Instrumentelle Mathematik, Münster, Germany, 2003. Available online from http://www.arachne.uni-muenster.de:8000/num/Preprints/2003/kleinje/.
[14] R. Krueger, D. Reinecke, and G. Krueger, *Thermoacoustic computed tomography*, Med. Phys., 26 (1999), pp. 1832–1837.
[15] M. Lavrentiev, V. Romanov, and V. Vasiliev, *Multidimensional Inverse Problems for Differential Equations*, Lecture Notes in Math. 167, Springer-Verlag, New York, 1970.
[16] A. Louis, *Approximate inverse for linear and some nonlinear problems*, Inverse Problems, 12 (1996), pp. 175–190.
[17] A. Louis, *A unified approach to regularization methods for linear ill-posed problems*, Inverse Problems, 15 (1999), pp. 489–498.
[18] A. Louis and P. Maass, *A mollifier method for linear operator equations of the first kind*, Inverse Problems, 6 (1990), pp. 427–440.
[19] A. Louis and E. T. Quinto, *Local tomographic methods in SONAR*, in Surveys on Solution Methods for Inverse Problems, D. Colton, H. Engl, A. Louis, J. McLaughlin, and W. Rundell, eds., Springer, Vienna, 2000, pp. 147–154.
[20] M. Nessibi, L. Rachdi, and K. Trimeche, *Ranges and inversion formulas for spherical mean operator and its dual*, J. Math. Anal. Appl., 196 (1995), pp. 861–884.
[21] S. Norton, *Reconstruction of a reflectivity field from line integrals over circular paths*, J. Acoust. Soc. Amer., 67 (1980), pp. 853–863.
[22] S. Norton, *Reconstruction of a two-dimensional reflecting medium over a circular domain: Exact solution*, J. Acoust. Soc. Amer., 67 (1980), pp. 1266–1273.
[23] S. Norton and M. Linzer, *Ultrasonic reflectivity imaging in three dimensions: Exact inverse scattering solutions for plane, cylindrical, and spherical apertures*, IEEE Trans. Biomed. Engrg., 28 (1981), pp. 200–202.
[24] V. Palamodov, *Reconstruction from limited data of arc means*, J. Fourier Anal. Appl., 6 (2000), pp. 26–42.
[25] E. T. Quinto, *Singularities of the X-ray transform and limited data tomography in* $\mathbb{R}^2$ *and* $\mathbb{R}^3$, SIAM J. Math. Anal., 24 (1993), pp. 1215–1225.
[26] A. Ramm, *Injectivity of the spherical mean operator*, C. R. Math. Acad. Sci. Paris, 335 (2002), pp. 1033–1038.
[27] A. Rieder and T. Schuster, *The approximate inverse in action* II, Math. Comp., 72 (2003), pp. 1399–1415.
[28] A. Rieder and T. Schuster, *The approximate inverse in action* III: 3*D-Doppler tomography*, Numer. Math., 97 (2004), pp. 353–378.
[29] W. Rudin, *Functional Analysis*, McGraw-Hill, New York, 1973.
[30] E. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

SIAM J. APPL. MATH.
Vol. 65, No. 4, pp. 1388–1406

MODELING OF SEISMIC DATA IN THE DOWNWARD CONTINUATION APPROACH*

CHRISTIAAN C. STOLK† AND MAARTEN V. DE HOOP‡

Abstract. Seismic data are commonly modeled by a high-frequency single scattering approximation. This amounts to a linearization in the medium coefficient about a smooth background. The discontinuities are contained in the medium perturbation. The high-frequency part of the wavefield in the background medium is described by a geometrical optics representation. It can also be described by a one-way wave equation. Based on this we derive a downward continuation operator for seismic data. This operator solves a pseudodifferential evolution equation in depth, the so-called double-square-root equation. We consider the modeling operator based on this equation. If the rays in the background that are associated with the reflections due to the perturbation are nowhere horizontal, the singular part of the data is described by the solution to an inhomogeneous double-square-root equation.

Key words. seismic modeling, microlocal analysis, double-square-root equation

AMS subject classifications. 86A15, 35R30

DOI. 10.1137/S0036139904439545

1. Introduction.

In reflection seismology one places point sources and point receivers on the earth's surface. The source generates acoustic waves in the subsurface, which are reflected where the medium properties vary discontinuously. In seismic imaging, one tries to reconstruct the properties of the subsurface from the reflected waves that are observed. There are various approaches to seismic imaging, each based on a different mathematical model for seismic reflection data with underlying assumptions. In general, seismic scattering and inverse scattering have been formulated in the form of a linearized inverse problem for the medium coefficient in the acoustic wave equation. The linearization is around a smoothly varying background, called the velocity model, which is a priori also unknown.

In this paper and a companion paper [24] we study a method of seismic imaging introduced by Clayton [6] and Claerbout [5]. The key concept in this method is the construction of data of fictitious experiments carried out in the subsurface, at increasing depths, from data observed at the earth's surface. These so-called downward continued data are then used for imaging the medium contrast as well as for a reflection tomographic procedure to estimate the smoothly varying background (known as migration velocity analysis). The downward continuation approach to seismic imaging has received much attention in the geophysical research literature, and it is currently widely used in practice in various approximations [3, 19, 16].

The downward continuation of data is derived from the factorization of the wave equation into two one-way wave equations. This factorization is closely connected to the notion of wave splitting [28]. One-way wave equations, in various approximations,

*Received by the editors January 7, 2004; accepted for publication (in revised form) October 8, 2004; published electronically April 26, 2005. This work was supported in part by the Mathematical Sciences Research Institute through National Science Foundation grant DMS-9810361.
http://www.siam.org/journals/siap/65-4/43954.html
†Department of Applied Mathematics, University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands (c.c.stolk@ewi.utwente.nl).
‡Center for Wave Phenomena, Colorado School of Mines, Golden, CO 80401 (mdehoop@Mines.EDU).

have been extensively used in applications other than seismics: for integrated optics (see, e.g., [12]) and for underwater acoustics (see, e.g., [25, 7]).

There are basically two categories of seismic imaging methods. One category is associated with the evolution of waves and data in time; the other is associated with the evolution in depth (or another principal spatial direction). The first category contains approaches known under the collective names of Kirchhoff migration [4] or generalized Radon transform inversion, and reverse-time migration [21]; the second category comprises the downward continuation approach. There are great computational advantages of the downward continuation approach to seismic imaging over the Kirchhoff approaches. There are fundamental, theoretical advantages as well, in particular with a view to the problem of estimating the smoothly varying background. These are analyzed in a separate paper [24]. For the Kirchhoff approach to seismic imaging there is a solid mathematical theory, which treats seismic imaging as an inverse problem and shows that singularities can be reconstructed [2, 20]. For the downward continuation approach much research has gone into the development of numerical one-way wave equations, but little is known from an analysis point of view. For a constant coefficient background, the downward continuation method was cast into an inverse problem in [1]. For the case of variable coefficients, which of course is the case of interest in practice, there has been no such theory.

The purpose of this paper is to develop a mathematical theory for modeling seismic reflection data in the downward continuation approach. As was done in the analysis of Kirchhoff methods, we make use of techniques and concepts from microlocal analysis, such as wave front set, denoted by $\mathrm{WF}(\cdot)$, and Fourier integral operators; see, e.g., [10] for background information on these concepts. We introduce the main concepts and operators involved in the method. We then study the double-square-root modeling operator. This modeling operator and its properties will be the point of departure for the development of an inverse scattering theory [24].

In our notation we will distinguish the vertical coordinate $z \in \mathbb{R}$ from the horizontal coordinates $x \in \mathbb{R}^{n-1}$ and write $(z, x) \in \mathbb{R}^n$. In these coordinates the scalar acoustic wave equation with wave speed function $c_0(z, x)$ is given by

$$Pu = f, \qquad P = c_0(z,x)^{-2}\partial_t^2 - \partial_z^2 - \sum_{j=1}^{n-1} \partial_{x_j}^2, \tag{1.1}$$

where $u = u(z, x, t)$ is the acoustic pressure. The equation is considered for t in a time interval $]0, T[$, together with an initial condition $u(\cdot, \cdot, 0) = 0$. The solution to (1.1) can be written as

$$u(z,x,t) = \int_0^t \int G(z,x,t-t_0,z_0,x_0) f(z_0,x_0,t_0)\, \mathrm{d}z_0 \mathrm{d}x_0 \mathrm{d}t_0, \tag{1.2}$$

where G is the Green's function of (1.1). The source f can be a distribution.

To model the scattering of waves, we adopt the linearized scattering or Born approximation. The linearization is in the wavespeed, around a smooth (C^∞) background c_0; for the full wavespeed function we write $c = c_0 + \delta c$. The perturbation δc may contain singularities. The perturbation in G at the acquisition surface $z = 0$ is given by (see, e.g., [2])

$$\begin{aligned} \delta G(0,r,t,0,s) = \int_{\mathbb{R}_+ \times \mathbb{R}^{n-1}} \int_0^t G(0,r,t-t_0,z_0,x_0)\, 2c_0^{-3}(z_0,x_0)\delta c(z_0,x_0) \\ \times \partial_{t_0}^2 G(z_0,x_0,t_0,0,s)\, \mathrm{d}t_0 \mathrm{d}z_0 \mathrm{d}x_0, \end{aligned} \tag{1.3}$$

where both $s, r \in \mathbb{R}^{n-1}$. We assume that the acquisition manifold Y, which contains the set of values of (s, r, t) used in the acquisition, is a bounded open subset of $\mathbb{R}^{2n-2} \times \mathbb{R}_+$. The modeled data are then a function of $(s, r, t) \in Y$ given by (1.3). We define the Born modeling map F through (1.3) as the map from δc to δG evaluated at $z = 0$. Since Y is bounded and the waves propagate with finite speed we may assume that δc is supported in a bounded open subset X of $\mathbb{R}_+ \times \mathbb{R}^{n-1}$. Furthermore, we assume that $\overline{X} \cap \{z = 0\} = \emptyset$. Naturally, (1.3) is, in general, not a complete model for raw data measured in seismic experiments. It models data that are the input for imaging and inversion and have undergone some processing.

We summarize some results in the literature about the modeling map, F. The solution operator (1.2) is such that singularities in the solution propagate along bicharacteristics. Denote by $p(z, x, \zeta, \xi, \tau) = -c(z, x)^{-2}\tau^2 + \zeta^2 + \|\xi\|^2$ the principal symbol of P. Propagating singularities are in the characteristic set, given by the points $(z, x, t, \zeta, \xi, \tau) \in T^*\mathbb{R}^{n+1}$ with

$$p(z, x, \zeta, \xi, \tau) = -c(z, x)^{-2}\tau^2 + \zeta^2 + \|\xi\|^2 = 0. \tag{1.4}$$

The bicharacteristics are the solution curves of a Hamilton system with Hamiltonian given by p,

$$\frac{\mathrm{d}(z, x, t)}{\mathrm{d}\lambda} = \frac{\partial p}{\partial(\zeta, \xi, \tau)}, \quad \frac{\mathrm{d}(\zeta, \xi, \tau)}{\mathrm{d}\lambda} = -\frac{\partial p}{\partial(z, x, t)}. \tag{1.5}$$

Assuming that $\tau \neq 0$, the time t is strictly increasing or decreasing with λ and can be used as parameter for the solution curve. To parameterize points on the solution curves, we use the initial position (z_0, x_0), the take-off direction $\alpha \in S^{n-1}$, the frequency τ, which together define the initial cotangent vector $(\zeta_0, \xi_0) = -\tau c(z_0, x_0)^{-1}\alpha$, and the time t (instead of λ). Points on the solution curves will be denoted by

$$\begin{aligned}\eta(t, z_0, x_0, \alpha, \tau) = (&\eta_z(t, z_0, x_0, \alpha, \tau), \eta_x(t, z_0, x_0, \alpha, \tau), t,\\ &\eta_\zeta(t, z_0, x_0, \alpha, \tau), \eta_\xi(t, z_0, x_0, \alpha, \tau), \tau).\end{aligned} \tag{1.6}$$

The variable τ is invariant along the Hamilton flow. We take $t = 0$ as the initial value for t (note that (1.5) are time translation invariant).

To ensure that δG defines a continuous map from $\mathcal{E}'(X)$ to $\mathcal{D}'(\mathbb{R}^n \times \mathbb{R}^n \times]0, T[)$ and that the restriction of δG to Y is a Fourier integral operator we make the following assumption on c_0.

ASSUMPTION 1. *There are no rays from $(0, s)$ to $(0, r)$ with travel time t such that $(s, r, t) \in Y$. For all ray pairs connecting $(0, r)$ via some $(z, x) \in X$ to $(0, s)$ with total time t such that $(s, r, t) \in Y$, the rays intersect the plane $z = 0$ transversally at r and s.*

We also assume that rays from such a point $(z, x) \in X$ intersect the surface $z = 0$ only once, because all reflections must come from the region $z > 0$ (the subsurface). The first part of the assumption excludes direct rays, or a pair of incident and reflected rays with scattering angle π. The second part of the assumption excludes rays grazing the plane $z = 0$. Concerning the second part, strictly only caustics grazing the plane $z = 0$ have to be excluded. In practice the wave speed near the surface is much lower than in the interior of the earth, and waves from the interior arrive under small angles with the vertical. So from a geophysical point of view one is only interested in incoming rays that intersect the measurement surface transversally. We have the following theorem. (See [10] for a general reference on Fourier integral operators.)

THEOREM 1.1 (see [20, 17]). *With Assumption* 1 *the map* F *is a Fourier integral operator* $\mathcal{E}'(X) \to \mathcal{D}'(Y)$ *of order* $(n-1)/4$ *with canonical relation*

(1.7)
$$\big\{(\eta_x(t_{\rm s},z,x,\beta,\tau),\eta_x(t_{\rm r},z,x,\alpha,\tau),t_{\rm s}+t_{\rm r},\eta_\xi(t_{\rm s},z,x,\beta,\tau),\eta_\xi(t_{\rm r},z,x,\alpha,\tau),\tau;z,x,\zeta,\xi)|$$
$$t_{\rm s},t_{\rm r}>0,\ \eta_z(t_{\rm s},z,x,\beta,\tau)=\eta_z(t_{\rm r},z,x,\alpha,\tau)=0,\ (\zeta,\xi)=-\tau c_0(z,x)^{-1}(\alpha+\beta),$$
$$(z,x,\alpha,\beta,\tau)\in \textit{ subset of } X\times(S^{n-1})^2\times\mathbb{R}\backslash 0\big\}\subset T^*\mathbb{R}^{2n-1}_{(s,r,t)}\times T^*\mathbb{R}^n_{(z,x)}.$$

In this paper, we express F in terms of a depth-continuation operator, and we study the properties of this operator. The main contributions of this paper are the following:

(i) We define an *upward continuation operator* $H(z,z_0)$ using the solution operators to one-way wave equations. Its adjoint will be the downward continuation operator. Intuitively this operator maps data from a fictitious experiment carried out at depth z_0 to data from an experiment carried out at depth z, $z<z_0$. Subject to Assumption 2 in the main text—stating, essentially, that the rays in the background that are associated with the reflections are nowhere tangent to horizontal—we prove that the data $F\delta c$ are given by $\int_0^\infty(\dots)H(0,z)(\dots)g(z,\cdot,\cdot,\cdot)\,{\rm d}z$, where the dots are pseudodifferential factors specified below and $g=g(z,s,r,t)$ is given by mapping $c_0^{-3}\delta c$ to a function $E_2E_1(c_0^{-3}\delta c)$ of (z,s,r,t) using the maps

(1.8)
$$E_1:\mathcal{D}'(\mathbb{R}^n)\to\mathcal{D}'(\mathbb{R}^{2n-1})\ :\ (c_0^{-3}\delta c)(z,x)\mapsto h(z,\bar x,x)=\delta(x-\bar x)(c_0^{-3}\delta c)(z,\tfrac{\bar x+x}{2}),$$
$$E_2:\mathcal{D}'(\mathbb{R}^{2n-1})\to\mathcal{D}'(\mathbb{R}^{2n})\ :\ h(z,\bar x,x)\mapsto\delta(t)h(z,\bar x,x)$$

(Theorem 5.1).

(ii) We show that the operator $H(z,z_0)$ solves the initial value problem for a first-order pseudodifferential evolution equation in depth, known as the *double-square-root (DSR) equation*. The data can be identified with the solution to an inhomogeneous DSR equation, with inhomogeneous term g (section 3). The computation of the map from g to data and the computation of its adjoint can be done by marching in depth using the DSR equation. This is the basis of DSR modeling and imaging methods in geophysics.

(iii) The modeling operator can be written as the composition of a Fourier integral operator representing *depth-to-time conversion*, with a locally invertible canonical relation (Theorem 4.2) and the operator E_1.

It should be mentioned that our Assumption 2 can be quite restrictive. However, the limited aperture of seismic acquisition yields a natural cutoff so that, in general, a large part of the *observed* data can be modeled with the approach presented in this paper.

In general, the downward continuation approach results in a more complete computation of the wave propagation and diffraction in the modeling of seismic reflection data than the one based on the geometrical optics approximation underlying the Kirchhoff approach. Fast algorithms have been designed to solve the DSR equation; as compared with numerical algorithms solving the full wave equation, the advantage of using the DSR equation becomes significant in space dimension 3 (and higher).

The outline of the paper is as follows. In section 2 we discuss one-way acoustic equations. In section 3 we use these to define the upward/downward continuation operator H, and we describe some of its properties. Section 4 contains our result on

depth-to-time conversion. In section 5 we show that the data can be modeled using the downward continuation method. The last section is about the relation between our assumption and the Bolker condition that occurs in the inversion.

2. Directional decomposition, single-square-root equations. Singularities of solutions to the wave equation, which propagate with velocity with nonzero vertical (z) component, are described by a first-order pseudodifferential evolution equation in z. This follows from a well-known factorization argument; see, e.g., [26]. In [22] the approximation of solutions to the wave equation by solutions to an evolution equation in z is discussed. Such an equation is called a one-way wave equation or single-square-root (SSR) equation. We summarize the structure and properties of this one-way wave equation that we need for the upward/downward continuation approach to seismic data processing.

To determine whether the velocity vector at some point of a ray (cf. (1.5)) is close to horizontal, we use the angle with the vertical, defined to be in $[0, \pi/2]$ and given by $\tan(\theta) = \frac{\|\xi\|}{|\zeta|}$. We recall that the propagating singularities are microlocally in the characteristic set given by (1.4). Given a point (z, x, ξ, τ) with $\|\xi\| < c(z,x)^{-1}|\tau|$, there are two solutions ζ to (1.4), given by $\zeta = \pm b$, where $b = b(z, x, \xi, \tau)$ is defined by

$$b(z, x, \xi, \tau) = -\tau\sqrt{c(z,x)^{-2} - \tau^{-2}\xi^2}. \tag{2.1}$$

The sign is chosen such that $\zeta = \pm b$ corresponds to propagation with $\pm\frac{\mathrm{d}z}{\mathrm{d}t} > 0$. There is also an angle associated with (z, x, ξ, τ) given by the solution $\theta \in [0, \pi/2]$ of the equation

$$\sin(\theta) = c(z,x)\|\tau^{-1}\xi\|. \tag{2.2}$$

When this angle is smaller than $\pi/2$ along a ray segment, then the vertical velocity $\frac{\mathrm{d}z}{\mathrm{d}t}$ does not change sign, and the ray segment can be parameterized by z. The maximal z-interval such that $\arcsin(c(z,x)\|\tau^{-1}\xi\|) < \theta$ for given θ along the bicharacteristic determined by the initial values $(z, x, \pm b, \xi, \tau)$ will be denoted by

$$]z_{\min,\pm}, z_{\max,\pm}[=]z_{\min,\pm}(z, x, \xi, \tau, \theta), z_{\max,\pm}(z, x, \xi, \tau, \theta)[; \tag{2.3}$$

see also Figure 1. Furthermore, we define a set

$$I_\theta = \{(z, x, t, \zeta, \xi, \tau) \mid \arcsin(c(z,x)\|\tau^{-1}\xi\|) < \theta, |\zeta| < C|\tau|\}, \tag{2.4}$$

where C is some constant that is everywhere larger than $c(z,x)^{-1}$.

The SSR equation. To obtain a one-way wave equation, the wave equation is written as the following first-order system in z:

$$\frac{\partial}{\partial z}\begin{pmatrix} u \\ \frac{\partial u}{\partial z}\end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -A(z, x, D_x, D_t) & 0\end{pmatrix}\begin{pmatrix} u \\ \frac{\partial u}{\partial z}\end{pmatrix} + \begin{pmatrix} 0 \\ f\end{pmatrix}, \tag{2.5}$$

where $D_x = -\mathrm{i}\frac{\partial}{\partial x}, D_z = -\mathrm{i}\frac{\partial}{\partial z}$, and $A(z, x, D_x, D_t) = c_0(z,x)^{-2}D_t^2 - D_x^2$. Then the system is transformed by using a family of matrix pseudodifferential operators $\mathsf{Q}(z) = \mathsf{Q}(z, x, D_x, D_t)$ with

$$\begin{pmatrix} u_+ \\ u_-\end{pmatrix} = \mathsf{Q}(z)\begin{pmatrix} u \\ \frac{\partial u}{\partial z}\end{pmatrix}, \qquad \begin{pmatrix} f_+ \\ f_-\end{pmatrix} = \mathsf{Q}(z)\begin{pmatrix} 0 \\ f\end{pmatrix}. \tag{2.6}$$

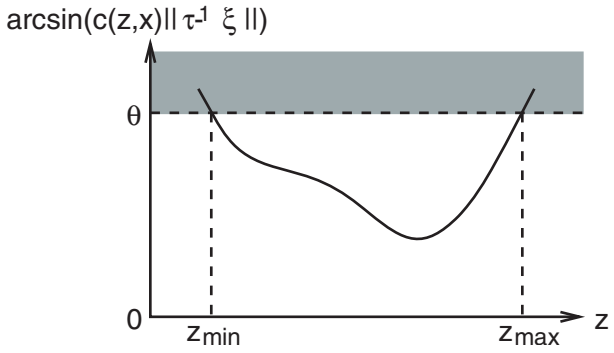


FIG. 1. *Definition of $z_{\min,\pm}$ and $z_{\max,\pm}$, which give the maximal interval where* $\arcsin(c(z,x)\|\tau^{-1}\xi\|)$ *is in the interval* $[0,\theta]$. *Here,* (z,x,ξ,τ) *lies on a bicharacteristic.*

The functions (u_+,u_-) satisfy a pseudodifferential system of equations. Let $\theta_2<\pi/2$ be a given angle. (In the next subsection we need another angle, θ_1, with $0<\theta_1<\theta_2<\pi/2$, hence the subscript 2.) With suitably chosen Q it is shown in [22] that the system that results from applying the transformation (2.6) to (2.5) is diagonal on I_{θ_2}. It then follows that (2.5) is equivalent to two equations of the form

$$\left(\frac{\partial}{\partial z}-\mathrm{i}B_\pm(z,x,D_x,D_t)\right)u_\pm=f_\pm, \tag{2.7}$$

microlocally on I_{θ_2}. These are called the one-way wave or SSR equations. The principal part of $B_\pm$ is equal to $\pm b$, while its subprincipal part depends on the normalization of $\mathsf{Q}(z)$. We choose the normalization such that $B_\pm$ are self-adjoint and Q satisfies

$$\begin{aligned}\mathsf{Q}(z,x,\xi,\tau)&=\tfrac{1}{2}\begin{pmatrix}a^{1/4}&-\mathrm{i}\,\mathrm{sgn}(\tau)a^{-1/4}\\ a^{1/4}&\mathrm{i}\,\mathrm{sgn}(\tau)a^{-1/4}\end{pmatrix}+\text{order}\begin{pmatrix}-\frac12&-\frac32\\-\frac12&-\frac32\end{pmatrix},\\ \mathsf{Q}(z,x,\xi,\tau)^{-1}&=\begin{pmatrix}a^{-1/4}&a^{-1/4}\\ \mathrm{i}\,\mathrm{sgn}(\tau)a^{1/4}&-\mathrm{i}\,\mathrm{sgn}(\tau)a^{1/4}\end{pmatrix}+\text{order}\begin{pmatrix}-\frac32&-\frac32\\-\frac12&-\frac12\end{pmatrix}\end{aligned} \tag{2.8}$$

with $a=a(z,x,\xi,\tau)=c_0(z,x)^{-2}\tau^2-\xi^2$.

It appears that only two components of $\mathsf{Q}(z)$ and $\mathsf{Q}(z)^{-1}$ are needed in the analysis. To clarify this, we first observe that multiplication by $\mathrm{i}\,\mathrm{sgn}(\tau)$ in the frequency domain corresponds to the application of a Hilbert transform with respect to the time variable, which we denote by $\mathcal{H}$. Next, we use the relation between $\mathsf{Q}(z)^*$ and $\mathsf{Q}(z)^{-1}$,

$$\begin{pmatrix}1&0\\0&-1\end{pmatrix}\mathsf{Q}(z)^{-1*}=2\mathsf{Q}(z)\begin{pmatrix}0&-\mathcal{H}\\ \mathcal{H}&0\end{pmatrix},$$

shown to hold microlocally in [22, (59)]. (This relation also appears in [8, (II.49)].) We let

$$Q_+=Q_+(z,x,D_x,D_t)=2\mathsf{Q}_{1,2}\mathcal{H},\ Q_-=Q_-(z,x,D_x,D_t)=-2\mathsf{Q}_{2,2}\mathcal{H},$$

where we choose a convenient normalization such that both $Q_\pm$ have principal symbol $a^{1/4}$. It follows with these definitions that

$$u=Q_+^*u_++Q_-^*u_-, \tag{2.9}$$

$$f_\pm=\mp\tfrac{1}{2}\mathcal{H}Q_\pm f. \tag{2.10}$$

The above procedure does not prescribe the symbol of the operator B_- for $\arcsin(c(z,x)\|\tau^{-1}\xi\|) > \theta_2$. We will assume that B_- is a first-order family of pseudodifferential operators with real homogeneous principal symbol. This implies that the evolution problem (2.7) has well-defined solutions satisfying energy estimates.

Propagation of singularities and introduction of a microlocal cutoff. Here, we discuss how the wave field is approximated by solutions of (2.7). This approximation is valid microlocally on part of the cotangent bundle $T^*\mathbb{R}^{n+1}_{(z,x,t)}$. We consider the approximation of upward traveling waves using the equation for u_-, where we assume that there are only upward traveling singularities at depth z_0, hence $u_+(z_0,\cdot) \in C^\infty$. The treatment of downward traveling waves using the equation for u_+ is analogous.

Consider the initial value problem for $P_{0,-} \stackrel{\text{def}}{=} \partial_z - \mathrm{i}B_-$,

$$P_{0,-}u_- = 0, \quad z < z_0, \quad Q_-^* u_-(z_0,\cdot) = u(z_0,\cdot). \tag{2.11}$$

Let $J_-(z_0,\theta)$ be defined by

$$J_-(z_0,\theta) = \{(z,x,t,\zeta,\xi,\tau) \in I_\theta \,|\, \tau^{-1}\zeta > 0 \text{ and } z_{\max,-}(z,x,\xi,\tau,\theta) \geq z_0\}. \tag{2.12}$$

The solutions to (2.11) agree with the solutions to the original wave equation microlocally on the set $J_-(z_0,\theta_2)$ in the following way. Suppose that $\mathrm{WF}(u) \cap \{z = z_0, \tau^{-1}\zeta < 0\} = \emptyset$ (i.e., at depth z_0 all singularities are propagating in the $-$ direction), and let u_- be a solution to (2.11); then it follows from the propagation of regularity/propagation of singularities result that

$$u \equiv Q_-^* u_- \tag{2.13}$$

microlocally on the set $J_-(z_0,\theta_2)$ [22]. Here, we say that $u \equiv v$ microlocally on a set $\Gamma \subset T^*\mathbb{R}^n$ if $\mathrm{WF}(u-v) \cap \Gamma = \emptyset$.

The solutions to (2.11) have propagating singularities, also in the part of the phase space where $\arcsin(c(z,x)\|\tau^{-1}\xi\|) \geq \theta_2$, but there the singularities of the solution are in general incorrect in the sense that they do not correspond to solutions of the original wave equation. For such singularities we introduce a pseudodifferential cutoff. Let θ_1 be given with $0 < \theta_1 < \theta_2$. We assume we have a pseudodifferential cutoff $\psi_1 = \psi_1(z,z_0,x,D_x,D_t)$ with symbol satisfying

$$\psi_1(z,z_0,x,\xi,\tau) \sim 1 \text{ on } J_-(z_0,\theta_1), \tag{2.14}$$

$$\psi_1(z,z_0,x,\xi,\tau) \in S^\infty \text{ outside } J_-(z_0,\theta_2), \text{ if } z - z_0 > \delta > 0. \tag{2.15}$$

Then we have

$$\psi_1 u \equiv \psi_1 Q_-^* u_-. \tag{2.16}$$

We reformulate this result in terms of the solution operators, the propagators. By $G_{0,-}(z,z_0)$ we will denote the solution operator to the evolution problem (2.11), defined to map $u_-(z_0,\cdot)$ to $u_-(z,\cdot)$. We assume that the full one-way propagator is then given by

$$G_-(z,z_0) = \psi_1(z,z_0) G_{0,-}(z,z_0). \tag{2.17}$$

Here, we let $z < z_0$. This can also be written as a pseudodifferential cutoff applied prior to $G_{0,-}$. We denote this different cutoff also by ψ_1 but with the order of z, z_0 interchanged, so that

$$G_-(z, z_0) = G_{0,-}(z, z_0)\psi_1(z_0, z). \tag{2.18}$$

In this paper this is all we need to know about the pseudodifferential cutoff ψ_1. But it raises the question of an explicit recipe for computing ψ_1: Can it, for example, be computed with a modified evolution equation in depth? This is indeed the case. It was established in [22, 23] that such a pseudodifferential cutoff can be generated by adding a dissipative term to $P_{0,-}$. Instead of $P_{0,-}$ one considers the operator

$$P_- = \partial_z - \mathrm{i}B_\pm(z, x, D_x, D_t) - C(z, x, D_x, D_t) \tag{2.19}$$

with C a first-order pseudodifferential operator with homogeneous, nonnegative real principal symbol, satisfying certain conditions. The operator $\psi_1(z, z_0)$ is then a (z, z_0)-family of pseudodifferential operators with symbol in $S^0_{\rho,1-\rho}(\mathbb{R}^n \times \mathbb{R}^n)$, such that the derivatives $\frac{\partial^{j+k}\psi_1}{\partial z_0^j \partial_z^k}$ are in $S^{(j+k)(1-\rho)}_{\rho,1-\rho}(\mathbb{R}^n \times \mathbb{R}^n)$ for $z \neq z_0$, where ρ can be any number satisfying $\frac{1}{2} < \rho < 1$ (see [23]). For the theory of such operators, see, e.g., [27, 14].

Let the elements $(z, x, t, \zeta, \xi, \tau)$ of the wave front set of f be such that $\tau^{-1}\zeta > 0$ (corresponding to propagation direction $\frac{\partial z}{\partial t} < 0$). Consider u_- defined by

$$u_-(z, \cdot) = \int_z^\infty G_-(z, z_0) \left(\frac{1}{2}\mathcal{H}Q_-(z_0)\right) f(z_0, \cdot)\, \mathrm{d}z_0, \tag{2.20}$$

assuming also that $f = 0$ on a neighborhood of the plane given by z. We have that $Q_-^* u_-(z, \cdot) \equiv u(z, \cdot)$, where u is the solution to (1.1) with f replaced by $(\psi_1(z_0, z) - Q_-^{-1}[Q_-, \psi_1(z_0, z)])f$. Here the square brackets denote a commutator.

We use the notation $\gamma(z, z_0, x_0, t_0, \xi_0, \tau)$ for the bicharacteristic of $P_{0,-}$ parameterized by z. In components we write them as (note that they are time translation invariant)

$$\begin{aligned}\gamma(z, z_0, x_0, t_0, \xi_0, \tau) = (&z, \gamma_x(z, z_0, x_0, \xi_0, \tau), \gamma_t(z, z_0, x_0, \xi_0, \tau) + t_0,\\ &- b(z, \gamma_x, \gamma_\xi, \tau), \gamma_\xi(z, z_0, x_0, \xi_0, \tau), \tau).\end{aligned} \tag{2.21}$$

Properties of G_-. The operator $G_-(z, z_0)$ is a Fourier integral operator with canonical relation

$$\{(\gamma_x, t_0 + \gamma_t, \gamma_\xi, \tau; x_0, t_0, \xi_0, \tau)\} \subset T^*\mathbb{R}^n \times T^*\mathbb{R}^n, \tag{2.22}$$

where $\gamma_x = \gamma_x(z, z_0, x_0, \xi_0, \tau)$ and the same for γ_t, γ_ξ as in (2.21).

The operators $B_\pm$ are self-adjoint. It follows that $G_{0,-}(z, z_0)$ is unitary. But then

$$G_-(z, z_0)^* G_-(z, z_0) = \psi_1(z_0, z)^* \psi_1(z_0, z), \tag{2.23}$$

and $G_-(z, z_0)^* G_-(z, z_0)$ is one microlocally where $\psi_1(z_0, z)$ is one.

Numerical methods for one-way wave propagation are described, e.g., in [9] and [13] and in the references given in those papers.

3. Downward/upward continuation and the DSR equation. In this section we construct the data downward/upward continuation operator, and we establish some of its properties.

Data model. In preparation of the downward/upward continuation approach to seismic data modeling, we rewrite (1.3) in the form

$$\begin{aligned}\delta G(0,r,t,0,s) = \int_{\mathbb{R}^{n-1}\times\mathbb{R}_+}\int_{\mathbb{R}^{n-1}}\int_{-\infty}^{t}\int_{\mathbb{R}_+} & G(0,r,t-t_0,z,x)\\ & \times 2\partial_{t_0}^2 R(z,x,\bar{x},t_0-\bar{t}_0)\\ & \times G(z,\bar{x},\bar{t}_0,0,s)\,\mathrm{d}\bar{t}_0\,\mathrm{d}t_0\,\mathrm{d}\bar{x}\,\mathrm{d}x\mathrm{d}z,\end{aligned} \tag{3.1}$$

where

$$R(z,x,\bar{x},t_0) = \delta(t_0)\delta(x-\bar{x})\left(\frac{\delta c}{c_0^3}\right)\left(z,\frac{\bar{x}+x}{2}\right) \tag{3.2}$$

so that

$$R = E_2E_1c_0^{-3}\delta c \tag{3.3}$$

with the definitions in (1.8). Changing variables of integration, i.e., $t_0 \mapsto t_0' = t_0 - \bar{t}_0$, (3.1) can be written in the form of an integral operator acting on the distribution R,

$$\begin{aligned}\delta G(0,r,t,0,s) = \int_{\mathbb{R}_+}\Big\{\int_{\mathbb{R}}\int_{\mathbb{R}^{n-1}}\int_{\mathbb{R}^{n-1}}\Big(\int_{\mathbb{R}_+} & G(0,r,t-t_0'-\bar{t}_0,z,x)\\ & \times G(z,\bar{x},\bar{t}_0,0,s)\,\mathrm{d}\bar{t}_0\Big)\\ & \times 2\partial_{t_0'}^2 R(z,x,\bar{x},t_0')\,\mathrm{d}\bar{x}\,\mathrm{d}x\,\mathrm{d}t_0'\Big\}\mathrm{d}z,\end{aligned} \tag{3.4}$$

in between the braces, the contributions of which are integrated over depth z.

Using the reciprocity relation of the time-convolution type for the Green's function, we arrive at the integral representation

$$\begin{aligned}\delta G(0,r,t,0,s) = \int_{\mathbb{R}_+}\Big\{\int_{\mathbb{R}^{n-1}}\int_{\mathbb{R}^{n-1}}\int_{\mathbb{R}}\Big(\int_0^{t-t_0} & G(0,r,t-t_0-\bar{t}_0,z,x)\\ & \times G(0,s,\bar{t}_0,z,\bar{x})\,\mathrm{d}\bar{t}_0\Big)\\ & \times 2\partial_{t_0}^2 R(z,x,\bar{x},t_0)\,\mathrm{d}\bar{x}\,\mathrm{d}x\,\mathrm{d}t_0\Big\}\mathrm{d}z.\end{aligned} \tag{3.5}$$

Upon substituting (3.3) into this representation we obtain a mapping $\delta c(z,x) \to \delta G(0,r,t,0,s)$ as encountered in Theorem 1.1. The associated operator kernel appears to propagate singularities from two different scattering points, $\bar{x}$ and x, at each depth z, to the surface at $z=0$.

To arrive at an upward continuation formulation of data modeling, the idea is to substitute in (3.5) for the Green's functions their upward propagating constituents. Thus we replace these Green's functions in accordance with (2.9), (2.10), using only the u_- constituent. So, for the Green's functions $G(z,x,t-t_0,z_0,x_0)$ in (3.5) we substitute the kernel of the operator

$$\frac{1}{2}\mathcal{H}Q_-^*(z,x,D_x,D_t)G_-(z,z_0)Q_-(z_0,x_0,D_{x_0},D_{t_0}), \tag{3.6}$$

viewed as a function of (z, x, t, z_0, x_0, t_0). Naturally, a time convolution of two one-way Green's functions appears. This is the motivation of the definition, by its kernel, of an operator $H(z, z_0), z < z_0$ on functions of (s, r, t),

$$(H(z,z_0))(s,r,t,s_0,r_0,t_0) = \int_{\mathbb{R}} (G_-(z,z_0))(s,t-t_0-\bar{t}_0,s_0)(G_-(z,z_0))(r,\bar{t}_0,r_0)\,\mathrm{d}\bar{t}_0. \tag{3.7}$$

Here $(G_-(z,z_0))(r,\bar{t}_0,r_0,0)$ denotes the distribution kernel of $G_-(z,z_0)$, and $(H(z,z_0))(s,r,t,s_0,r_0,t_0)$ denotes the distribution kernel of $H(z,z_0)$.

As an alternative formulation, we can write the operator $H(z,z_0)$ as the composition of two operators obtained by a tensor product. We recall that if ψ_1, ψ_2 are two operators with kernels $K_{\psi_1}(x,\bar{x})$, $K_{\psi_2}(y,\bar{y})$, then their tensor product $\psi_1 \otimes \psi_2$ has kernel given by the product $K_{\psi_1}(x,\bar{x})K_{\psi_2}(y,\bar{y})$ and maps functions of $(\bar{x},\bar{y})$ to functions of (x,y). We denote the identity operator acting on functions of s by Id_s and similarly for Id_r. If ψ is an operator acting in the (x,t) variables, then we will write ψ_s, ψ_r for the operator acting in the (s,t) variables or the (r,t) variables, respectively. Then we can also write (3.7) as

$$H(z,z_0) = (\mathrm{Id}_s \otimes G_{-,r}(z,z_0)) \circ (G_{-,s}(z,z_0) \otimes \mathrm{Id}_r). \tag{3.8}$$

Since the tensor product of two operators is a well-defined operator, this shows that $H(z,z_0)$ is well defined. If ψ is an operator on functions of (x,t), then we will often simply write ψ_s instead of $\psi_s \otimes \mathrm{Id}_r$. The map $H(z,z_0)$, $z < z_0$, is the upward continuation operator.

If ψ_1 and ψ_2 are operators on functions of (x,t) and are time translation invariant, then $\psi_{1,s}$ and $\psi_{2,r}$ commute, which can be derived by writing out the distribution kernel of the compositions. The factors $G_{-,s}$ and $G_{-,r}$ can be written as compositions $\psi_{1,s}G_{0,-,s}$, $G_{0,-,s}\psi_{1,s}$ (and similarly for r) using (2.17), (2.18). It follows that the operator H can be written as a composition $\psi_2(z,z_0)H_0(z,z_0)$, where H_0 is given by (3.7) with G_- replaced by $G_{-,0}$ and $\psi_2(z,z_0) = \psi_{1,s}(z,z_0)\psi_{1,r}(z,z_0)$. The operator $\psi_2(z,z_0)$ is pseudodifferential with symbol

$$\psi_2(z,z_0,s,r,\sigma,\rho,\tau) = \psi_1(z,z_0,s,\sigma,\tau)\psi_1(z,z_0,r,\rho,\tau). \tag{3.9}$$

We can also write $H(z,z_0) = H_0(z,z_0)\psi_2(z_0,z)$ with ψ_2 defined by (3.9) as well, but with z, z_0 interchanged.

Replacing both source and receiver Green's functions, the result is the replacement of the integral in the parentheses of (3.5) by $-\frac{1}{4}Q^*_{-,s}(0)Q^*_{-,r}(0)H(0,z)Q_{-,s}(z)Q_{-,r}(z)$, where we denote $Q_{-,s}(z) = Q_-(z,s,D_s,D_t)$, and similarly for $Q_{-,r}(z)$. Therefore, we define the DSR modeling operator as

$$F_{\mathrm{D}}\delta c = Q^*_{-,s}(0)Q^*_{-,r}(0)\int_0^Z H(0,z)Q_{-,s}(z)Q_{-,r}(z)\tfrac{1}{2}D_t^2(E_2E_1c_0^{-3}\delta c)(z,\cdot,\cdot,\cdot)\mathrm{d}z, \tag{3.10}$$

where Z is some large number such that $\mathrm{supp}(\delta c)$ is contained in $]0,Z[\times\mathbb{R}^{n-1}$.

In Theorem 5.1 we will show that, in general, F_{D} differs from F by a pseudodifferential cutoff and that under a certain assumption F_{D} models the singular part of the data. We first derive some important properties of H.

The DSR equation. It follows from differentiating expression (3.8) for H with respect to z, using the fact that $B_-(z,r,D_r,D_t)$ and $G_{-,s}(z,z_0)$ commute, that the operator $H_0(z,z_0)$ is a solution operator for the Cauchy initial value problem for the so-called DSR equation, given by

$$\left(\frac{\partial}{\partial z} - \mathrm{i}B_-(z,s,D_s,D_t) - \mathrm{i}B_-(z,r,D_r,D_t)\right) u = 0. \tag{3.11}$$

Using Duhamel's principle (cf. (1.2)), it follows that

$$u(z,s,r,t) = \int_z^Z (H(z,z_0)g(z_0,\cdot,\cdot,\cdot))(s,r,t)\,\mathrm{d}z_0 \tag{3.12}$$

solves the inhomogeneous DSR equation,

$$\begin{aligned}&\left(\frac{\partial}{\partial z} - \mathrm{i}B_-(z,s,D_s,D_t) - \mathrm{i}B_-(z,r,D_r,D_t) - C(z,s,D_s,D_t) - C(z,r,D_r,D_t)\right) u\\ &\quad = g(z,s,r,t), \quad 0 \le z < Z,\end{aligned} \tag{3.13}$$

with zero initial condition, $u(Z,s,r,t) = 0$. It follows from (3.10) that $F_{\mathrm{D}}\delta c$ is given by $Q^*_{-,s}(0)Q^*_{-,r}(0)$ acting on the solution u at $z = 0$ of an inhomogeneous DSR equation with

$$g = Q_{-,s}(z)Q_{-,r}(z)\tfrac{1}{2}D_t^2 R \tag{3.14}$$

and Z such that δc is supported in $0 < \delta < z < Z$ as before.

The bicharacteristics associated with (3.13) are, in the notation of (2.21), given by

$$\begin{aligned}&\Gamma(z,z_0;s_0,r_0,t_0,\sigma_0,\rho_0,\tau) = (\gamma_x(z,z_0,s_0,\sigma_0,\tau),\gamma_x(z,z_0,r_0,\rho_0,\tau),t_0\\ &\quad + \gamma_t(z,z_0,s_0,\sigma_0,\tau) + \gamma_t(z,z_0,r_0,\rho_0,\tau),\gamma_\xi(z,z_0,s_0,\sigma_0,\tau),\gamma_\xi(z,z_0,r_0,\rho_0,\tau),\tau).\end{aligned} \tag{3.15}$$

They are defined on the intersection of the maximal intervals associated with source ray coordinates (z,s,σ,τ) and receiver ray coordinates (z,r,ρ,τ); let θ be given as in the previous section. The intersection will be denoted by $]Z_{\min},Z_{\max}[=]Z_{\min}(z,s,r,\sigma,\rho,\tau,\theta),Z_{\max}(z,s,r,\sigma,\rho,\tau,\theta)[$, where we have

$$Z_{\min}(z,s,r,\sigma,\rho,\tau,\theta) = \max(z_{\min,-}(z,s,\sigma,\tau,\theta),z_{\min,-}(z,r,\rho,\tau,\theta)), \tag{3.16}$$

$$Z_{\max}(z,s,r,\sigma,\rho,\tau,\theta) = \min(z_{\max,-}(z,s,\sigma,\tau,\theta),z_{\max,-}(z,r,\rho,\tau,\theta)). \tag{3.17}$$

Let $g(z,s,r,t)$ be supported in the set $0 < \delta < z < Z$. As mentioned, the map $g \mapsto u$ given by (3.12) maps g to the solution of the inhomogeneous DSR equation (3.13) at $z = 0$. Motivated by (3.10), we define an operator L by modifying (3.12) with pseudodifferential factors $Q_{-,s}, Q_{-,r}$ and setting $z = 0$ as follows:

$$Lg = Q^*_{-,s}(0)Q^*_{-,r}(0)\int_0^Z H(0,z)Q_{-,s}(z)Q_{-,r}(z)g(z,\cdot,\cdot,\cdot)\,\mathrm{d}z. \tag{3.18}$$

Our next result states that H and L are Fourier integral operators and gives a representation of the kernel of H as an oscillatory integral. Consider the following set:

$$\begin{aligned}&\{(\Gamma(0,z,s,r,t,\sigma,\rho,\tau);z,s,r,t,-b(z,s,\sigma,\tau) - b(z,r,\rho,\tau),\sigma,\rho,\tau)\,|\\ &\quad (s,r,t,\sigma,\rho,\tau) \in T^*\mathbb{R}^{2n-1}_{(s,r,t)}, 0 > Z_{\min}(z,s,r,t,\sigma,\rho,\tau,\theta_2)\}.\end{aligned} \tag{3.19}$$

As will be clear from the proof below, this set is a canonical relation. Let $y_0 = (s_0, r_0, t_0)$, $\eta_0 = (\sigma_0, \rho_0, \tau)$. A convenient choice of phase function for the canonical relation is described by Maslov and Fedoriuk [18]. They state that one can always use a subset of the cotangent vector components as phase variables. There is always a set of local coordinates for the canonical relation of the form

$$(z, y_{0I}, \eta_{0J}, s, r, t), \tag{3.20}$$

where $I \cup J$ is a partition of $\{1, \dots, 2n-1\}$. It follows from Theorem 4.21 in Maslov and Fedoriuk [18] that there is a function $S = S(z, y_{0I}, \eta_{0J}, s, r, t)$, such that locally the canonical relation (3.19) is given by

$$y_{0J} = -\frac{\partial S}{\partial \eta_{0J}}, \qquad \zeta = \frac{\partial S}{\partial z}, \tag{3.21}$$

$$\eta_{0I} = \frac{\partial S}{\partial y_{0I}}, \qquad (\sigma, \rho, \tau) = -\frac{\partial S}{\partial (s, r, t)}. \tag{3.22}$$

Here we take into account the fact that we have a canonical relation, which introduces a minus sign for (σ, ρ, τ).

LEMMA 3.1. *$H(z, z_0)$ is a Fourier integral operator with canonical relation*

$$\{(\Gamma(z, z_0, s, r, t, \sigma, \rho, \tau); s, r, t, \sigma, \rho, \tau) \mid (s, r, t, \sigma, \rho, \tau) \in T^*\mathbb{R}^{2n-1}_{(s,r,t)} \backslash 0,\ z_0 > Z_{\min}(z, s, r, t, \sigma, \rho, \tau, \theta_2)\}. \tag{3.23}$$

The operator L is a Fourier integral operator with canonical relation (3.19). *The kernel of $H(0, z)$ admits microlocally an oscillatory integral representation with phase variables η_{0J}, given by*

$$(H(0, z))(s_0, r_0, t_0, s, r, t) = (2\pi)^{-(2n-1+|I|)/2} \int A(z, y_0, \eta_{0J}, s, r, t) \exp[\mathrm{i}(S(z, y_{0I}, \eta_{0J}, s, r, t) + \langle \eta_{0J}, y_{0J} \rangle)] \, \mathrm{d}\eta_{0J} \tag{3.24}$$

such that the principal part a of the amplitude A satisfies

$$|a(z, y_0, \eta_{0J}, s, r, t)| = \left| \frac{\partial(\sigma, \rho, \tau)}{\partial(y_{0I}, \eta_{0J})} \right|^{1/2} \tag{3.25}$$

with

$$(\sigma(z, y_{0I}, \eta_{0J}, s, r, t), \rho(z, y_{0I}, \eta_{0J}, s, r, t), \tau(z, y_{0I}, \eta_{0J}, s, r, t)) = -\frac{\partial S}{\partial (s, r, t)}(z, y_{0I}, \eta_{0J}, s, r, t) \tag{3.26}$$

in accordance with (3.22).

Proof. The operators $G_{-,s}(z, z_0)$ and $G_{-,r}(z, z_0)$ are Fourier integral operators as noted at the end of section 2 (subject to the substitution of x by s or r, respectively). We consider $G_{-,s}(z, z_0)$. Locally there are Maslov phase functions for its canonical relation (cf. (2.22)), similar to the one described above, here with phase variables $(\tau, \sigma_{0J'})$, where $I' \cup J'$ is a partition of $\{1, \dots, n-1\}$. Thus $G_{-,s}(z, z_0)$ is a locally

finite sum $\sum_j G^{(j)}_{-,s}(z,z_0)$, where the kernels of $G^{(j)}_{-,s}(z,z_0)$ admit oscillatory integral representations of the form

$$(3.27)\quad (G^{(j)}_{-,s}(z,z_0))(s,t,s_0) = \int A'(s,s_0,\sigma_{0,J'},\tau)\exp[\mathrm{i}(S'(z,z_0,s,s_{0I'},\sigma_{0J'},\tau)-\langle\sigma_{0J'},s_{0J'}\rangle-\tau t)]\mathrm{d}\sigma_{0J'}\mathrm{d}\tau.$$

We denote the canonical relation of $G^{(j)}_{-,s}(z,z_0)$ by $\Lambda^{(j)}_s$ (cf. (2.22)). Similarly, we have $G_{-,r}(z,z_0)=\sum_k G^{(k)}_{-,r}(z,z_0)$ in which the kernels of $G^{(k)}_{-,r}(z,z_0)$ admit oscillatory integral representations of the above type with phase variables $(\tau,\rho_{0J''})$, amplitude A'', and phase function $S''(z,z_0,r,r_{0I''},\rho_{0J''},\tau)-\langle\rho_{0J''},r_{0J''}\rangle-\tau t$. We denote the canonical relation of $G^{(k)}_{-,r}(z,z_0)$ by $\Lambda^{(k)}_r$. But then the kernel of $H(z,z_0)$ is given by a sum $\sum_{j,k}H^{(j,k)}(z,z_0)$. Entering expressions of the type (3.27) for $G^{(j)}_{-,s}(z,z_0)$ and $G^{(k)}_{-,r}(z,z_0)$ into (3.7), and performing the $\bar{t}_0$ integration, we find the following expression for the kernel of $H^{(j,k)}(z,z_0)$:

$$\begin{aligned}(3.28)\quad (H^{(j,k)}(z,z_0))(s,r,t,s_0,r_0,t_0) = \int 2\pi A'(s,s_0,\sigma_{0,J'},\tau)A''(r,r_0,\rho_{0,J''},\tau)\\ \times\exp[\mathrm{i}(S'(z,z_0,s,s_{0I'},\sigma_{0J'},\tau)-\langle\sigma_{0J'},s_{0J'}\rangle\\ +S''(z,z_0,r,r_{0I''},\rho_{0J''},\tau)-\langle\rho_{0J''},r_{0J''}\rangle-\tau t)]\,\mathrm{d}\sigma_{0J'}\,\mathrm{d}\rho_{0J''}\,\mathrm{d}\tau.\end{aligned}$$

It is not difficult to verify that $-\mathrm{i}$ times the argument in the exponent is a nondegenerate phase function. Because A' and A'' are symbols supported inside a region with $\|\sigma\|<C|\tau|$ and $\|\rho\|<C|\tau|$ it follows that $A'A''$ is a symbol and that (3.28) is a Fourier integral operator. From the phase function it follows that the canonical relation of $H^{(j,k)}(z,z_0)$ is given by the points

$$(s,r,t_0+t_1+t_2,\sigma,\rho,\tau;s_0,r_0,t_0,\sigma_0,\rho_0,\tau)$$

with

$$(s,t_1,\sigma,\tau;s_0,0,\sigma_0,\tau)\in\Lambda^{(j)}_s \text{ and } (r,t_2,\rho,\tau;r_0,0,\rho_0,\tau)\in\Lambda^{(k)}_r.$$

Taking the union over (j,k) results in (3.23).

Using (3.18) and the fact that H is given by a sum of terms of the form (3.28), it also follows that L is a Fourier integral operator with canonical relation (3.19), as usual for the solution operators of first-order hyperbolic equations.

The phase function $S(z,y_{0I},\eta_{0J},s,r,t)-\langle\eta_{0J},y_{0J}\rangle)$, with S as described in (3.21)–(3.22), describes locally the canonical relation of $H(0,z)$. Therefore the kernel of $H(0,z)$ has microlocally an oscillatory integral representation of the form

$$\begin{aligned}(3.29)\quad (H(0,z))(y_0,s,r,t) = (2\pi)^{-(2n-1+|I|)/2}\\ \times\int A(z,y_0,\eta_{0J},s,r,t)\exp[\mathrm{i}(S(z,y_{0I},\eta_{0J},s,r,t)+\langle\eta_{0J},y_{0J}\rangle)]\,\mathrm{d}\eta_{0J}.\end{aligned}$$

Then the adjoint $H(0,z)^*$ has amplitude $\overline{A(z,y_0,\eta_{0J},s,r,t)}$ and phase $-S(z,y_{0I},\eta_{0J},s,r,t)-\langle\eta_{0J},y_{0J}\rangle$. Hence, the kernel of the composition $H(0,z)^*H(0,z)$ has the

oscillatory integral representation

$$(3.30)\quad (2\pi)^{-(2n-1)}\int \overline{A(z,y_0,\eta_{0J},s',r',t')}A(z,y_0,\eta_{0J},s,r,t)\\ \times\exp(\mathrm{i}[-S(z,y_{0I},\eta_{0J},s',r',t')+S(z,y_{0I},\eta_{0J},s,r,t)])\,\mathrm{d}y_{0I}\mathrm{d}\eta_{0J}.$$

We expand the phase as a function of (s',r',t') in a Taylor series about (s,r,t) and identify the gradient

$$(3.31)\quad -\frac{\partial S}{\partial(s,r,t)}(z,y_{0I},\eta_{0J},s,r,t)\\ =(\sigma(z,y_{0I},\eta_{0J},s,r,t),\rho(z,y_{0I},\eta_{0J},s,r,t),\tau(z,y_{0I},\eta_{0J},s,r,t)).$$

Applying a change of variables, $(y_{0I},\eta_{0J})\mapsto(\sigma,\rho,\tau)$, the phase takes the form

$$(3.32)\qquad \langle(\sigma,\rho,\tau),(s'-s,r'-r,t'-t)\rangle.$$

In the text preceding (2.23) it was noted that $G_{0,-}(z,z_0)$ is unitary. It follows using (3.8) that $H_0(z,z_0)$ is also unitary. Therefore, the operator $H(0,z)^*H(0,z)$ must be a pseudodifferential operator (in (s,r,t)) with symbol 1 in the set of (s,r,t,σ,ρ,τ), where ψ_2 is equal to 1. We conclude that the principal part a of the amplitude A is given by

$$(3.33)\qquad |a(z,y_0,\eta_{0J},s,r,t)|=\left|\frac{\partial(\sigma,\rho,\tau)}{\partial(y_{0I},\eta_{0J})}\right|^{1/2}. \qquad \square$$

4. Depth-to-time conversion. For $h=h(z,s,r)$ we consider the mapping

$$(4.1)\qquad K\ :\ h\mapsto Q^*_{-,s}(0)Q^*_{-,r}(0)\int_0^Z H(0,z)Q_{-,s}(z)Q_{-,r}(z)(E_2h)(z,\cdot,\cdot,\cdot)\mathrm{d}z;$$

we have $K=LE_2$ (cf. (3.18)). The DSR modeling operator (cf. (3.10)) is then given by

$$(4.2)\qquad F_{\mathrm{D}}\delta c=\frac{1}{2}D_t^2KE_1c_0^{-3}\delta c.$$

This factorization is exploited in seismic applications such as imaging.

First we make the following observation. We use the notation $\Theta=\Theta(z,s,r,\sigma,\rho,\tau)$ for the sum $-b(z,s,\sigma,\tau)-b(z,r,\rho,\tau)$ appearing in the canonical relation (3.19) of L,

$$(4.3)\qquad \Theta(z,s,r,\sigma,\rho,\tau)=-b(z,s,\sigma,\tau)-b(z,r,\rho,\tau).$$

Because of expression (2.1) the map $\tau\mapsto\zeta=\Theta$ is strictly monotone when Θ is real. Taking as domain only the τ where the two square roots are real, we find the following lemma. The inverse of this map will be denoted by Θ^{-1}.

LEMMA 4.1. *Suppose (z,s,r,σ,ρ) are given, let $c=\max(c(z,s)\|\sigma\|,c(z,r)\|\rho\|)$, and let $d=\sqrt{|\sigma^2\frac{c(z,s)^2}{c(z,r)^2}-\rho^2|}$ if $c(z,s)\|\sigma\|\geq c(z,r)\|\rho\|$ and $d=\sqrt{|\rho^2\frac{c(z,r)^2}{c(z,s)^2}-\sigma^2|}$ otherwise. The map $\tau\mapsto\Theta(z,s,r,\sigma,\rho,\tau)$ is a diffeomorphism $]-\infty,-c[\cup]c,\infty[\ \to\ ]-\infty,-d[\cup]d,\infty[$.*

The maximal depth $Z_{\max}$ associated with $(z,s,r,\sigma,\rho,\tau,\theta)$ also has an associated maximal time, given by

$$(4.4)\qquad T_{\max}(z,s,r,\sigma,\rho,\tau,\theta)=-\Gamma_t(Z_{\max}(z,s,r,\sigma,\rho,\tau,\theta),s,r,\sigma,\rho,\tau).$$

We define a subset Ω_θ of $T^*\mathbb{R}^{2n-1}_{(s,r,t)}$, such that t is bounded by $T_{\max}$,

$$\Omega_\theta = \{(s,r,t,\sigma,\rho,\tau)\,|\,0<t<T_{\max}(0,s,r,\sigma,\rho,\tau,\theta)\}. \tag{4.5}$$

We have the following result about K.

THEOREM 4.2. *The operator K is microlocally a Fourier integral operator with canonical relation consisting of a set of points,*

$$\{(\Gamma(0,z,s,r,0,\sigma,\rho,\tau);z,s,r,\Theta(z,s,r,\sigma,\rho,\tau),\sigma,\rho)\,| \\ z,s,r,\sigma,\rho,\tau\in\mathbb{R}^{4n-2}, 0<Z_{\min}(z,s,r,\sigma,\rho,\tau,\theta_2)\}. \tag{4.6}$$

This canonical relation is the graph of an invertible map Σ:

$$\{(z,s,r,\zeta,\sigma,\rho)\,|\,0\langle z,0\rangle Z_{\min}(z,s,r,\sigma,\rho,\Theta^{-1}(z,s,r,\zeta,\sigma,\rho),\theta_2)\}\to\Omega_{\theta_2}. \tag{4.7}$$

The map K converts depth to time, which is indeed the way seismologists often look at modeling.

Proof. The operator K is the composition of (3.18) and E_2. The first is a Fourier integral operator with canonical relation given by (3.19). The operator E_2 is a Fourier integral operator with canonical relation given by

$$\{(z,s,r,0,\zeta,\sigma,\rho,\tau;z,s,r,\zeta,\sigma,\rho)\,|\,(z,s,r,\zeta,\sigma,\rho)\in T^*\mathbb{R}^{2n-1}_{(z,s,r)}\backslash 0,\tau\in\mathbb{R}\backslash\{0\}\}. \tag{4.8}$$

In general, the composition of two canonical relations $\Lambda_1\subset T^*(X\times Y)\backslash 0$, $\Lambda_2\subset T^*(Y\times Z)\backslash 0$, X,Y,Z open subsets of $\mathbb{R}^{n_X},\mathbb{R}^{n_Y}$, respectively, $\mathbb{R}^{n_Z}$, is said to be transversal if

$$\Lambda_1\times\Lambda_2 \text{ intersects } T^*X\times(\operatorname{diag}T^*Y)\times T^*Z \text{ transversally.}$$

In the particular case of the canonical relations of L and E_2, their composition is transversal if at the solution τ of

$$-b_s-b_r=\zeta \tag{4.9}$$

we have $\frac{\mathrm{d}\Theta}{\mathrm{d}\tau}\neq 0$; see, e.g., Theorem 2.4.1 in [10]. Because by the previous lemma this is the case, it then follows that the composition LE_2, hence K, is a Fourier integral operator. The composition of the canonical relations is equal to (4.6).

The canonical relation of K is parameterized by (z,s,r,σ,ρ,τ) in a subset of $\mathbb{R}^{4n-2}$. To show that it is invertible we must show that the projections of (4.6) on the two sets given in (4.7) are both diffeomorphisms. By the previous lemma this is clear for the projection on the right-hand side of (4.7). For the projection on the left-hand side of (4.7) it follows from Lemma 25.3.6 of [15] and the fact that the right projection has maximal rank that the linearization of this projection is invertible. Thus it remains to be shown that the equation

$$(s_0,r_0,t_0,\sigma_0,\rho_0,\tau_0)=\Gamma(0,z,s,r,0,\sigma,\rho,\tau) \tag{4.10}$$

determines a unique point (z,s,r,σ,ρ,τ) when $(s_0,r_0,t_0,\sigma_0,\rho_0,\tau_0)$ is in Ω_{θ_2}, the right-hand side of (4.7). The point $(s_0,r_0,t_0,\sigma_0,\rho_0,\tau_0)$ determines a DSR bicharacteristic $\Gamma(z,0,s_0,r_0,t_0,\sigma_0,\rho_0,\tau_0)$. The t component will be denoted here by $\Gamma_t = t_0-\gamma_t(z,0,s_0,\sigma_0,\tau)-\gamma_t(z,0,r_0,\rho_0,\tau)$. We have a solution to (4.10) if and only if

$$\Gamma(z,0,s_0,r_0,t_0,\sigma_0,\rho_0,\tau_0)-(s,r,0,\sigma,\rho,\tau)=0. \tag{4.11}$$

In particular, we have that $\Gamma_t(z, 0, s_0, r_0, t_0, \sigma_0, \rho_0, \tau_0) = 0$. Because Γ_t depends strictly monotonically on z, this equation uniquely determines z. The other equations uniquely determine s, r, σ, ρ, τ. If $z < Z_{\max}(0, s_0, r_0, \sigma_0, \rho_0, \tau)$, then there is a DSR bicharacteristic connecting $(s_0, r_0, t_0, \sigma_0, \rho_0, \tau)$ at depth 0 with $(s, r, t, \sigma, \rho, \tau) = \Gamma(z, 0, s, r, t_0, \sigma, \rho, \tau)$ at depth z; hence it follows that then $0 > Z_{\min}(z, s, r, \sigma, \rho, \tau)$, and vice versa. So using definition (4.4) this point $(z, s, r, \sigma, \rho, \tau)$ is such that $0 > Z_{\min}(z, s, r, \sigma, \rho, \tau, \theta_2)$ precisely when $t < T_{\max}(0, s_0, r_0, \sigma_0, \rho_0, \tau_0, \theta_2)$. This completes the proof of the theorem. □

5. Modeling in the single-scattering approximation. The replacement of the wave equation Green's function by a pair of one-way Green's functions leads to a cutoff in the modeling of the scattered wave field. To describe when all the singularities of the data are modeled by the DSR method, we need the following assumption. We use some angle θ, $0 < \theta < \pi/2$, with the vertical as introduced in section 2.

ASSUMPTION 2 (DSR assumption). *If $(z, x) \in X$ and $\alpha, \beta \in S^{n-1}$, $t_s, t_r > 0$ depending on (z, x, α, β) are such that $\eta_z(t_s, z, x, \beta, \tau) = \eta_z(t_r, z, x, \alpha, \tau) = 0$, then*

$$c(z, x)^{-1} \frac{\partial \eta_z}{\partial t}(t, z, x, \beta, \tau) < -\cos(\theta), t \in [0, t_s], \tag{5.1}$$

$$c(z, x)^{-1} \frac{\partial \eta_z}{\partial t}(t, z, x, \alpha, \tau) < -\cos(\theta), t \in [0, t_r]. \tag{5.2}$$

It is clear that this assumption is stronger than Assumption 1. In general the set of rays violating this assumption is not small, but it can contain an open subset of the canonical relation (1.7), depending on the properties of the background medium. This limits the applicability of the method discussed here, which however is still useful in many cases, as discussed in the introduction.

In the following theorem we give the DSR modeling formula, and we give the result in terms of a cutoff acting on $F\delta c$. The symbol $\psi_2(0, z, s, r, 0, \sigma, \rho, \tau)$ can be pulled back to a symbol that is a function of $(s, r, t, \sigma, \rho, \tau)$ by the inverse of the map Σ given by (4.6), (4.7).

THEOREM 5.1. *If Assumption 2 is satisfied with $\theta = \theta_1$, then $F_D \delta c \equiv F\delta c$. There is a pseudodifferential operator $\psi_D = \psi_D(s, r, t, D_s, D_r, D_t)$ with principal symbol given by the pull back mentioned just above of ψ_2, that is, 1 on Ω_{θ_1}, and is in $S^{-\infty}$ outside Ω_{θ_2}, such that*

$$F_D \delta c \equiv \psi_D F \delta c. \tag{5.3}$$

Proof. We reconsider the modeling operator F of Theorem 1.1 and use its description by (3.5). In this proof, we denote by G_s the map from a function $f(z, s, t)$ to $(G_s f)(z, s, t) = \int_{\mathbb{R}\times\mathbb{R}^{n-1}\times\mathbb{R}} G(z, s, t - t_0, z_0, s_0) f(z_0, s_0, t_0)\, dz_0 ds_0 dt_0$; cf. (1.2). Motivated by (3.5) and the introduction of H in (3.8), we consider the operator $M = (\mathrm{Id}_r \otimes G_s) \circ (G_r \otimes \mathrm{Id}_s)$, which maps functions of (z_s, z_r, s, r, t) to functions of (z_s, z_r, s, r, t). In our application, we consider M as a map of functions in $z > \delta$ to functions on a small neighborhood of $z_s = 0, z_r = 0$. By an argument similar to the first part of the proof of Lemma 3.1, it follows that M is a Fourier integral operator, with canonical relation consisting of a set of points

$$\{(\eta_{z,s}, \eta_{z,r}, \eta_{x,s}, \eta_{x,r}, t + t_s + t_r, \eta_{\zeta,s}, \eta_{\zeta,r}, \eta_{\xi,s}, \eta_{\xi,r}, \tau; z_s, z_r, s, r, t, \zeta_s, \zeta_r, \rho, \sigma, \tau)\}, \tag{5.4}$$

where $(\zeta_s,\sigma) = -\tau c(z_s,s)^{-1}\beta$, $(\zeta_r,\rho) = -\tau c(z_r,r)^{-1}\alpha$, $\eta_{z,s} = \eta_z(t_s,z_s,s,\beta,\tau)$, and similar for the other components, and for the r-components, cf. (1.6); $\alpha,\beta \in S^{n-1}$ as in Theorem 1.1.

Denote by $R_4(z)$ the restrictions to $z_s = z$ and $z_r = z$ of functions $f(z_s,z_r,s,r,t)$ and by $E_4(z)$ the map that maps a function $f(s,r,t)$ to $(E_4(z)f)(z_s,z_r,s,r,t) = \delta(z_s - z)\delta(z_r - z)f(s,r,t)$. It follows, from writing out the distribution kernel of M, and using the remark below (2.20), that for distributions in (z_s,z_r,s,r,t) with singularities with $\tau^{-1}\zeta_s > 0$ and $\tau^{-1}\zeta_r > 0$, we have

$$R_4(0)ME_4(z)\psi_2'(z,0) = -\frac{1}{4}Q_{-,s}^*(0)Q_{-,r}^*(0)H(0,z)Q_{-,s}(z)Q_{-,r}(z), \tag{5.5}$$

modulo a regularizing operator, where $\psi_2' = \psi_2 - Q_{-,s}^{-1}Q_{-,r}^{-1}[Q_{-,s}Q_{-,r},\psi_2]$. Since for F the rays come from one side of the surface $z = 0$, we can apply this to (3.5). Denote by E_5 the map that maps a function $f(z,s,r,t)$ to $(E_5(z)f)(z_s,z_r,s,r,t) = \delta(z_s - z_r)f(\frac{z_s+z_r}{2},s,r,t)$. It follows that we have

$$R_4(0)ME_5\psi_2'(z,0)E_2E_1(c_0^{-3}\delta c) \equiv \tfrac{1}{4}KE_1(c_0^{-3}\delta c), \tag{5.6}$$

modulo a regularizing operator.

We can find an operator $\psi'(z,s,r,D_z,D_s,D_r)$ such that the principal symbol $\psi_2' - \psi'$ is zero on the set $\zeta = -b(z,s,\sigma,\tau) - b(z,r,\rho,\tau)$. Namely, first set (for the principal symbol) $\psi'(z,s,r,\zeta,\sigma,\rho) = \psi_2'(z,0,s,r,\sigma,\rho,\Theta^{-1})$. Then the map $ME_5(\psi_2' - \psi')$ is a Fourier integral operator with highest-order amplitude equal to zero. With lower-order terms in ψ' we find that we can replace ψ_2' in (5.6) by an operator $\psi' = \psi'(z,s,r,D_z,D_s,D_r)$. The operator ψ' commutes with E_2. Hence, if $h = h(z,s,r)$, we have that

$$R_4(0)ME_5E_2\psi'h = Kh, \tag{5.7}$$

modulo a smoothing operator. Because of equality (5.7), $R_4(0)ME_5E_2$ is an invertible Fourier integral operator with canonical relation given by (4.6), microlocally on a neighborhood of the set where ψ' is not in $S^{-\infty}$. Now define microlocally on a neighborhood of the set where ψ' is not in $S^{-\infty}$,

$$\psi_{\mathrm{D}} = R_4(0)ME_5E_2\psi'(R_4(0)ME_5E_2)^{-1}. \tag{5.8}$$

By Egorov's theorem this is a pseudodifferential operator with symbol as in the theorem and we have

$$\psi_{\mathrm{D}}R_4(0)ME_5E_2 = K, \tag{5.9}$$

modulo a smoothing operator. It follows that (5.3) is satisfied. □

6. The Bolker condition. It follows from (4.2) and from Theorem 4.2 that the canonical relation of F_{D} in (3.10) satisfies Guillemin's [11] Bolker condition: The projection of the canonical relation (1.7) on $T^*Y\backslash 0$ is an embedding.

Indeed, Assumption 2 is stronger than this condition, as can be seen from the

arguments in the proof of Theorem 4.2. This fact will be important for the inverse scattering based on modeling data by F_{D}.

Acknowledgments. We thank Margaret Cheney for her interest in the work presented in this paper and for encouraging us to write up the results in the present form. We also thank the members of the Mathematical Sciences Research Institute, and in particular Gunther Uhlmann, for providing a very stimulating environment during the Inverse Problems program in Fall 2001.

REFERENCES

[1] A. J. Berkhout and C. P. A. Wapenaar, *A unified approach to acoustical reflection imaging. II: The inverse problem*, J. Acoust. Soc. Amer., 93 (1993), pp. 2017–2023.
[2] G. Beylkin, *Imaging of discontinuities in the inverse scattering problem by inversion of a causal generalized Radon transform*, J. Math. Phys., 26 (1985), pp. 99–108.
[3] B. Biondo and G. Palacharla, *3-D prestack migration of common-azimuth data*, Geophysics, 61 (1996), pp. 1822–1832.
[4] N. Bleistein, J. K. Cohen, and J. W. Stockwell, *Mathematics of Multidimensional Seismic Imaging, Migration, and Inversion*, Springer-Verlag, New York, 2001.
[5] J. F. Claerbout, *Imaging the Earth's Interior*, Blackwell Scientific Publications, Oxford, UK, 1985.
[6] R. W. Clayton, *Common Midpoint Migration*, Technical report SEP-14, Stanford University, Stanford, CA, 1978.
[7] M. D. Collins, and R. B. Evans, *A two-way parabolic equation for acoustic backscattering in the ocean*, J. Acoust. Soc. Amer., 91 (1992), pp. 1357–1368.
[8] M. V. de Hoop, *Generalization of the Bremmer coupling series*, J. Math. Phys., 37 (1996), pp. 3246-3282.
[9] M. V. de Hoop, J. H. Le Rousseau, and R.-S. Wu, *Generalization of the phase-screen approximation for the scattering of acoustic waves*, Wave Motion, 31 (2000), pp. 43–70.
[10] J. J. Duistermaat, *Fourier Integral Operators*, Birkhäuser Boston, Boston, 1996.
[11] V. Guillemin, *On some results of Gel'fand in integral geometry*, in Pseudodifferential Operators and Applications, AMS, Providence, RI, 1985, pp. 149–155.
[12] G. R. Hadley, *Multistep method for wide-angle beam propagation*, Opt. Lett., 17 (1992), pp. 1743–1745.
[13] L. Halpern and L. N. Trefethen, *Wide-angle one-way wave equations*, J. Acoust. Soc. Amer., 84 (1988), pp. 1397–1404.
[14] L. Hörmander, *The Analysis of Linear Partial Differential Operators*, vol. 3, Springer-Verlag, Berlin, 1985.
[15] L. Hörmander, *The Analysis of Linear Partial Differential Operators*, vol. 4, Springer-Verlag, Berlin, 1985.
[16] S. Jin, C. C. Mosher, and R.-S. Wu, *Offset-domain pseudoscreen prestack depth migration*, Geophysics, 67 (2002), pp. 1895–1902.
[17] A. P. E. ten Kroode, D. J. Smit, and A. R. Verdel, *A microlocal analysis of migration*, Wave Motion, 28 (1998), pp. 149–172.
[18] V. P. Maslov and M. V. Fedoriuk, *Semi-Classical Approximation in Quantum Mechanics*, D. Reidel, Boston, 1981.
[19] A. M. Popovici, *Prestack migration by split-step DSR*, Geophysics, 61 (1996), pp. 1412–1416.
[20] Rakesh, *A linearized inverse problem for the wave equation*, Comm. Partial Differential Equations, 13 (1988), pp. 573–601.
[21] P. S. Schultz and J. W. C. Sherwood, *Depth migration before stack*, Geophysics, 45 (1980), pp. 376–393.
[22] C. C. Stolk, *A pseudodifferential equation with damping for one-way wave propagation in inhomogeneous acoustic media*, Wave Motion, 40 (2004), pp. 111–121.
[23] C. C. Stolk, *Parametrix for a hyperbolic initial value problem with dissipation in some region*, Asymptot. Anal., to appear.
[24] C. C. Stolk, and M. V. de Hoop, *Seismic Inverse Scattering in the Downward Continuation Approach*, 2004. preprint.
[25] F. D. Tappert, *The parabolic approximation method*, in Wave Propagation and Underwater Acoustics, Lecture Notes in Phys. 70, J. B. Keller and J. S. Papadakis, eds., Springer-Verlag, New York, 1977, pp. 224–287.

[26] M. E. Taylor, *Reflection of singularities of solutions to systems of differential equations*, Comm. Pure Appl. Math., 28 (1975), pp. 457–478.
[27] M. E. Taylor, *Pseudodifferential Operators*, Princeton University Press, Princeton, NJ, 1981.
[28] V. H. Weston, *Factorization of the wave equation in higher dimensions*, J. Math. Phys., 28 (1987), pp. 1061–1068.

SIAM J. APPL. MATH.
Vol. 65, No. 4, pp. 1407–1419

SOME PROPERTIES OF THE CAPACITY VALUE FUNCTION*

B. A. CHIERA†, A. E. KRZESINSKI‡, AND P. G. TAYLOR†

Abstract. In a previous paper [B. A. Chiera and P. G. Taylor, *Probab. Engrg. Inform. Sci.*, 16 (2002), pp. 513–522], two of the authors developed a method for ascribing a value to an extra unit of capacity on a telecommunications link. Specifically, they expressed the value of an extra unit of capacity as a function of current capacity, current occupancy, and a planning horizon. The intention was to use this function as an ingredient in a bandwidth reallocation scheme for ensuring efficient operation of a telecommunications network.

Unfortunately, direct evaluation of the function requires numerical inversion of a Laplace transform expressed in terms of Charlier polynomials, a task that is beyond the processing capabilities of typical switches in today's telecommunications networks. Because of this, it is desirable to have more easily computable methods of either calculating or approximating the capacity value function. We develop two approaches to this problem: the first is a recursive method of computing the Laplace transform of the capacity value function, and the second is a linear approximation to the capacity value function itself.

Key words. Erlang loss system, capacity value function, approximation

AMS subject classifications. 60K30, 41A10

DOI. 10.1137/S0036139903430859

1. Introduction.

A topical issue in recent telecommunications literature revolves around the problem of how network resources should be reallocated from underutilized to overutilized links. One possible way to approach this problem is by calculating the expected amount of extra revenue that would be earned on a specified end-to-end link over some planning horizon if extra capacity were present on that link. A reallocation scheme can then be designed in which capacity is transferred from places in the network where its earning capacity is temporarily low to places where it is high. We would expect that such a scheme should maximize the overall rate at which the network earns revenue.

Previous approaches to allocating a value to capacity in a telecommunications context cover an extensive range. They include a simple model in which the value of capacity is an exponential function of the amount of free capacity [8], a model that values customers in a dynamic loss system [10], multimarket pricing scenarios that price resources on the basis of current and future usage [7], and constrained producer-consumer linear programming models requiring simultaneous solution [11].

For any capacity valuation model to be implementable in practice, it must satisfy the following criteria:

1. It must require only local information; that is, knowledge of the current state of the entire network should not be required to compute the value of capacity on any given route.

*Received by the editors July 2, 2003; accepted for publication (in revised form) September 8, 2004; published electronically April 26, 2005. This research was supported by Australian Research Council grant A10033153.

http://www.siam.org/journals/siap/65-4/43085.html

†Department of Mathematics and Statistics, University of Melbourne, Victoria 3010, Australia (bchiera@ms.unimelb.edu.au, ptaylor@ms.unimelb.edu.au).

‡Department of Computer Science, University of Stellenbosch, 7600 Stellenbosch, South Africa (aek1@cs.sun.ac.za).

2. It should be scalable. This means that the model must allow for, and be easily adaptable to, changes in network size.
3. It should involve calculations that can be efficiently implemented by network switches.

Some of the above pricing schemes satisfy the first two requirements. However, with the exception of the exponential pricing function given in [8], it is not always clear that these schemes can be implemented and run in real time on network switches. In particular, the WALRAS model proposed by Wellman [11] possesses the disadvantage that the time for computation can potentially exceed the time it takes for the underlying market to change.

While the pricing function suggested by [8] does satisfy all three requirements, it is not derived from a realistic model for the value of capacity. Thus it is unlikely that a reallocation scheme based on this valuation would maximize the overall good of either the subnetwork to which the link belongs or the network as a whole.

An alternative method for valuing capacity was proposed by two of the authors in [4]. This model, derived from the basic tenets of renewal theory, is designed to reflect the difference between the amount of lost revenue that would ensue if a particular link were allocated an extra unit of capacity and if it were not. This difference can be thought of as the "value to the link" of the extra capacity.

In ongoing work, the authors are planning to incorporate this value function into a capacity reallocation scheme. The method allows for reallocations to occur between multiple-link routes and their constituent single-link routes, with the result that the model is decentralized and scalable while at the same time able to maximize the revenue in each part of the network.

A potential drawback, however, is that the calculation of the value function involves numerical computations that are beyond the capacity of today's switches. The numerical problems manifest themselves at two stages. First, it is necessary to calculate the Laplace transform of the capacity value function, which is expressed in terms of Charlier polynomials, in a stable and efficient manner. Second, it is necessary to invert the Laplace transform numerically.

In this paper, we shall address both of the above-mentioned difficulties. By concentrating on computing ratios of the Charlier polynomials, rather than the polynomials themselves, we shall develop a stable recursive method for computing the Laplace transform of the capacity value function. This can be used in conjunction with an efficient method of transform inversion if one is available. For situations when no such method is available, we propose a linear approximation to the capacity value function itself, together with a bound on its accuracy. Furthermore, we present efficient recursions for calculating the coefficients in the linear approximation. The value of capacity as given by the linear approximation can be thought of as consisting of an initial set-up cost and then a fixed per-unit cost. As such, it may prove to be useful in formulating optimization problems involving the allocation of capacity.

In section 2 we shall give a brief description of the value function of [4] and establish a preliminary result. In section 3 we shall introduce our recursion for the Laplace transform of the capacity value function. Section 4 contains our approximation of the capacity value function itself together with a discussion of the approximation error and how the various coefficients might be efficiently calculated, while section 5 presents a numerical comparison of the approximate model compared with the original pricing model. Finally, some conclusions are presented in section 6.

2. The valuation model. The model of [4] is calculated on the assumption that an end-to-end link in a telecommunications network is well modeled by an M/M/C/C loss system. This is a continuous-time birth-and-death process with state space $\{0, 1, \ldots, C\}$ [9]. For $0 \leq n \leq C$, the state n denotes the number of connections present at any one time. We assume that the arrival rate λ and mean holding time μ^{-1} of connections are known. In practice, this may not be true in any a priori sense. However, methods for online evaluation of these parameters are currently a subject of great interest in the literature (see, for example, [3]) and it is reasonable to expect that estimates will be available. We denote the expected amount of revenue earned per connection by θ.

In [4], two of the authors derived the function $R_{n,C}(T)$ that gives the expected revenue lost in the interval $[0, T]$ due to arriving connections being rejected, when the capacity is C and the occupancy at time 0 is n. This model can be converted to a set of value functions for capacity via the relations

$$B_n(T) = R_{n,C}(T) - R_{n,C+1}(T), \tag{2.1}$$

$$S_n(T) = \begin{cases} R_{n,C-1}(T) - R_{n,C}(T) & \text{when} \quad n < C, \\ R_{C-1,C-1}(T) - R_{C,C}(T) & \text{when} \quad n = C, \end{cases} \tag{2.2}$$

respectively. The function $B_n(T)$ gives the amount that the link should "pay" for an extra unit of capacity. The logic behind (2.1) is that the value of an extra unit of capacity is given by the expected difference between the revenue that would be lost over the planning horizon $[0, T]$ if the extra capacity were present at time 0 and if it were not. Similarly, the function $S_n(T)$ gives the expected amount of extra revenue that would be lost over the planning horizon $[0, T]$ if the link were to give up a unit of capacity at time 0.

The formula (2.2) was not explicitly given in [4] for the case $n = C$. In this case, if the link were to give up a unit of capacity, it would also have to eject one of its current customers. The issue then arises as to whether an extra penalty value should be added to reflect the negative consequences of such a decision. The right-hand side of (2.2) reflects the situation in which no such penalty is added. The opposite extreme would be to incorporate an infinite penalty, which would have the effect of precluding any capacity reallocation away from a full link.

We might envisage a capacity reallocation scheme in which the links act in a cooperative manner. If $S_n(T)$ for one link is less than or equal to $B_n(T)$ for another link with whom it shares physical capacity, then a unit of capacity will be reallocated from the first link to the second. Via this mechanism, capacity is moved to that part of the network in which it can have the greatest effect in reducing loss of revenue.

To evaluate $R_{n,C}(T)$, it is necessary to compute the numerical inversion of the Laplace transform,

$$\widetilde{R}_{n,C}(s) = \left(\frac{1}{s}\right)\left(\frac{\theta\lambda}{(s + C\mu)P_C\left(s/\lambda\right) - C\mu P_{C-1}\left(s/\lambda\right)}\right) P_n\left(\frac{s}{\lambda}\right), \tag{2.3}$$

where

$$P_n(s/\lambda) = (-\mu/\lambda)^n C_n^{(\lambda/\mu)}\left(-s/\mu\right) \tag{2.4}$$

and

$$C_n^{(\lambda/\mu)}\left(-s/\mu\right) = \sum_{k=0}^{n} \binom{n}{k}\binom{-s/\mu}{k}\left(\frac{-\lambda}{\mu}\right)^{n-k} k! \tag{2.5}$$

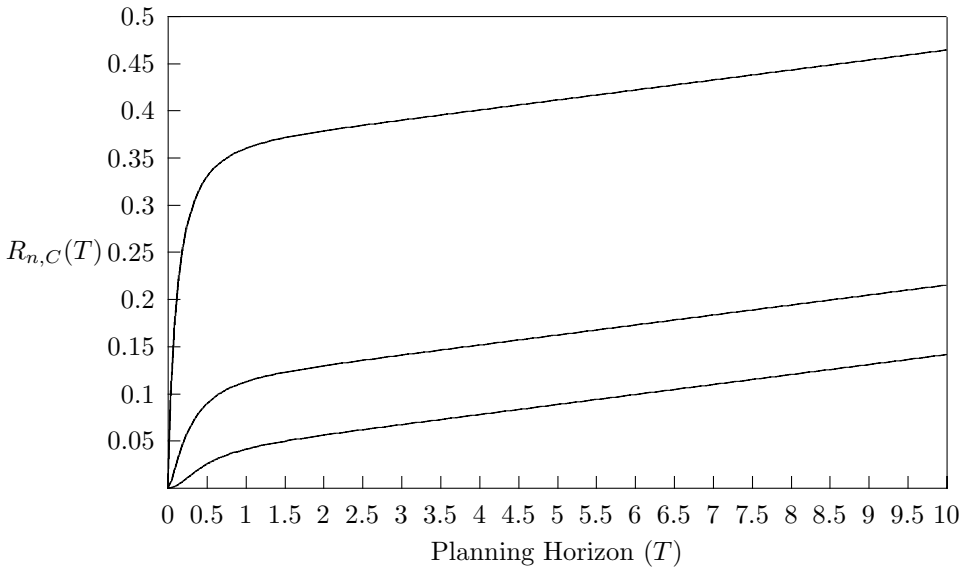


FIG. 1. $R_{n,C}(T)$ *with* $n = 4, 5, 6, C = 6, \lambda = 3$, *and* $\mu^{-1} = 0.5$.

is a Charlier polynomial as defined in [5]. For some purposes below it will be convenient to write $P_n(s/\lambda)$ in the form

$$P_n(s/\lambda) = \sum_{k=0}^{n} \binom{n}{k} \left(\frac{1}{\lambda}\right)^k s(s+\mu)\dots(s+(k-1)\mu), \tag{2.6}$$

which is easily derived by substituting (2.5) into (2.4). In [4] it was recommended that the Euler method (see Abate and Whitt [1]) be used for the numerical inversion of $\widetilde{R}_{n,C}(s)$ to yield $R_{n,C}(T)$. This involves numerical integration of the function $\widetilde{R}_{n,C}(s)$ along a contour which lies in the right complex half-plane. Thus, to implement the method, we need to be able to evaluate the right-hand side of (2.3) for complex numbers s with $\Re(s) > 0$.

An example of the lost revenue curves described by $R_{n,C}(T)$ is given in Figure 1. This example is adapted from one presented in [4]. Here the expected loss revenue curves are given for the case where $C = 6, n \in \{4, 5, 6\}, \lambda = 3$, and $\mu^{-1} = 0.5$.

In practice, it is difficult to evaluate both the expression on the right-hand side of (2.3) and its inverse Laplace transform. The calculation of the Charlier polynomials and $\widetilde{R}_{n,C}(s)$, if not done carefully, may be subject to arithmetic overflow, while the Euler method, although straightforward from an implementation viewpoint, is numerically complex. It is therefore of interest to develop stable and efficient methods for the calculation of $\widetilde{R}_{n,C}(s)$ and $R_{n,C}(T)$. In the following sections, we develop a recursion for $\widetilde{R}_{n,C}(s)$ and a linear approximation for $R_{n,C}(T)$. Furthermore, we give a stable recursion for computing the coefficients and bounds on the accuracy of the approximation. To conclude this section, we prove the following lemma, which will be useful in the rest of the paper.

LEMMA 2.1. *Let*

$$G_C(s) = (s + C\mu)P_C(s/\lambda) - C\mu P_{C-1}(s/\lambda) \tag{2.7}$$

so that the denominator of the right-hand side of (2.3) *is* $sG_C(s)$. *Then*

(i) $G_C(s)$ *has a zero at* $s = 0$, *which implies that* $G_C(s) = sF_C(s)$, *where* $F_C(s)$ *is a polynomial of degree* C;

(ii) *the zeros* $-\sigma_C < -\sigma_{C-1} \cdots < -\sigma_1$ *of* $F_C(s)$ *are all real and negative; and*

(iii) *the maximal zero* $-\sigma_1$ *of* $F_C(s)$ *is less than* $-\mu$.

Proof. Clearly $G_C(s)$ is a polynomial of degree $C + 1$. To get (i), observe that

$$G_C(s) = (s + C\mu)\left(\frac{-\mu}{\lambda}\right)^C C_C^{(\lambda/\mu)}(-s/\mu) - C\mu\left(\frac{-\mu}{\lambda}\right)^{C-1} C_{C-1}^{(\lambda/\mu)}(-s/\mu), \tag{2.8}$$

where $C_n^{(\lambda/\mu)}(-s/\mu)$ is defined in (2.5). Substituting $s = 0$ in (2.5), we see that $C_n^{(\lambda/\mu)}(0) = (\frac{-\lambda}{\mu})^n$ and thus the right-hand side of (2.8) is zero at $s = 0$. This gives (i).

Now let us think about (ii). It is known [5] that the zeros of $C_n^{(\lambda/\mu)}(x)$ are all real and positive. Moreover they interleave, that is, with $x_{n,i}$ the ith zero of $C_n^{(\lambda/\mu)}(x)$,

$$0 < x_{n,1} < x_{n-1,1} < \cdots x_{n,i} < x_{n-1,i} < x_{n,i+1} < \cdots < x_{n-1,n-1} < x_{n,n} < \infty. \tag{2.9}$$

By (2.4), we can see that the zeros of $P_n(s/\lambda)$ occur at the points $s_{n,i} = -x_{n,i}\mu$, which are all real and negative. From (2.9) it thus follows that

$$-\infty < s_{n,n} < s_{n-1,n-1} < \cdots < s_{n,i+1} < s_{n-1,i} < s_{n,i} < \cdots < s_{n-1,1} < s_{n,1} < 0. \tag{2.10}$$

For notational convenience, define $s_{n,0} = 0$ and $s_{n,n+1} = -\infty$. From (2.6), it is easy to see that the lead coefficient of $P_n(s/\lambda)$ is positive and so $P_n(s/\lambda)$ is negative in intervals of the form $(s_{n,2k}, s_{n,2k-1})$ for $k = 1, \lceil n/2 \rceil$ and positive in intervals of the form $(s_{n,2k+1}, s_{n,2k})$ for $k = 0, \lfloor n/2 \rfloor$.

It is clear from the representation (2.6) that $-C\mu$ is not a zero of $P_{C-1}(s/\lambda)$, so the point $-C\mu$ must lie in one of the intervals $(s_{C-1,i+1}, s_{C-1,i})$ for $i = 0, \ldots, C-1$. Define $j \in \{0, \ldots, C-1\}$ to be such that the interval that contains $-C\mu$ is $(s_{C-1,j+1}, s_{C-1,j})$. It follows from the interleaving property (2.10) that for $i \neq j$, $(s + C\mu)P_C(s/\lambda)$ changes sign exactly once in each of the intervals $(s_{C-1,i+1}, s_{C-1,i})$. Since $F_C(s) = (s + C\mu)P_C(s/\lambda)$ at the zeros of $P_{C-1}(s/\lambda)$, it too must change sign at least once in each of the intervals $(s_{C-1,i+1}, s_{C-1,i})$ for $i \neq j$. Thus there is at least one zero of $F_C(s)$ in each of the intervals $(s_{C-1,i+1}, s_{C-1,i})$ for $i \neq j$.

Now consider the behavior of $F_c(s)$ in the interval $(s_{C-1,j+1}, s_{C-1,j})$. The function $(s + C\mu)P_C(s/\lambda)$ changes sign twice in this interval. Assume that j is odd and not equal to $C - 1$. Then, by the interleaving property (2.10) and by the fact that $s_{C-1,j+1} < -C\mu < s_{C-1,j}$, it follows that $F_C(s_{C-1,j}) = (s_{C-1,j} + C\mu)P_C(s_{C-1,j}/\lambda)$ and $F_C(s_{C-1,j+1}) = (s_{C-1,j+1} + C\mu)P_C(s_{C-1,j+1}/\lambda)$ must both be negative. Moreover, $P_{C-1}(s/\lambda)$ is negative for $s \in (s_{C-1,j+1}, s_{C-1,j})$, which gives us that $F_C(s_{C,j+1})$ and $F_C(-C\mu)$ are both positive. There must therefore exist two zeros of $F_C(s)$ in the interval $(s_{C-1,j+1}, s_{C-1,j})$. A similar argument holds if j is even, equal to 0, or equal to $C - 1$.

We have thus shown that the degree-C polynomial $F_C(s)$ has C real and negative zeros, $C - 2$ of them in intervals of the form $(s_{C-1,i+1}, s_{C-1,i})$, for $i \neq j$, and two of them in the interval $(s_{C-1,j+1}, s_{C-1,j})$. Part (ii) of the lemma is proved.

To prove part (iii), we substitute the form (2.6) of $P_n(s/\lambda)$ into (2.7). We have

$$\begin{aligned} G_C(s) &= s\sum_{k=0}^{C}\binom{C}{k}\left(\frac{1}{\lambda}\right)^k s(s+\mu)\dots(s+(k-1)\mu) \\ &\quad + C\mu\sum_{k=0}^{C-1}\left[\binom{C}{k}-\binom{C-1}{k}\right]\left(\frac{1}{\lambda}\right)^k s(s+\mu)\dots(s+(k-1)\mu) \\ &\quad + C\mu\left(\frac{1}{\lambda}\right)^C s(s+\mu)\dots(s+(C-1)\mu) \end{aligned}$$

and so $F_C(s) = G_C(s)/s$ can be expressed as

$$\begin{aligned} F_C(s) &= \sum_{k=0}^{C}\binom{C}{k}\left(\frac{1}{\lambda}\right)^k s(s+\mu)\dots(s+(k-1)\mu) \\ &\quad + C\mu\sum_{k=1}^{C-1}\binom{C-1}{k-1}\left(\frac{1}{\lambda}\right)^k (s+\mu)\dots(s+(k-1)\mu) \\ &\quad + C\mu\left(\frac{1}{\lambda}\right)^C (s+\mu)\dots(s+(C-1)\mu) \\ &= 1+\sum_{k=1}^{C-1}\binom{C-1}{k-1}[C(s/k+\mu)]\left(\frac{1}{\lambda}\right)^k (s+\mu)\dots(s+(k-1)\mu) \\ &\quad + \left(\frac{1}{\lambda}\right)^C (s+\mu)\dots(s+C\mu), \end{aligned}$$

which is easily seen to be positive for $s > -\mu$. All the zeros of $F_c(s)$ are thus less than $-\mu$ and part (iii) is proved. □

3. A stable method for computing $\widetilde{\boldsymbol{R}}_{\boldsymbol{n,C}}(\boldsymbol{s})$. In (11) of [4], it was shown that the polynomials $P_n(s/\lambda)$ satisfy the recurrence relation

$$P_{n+1}(s/\lambda) = \left(\frac{s}{\lambda}+\frac{\mu n}{\lambda}+1\right)P_n(s/\lambda) - \frac{\mu n}{\lambda}P_{n-1}(s/\lambda) \tag{3.1}$$

for $n \geq 1$. With $H_n(s) \equiv P_{n-1}(s/\lambda)/P_n(s/\lambda)$, it follows that

$$1 = \left(\frac{s}{\lambda}+\frac{\mu n}{\lambda}+1\right)H_{n+1}(s) - \frac{\mu n}{\lambda}H_n(s)H_{n+1}(s)$$

and so

$$H_{n+1}(s) = \frac{1}{\left(\frac{s}{\lambda}+\frac{\mu n}{\lambda}+1\right) - \left(\frac{\mu n}{\lambda}\right)H_n(s)}. \tag{3.2}$$

With the initial condition

$$H_0(s) = \frac{1}{s/\lambda+1}, \tag{3.3}$$

we can use (3.2) to calculate $H_n(s)$ for s with $\Re(s) > 0$.

Moreover, Lemma 3.1 below shows that $H_n(s)$ remains bounded for s with $\Re(s) > 0$ and so the recursion is stable. Having calculated $H_n(s)$ for $n = 1, \ldots, C$, we can then calculate $\widetilde{R}_{C,C}(s)$ via the equation

$$\widetilde{R}_{C,C}(s) = \left(\frac{1}{s}\right)\left(\frac{\theta}{\frac{s}{\lambda} + \frac{\mu C}{\lambda} - \frac{\mu C}{\lambda} H_C(s)}\right), \tag{3.4}$$

which is easily derived from (2.3), and $\widetilde{R}_{n,C}(s)$ from the relation

$$\widetilde{R}_{n,C}(s) = \widetilde{R}_{C,C}(s) \prod_{j=n+1}^{C} H_j(s). \tag{3.5}$$

LEMMA 3.1. *For s with $\Re(s) > 0$ and $n \geq 1$, the ratios $H_n(s)$ are such that $|H_n(s)| < 1$.*

Proof. From (3.3), it follows that $|H_0(s)|$ is clearly less than one for s with $\Re(s) > 0$. Now assume that $|H_n(s)|$ is less than one for s with $\Re(s) > 0$. Then, from (3.2),

$$\begin{aligned} |H_{n+1}(s)| &= \left| \frac{1}{\left(\frac{s}{\lambda} + \frac{\mu n}{\lambda} + 1\right) - \left(\frac{\mu n}{\lambda}\right) H_n(s)} \right| \\ &\leq \frac{1}{\left|\frac{s}{\lambda} + \frac{\mu n}{\lambda} + 1\right| - \left|\frac{\mu n}{\lambda} H_n(s)\right|} \\ &\leq \frac{1}{\left|\frac{s}{\lambda} + 1\right|}, \end{aligned}$$

which is less than one for s with $\Re(s) > 0$. The lemma is thus proved by mathematical induction. □

4. An approximation to $R_{n,C}(T)$. As $T \to \infty$, the occupancy of the M/M/C/C loss system will be distributed according to the system's stationary distribution. Hence, as pointed out in [4] we would expect that the loss curves described by $R_{n,C}(T)$ will have asymptotic slope $\theta\lambda\pi_{\rho,C}$, where the Erlang-B function

$$\pi_{\rho,C} = \frac{\rho^C/C!}{\sum_{i=0}^{C} \rho^i/i!} \tag{4.1}$$

gives the stationary probability that the link is full when the traffic is $\rho \equiv \lambda/\mu$ and the capacity is C.

In this section, we shall verify that this is indeed the case and develop a straight-line approximation to $R_{n,C}(T)$ of the form

$$\hat{R}_{n,C}(T) = \theta\lambda\pi_{\rho,C}T + a_{n,C} \tag{4.2}$$

that can be used to compute approximate buying and selling prices following a process similar to that used in (2.1) and (2.2), that is,

$$\hat{B}_n(T) = \hat{R}_{n,C}(T) - \hat{R}_{n,C+1}(T), \tag{4.3}$$

$$\hat{S}_n(T) = \begin{cases} \hat{R}_{n,C-1}(T) - \hat{R}_{n,C}(T) & \text{when} \quad n < C, \\ \hat{R}_{C-1,C-1}(T) - \hat{R}_{C,C}(T) & \text{when} \quad n = C. \end{cases} \tag{4.4}$$

As with (2.2), it may be appropriate to add a penalty function to the right-hand side of (4.4) when $n = C$.

A key step in the development of the approximation is given in the following theorem.

THEOREM 4.1. *The function $R_{n,C}(T)$ defined in section 2 satisfies the following:*

$$\lim_{T\to\infty} [R_{n,C}(T) - \theta\lambda\pi_{\rho,C}T] \tag{4.5}$$

$$= \frac{2\theta\lambda g_1(n) - \theta\lambda\pi_{\rho,C}\left[2g_1(C) + C\mu\left[g_2(C) - g_2(C-1)\right]\right]}{2\left[1 + C\mu[g_1(C) - g_1(C-1)]\right]}, \tag{4.6}$$

where, for $n \in \{0, 1, \ldots, C\}$,

$$g_1(n) = \frac{1}{\mu}\sum_{k=1}^{n}\binom{n}{k}\left(\frac{\mu}{\lambda}\right)^k (k-1)! \tag{4.7}$$

and

$$g_2(n) = \frac{2}{\mu^2}\sum_{k=2}^{n}\binom{n}{k}\left(\frac{\mu}{\lambda}\right)^k (k-1)! \sum_{m=1}^{k-1}\frac{1}{m}. \tag{4.8}$$

Proof. It is easily established by differentiating (2.6) that

$$g_1(n) = \left.\frac{dP_n(s/\lambda)}{ds}\right|_{s=0} \tag{4.9}$$

and

$$g_2(n) = \left.\frac{d^2P_n(s/\lambda)}{ds^2}\right|_{s=0}. \tag{4.10}$$

We also recall the previously used fact that $P_n(0) = 1$.

By parts (i) and (ii) of Lemma 2.1, the rational function $s\widetilde{R}_{n,C}(s)$ has one pole at $s = 0$ with all the other poles real and negative. Hence, as long as the function

$$A_{n,C}(s) \equiv \left[s\widetilde{R}_{n,C}(s) - \frac{\theta\lambda\pi_{\rho,C}}{s}\right] \tag{4.11}$$

does not have a pole at $s = 0$, all its poles are in the left half-plane; in particular, they are real and negative. It then follows by the final value theorem (see, for example, [6, pp. 110–111]) that the limit (4.5) exists and is equal to $\lim_{s\to 0} A_{n,C}(s)$.

Expansion of $A_{n,C}(s)$ gives

$$A_{n,C}(s) = \frac{s\theta\lambda P_n(\frac{s}{\lambda}) - \theta\lambda\pi_{\rho,C}\left[(s + C\mu)P_C(\frac{s}{\lambda}) - C\mu P_{C-1}(\frac{s}{\lambda})\right]}{s\left[(s + C\mu)P_C(\frac{s}{\lambda}) - C\mu P_{C-1}(\frac{s}{\lambda})\right]}.$$

Both the numerator and the denominator of this are equal to zero at $s = 0$. The derivative of the denominator is also equal to zero at $s = 0$, which implies that for the limit as $s \to 0$ to exist, the derivative of the numerator must be equal to zero at $s = 0$. Using (4.9), this derivative is equal to

$$\theta\lambda - \theta\lambda\pi_{\rho,C}\left[1 + C\mu\left[g_1(C) - g_1(C-1)\right]\right]. \tag{4.12}$$

Consider the factor $1 + C\mu\,[g_1(C) - g_1(C-1)]$. By (4.7), this is equal to

$$\begin{aligned}
&1 + C\left[\sum_{k=1}^{C}\binom{C}{k}\left(\frac{\mu}{\lambda}\right)^k (k-1)! - \sum_{k=1}^{C-1}\binom{C-1}{k}\left(\frac{\mu}{\lambda}\right)^k (k-1)!\right] \\
&= 1 + C\sum_{k=1}^{C-1}\binom{C-1}{k-1}\left(\frac{\mu}{\lambda}\right)^k (k-1)! + C!\left(\frac{\mu}{\lambda}\right)^C \\
&= 1 + C!\sum_{k=0}^{C-1}\left(\frac{\mu}{\lambda}\right)^{C-k}/k! \\
&= 1 + \frac{\sum_{k=0}^{C-1}\left(\frac{\lambda}{\mu}\right)^k/k!}{\left(\frac{\lambda}{\mu}\right)^C/C!} \\
&= \frac{1}{\pi_{\rho,C}}
\end{aligned} \tag{4.13}$$

by (4.1). It follows that (4.12) is equal to zero as required. This argument shows further that if the coefficient of T in (4.2) were anything other than $\theta\lambda\pi_{\rho,C}$, then $\lim_{s\to 0} A_{n,C}(s)$ would not exist. This verifies that the approximation (4.2) has the correct asymptotic slope.

Now, using a further application of l'Hôpital's rule, together with (4.9) and (4.10), it is easily seen that $\lim_{s\to 0} A_{n,C}(s)$ is equal to the right-hand side of (4.6), which gives us the result. □

Theorem 4.1 tells us that we should take the constant $a_{n,C}$ in (4.2) equal to the right-hand side of (4.6). Specifically, our linear approximation to $R_{n,C}(T)$ is given by

$$\hat{R}_{n,C}(T) = \theta\lambda\pi_{\rho,C}T + \frac{2\theta\lambda g_1(n) - \theta\lambda\pi_{\rho,C}\,[2g_1(C) + C\mu\,[g_2(C) - g_2(C-1)]]}{2\,[1 + C\mu[g_1(C) - g_1(C-1)]]}. \tag{4.14}$$

For large problems, we need to be careful in computing the coefficients in this linear approximation. There is certainly the potential for numerical problems if we attempt to calculate them directly using (4.1), (4.9), and (4.10). Fortunately, it is possible to design stable and efficient recursive methods for their evaluation.

It is well known (see, for example, [2]) that $\pi_{\rho,C}$ is most efficiently calculated using the recursion

$$\pi_{\rho,C+1} = \frac{\rho\pi_{\rho,C}}{C + 1 + \rho\pi_{\rho,C}} \tag{4.15}$$

with $\pi_{\rho,C+1} = 1$. This gives us the linear term in the approximation (4.14) and also, along with (4.13), the denominator of the constant term. Moreover, we can use (4.13) to derive the fact that, for $n \geq 1$,

$$g_1(n) = \frac{1}{\mu}\sum_{j=1}^{n}\frac{1 - \pi_{\rho,j}}{j\pi_{\rho,j}}, \tag{4.16}$$

which can be computed at the same time as we generate the terms in the recursion (4.15). This gives us both $g_1(n)$ and $g_1(C)$ in the constant term. Finally, let $\psi_{\rho,n} = [n\mu\,[g_2(n) - g_2(n-1)]]$.

Differentiating (3.1) twice and putting $s = 0$, we get, for $n \geq 2$,

$$\psi_{\rho,n+1} = \frac{n+1}{\rho} \left[2g_1(n) + \psi_{\rho,n} \right]. \tag{4.17}$$

This recursion can be initialized by observing that $\psi_{\rho,2} = 2/\lambda^2$.

Generally a link will be dimensioned so that $\lambda/C\mu$ is close to one. Thus the coefficient of the multiplicative part of the recursion in (4.17) will be less than one for the values of interest. The recursion can thus be expected to remain stable for these values.

The above observations demonstrate that we can compute both coefficients in the linear approximation (4.2) in a stable and efficient manner. The final question of interest concerns the accuracy of the approximation. We can use the information in Lemma 2.1 about the roots of $G_c(s)$ to approach this question. The result is contained in the following lemma.

LEMMA 4.2. *The difference between the functions* $R_{n,C}(T)$ *and* $\hat{R}_{n,C}(T)$, *as defined in* (4.14), *satisfies*

(i)

$$\lim_{T\to\infty} \frac{R_{n,C}(T) - \hat{R}_{n,C}(T)}{e^{-\sigma_1 T}} = K_1, \tag{4.18}$$

where

$$K_1 = \lim_{s\to-\sigma_1} (s+\sigma_1) \left[s\tilde{R}_{n,C}(s) - \frac{\theta\lambda\pi_{\rho,C}}{s} \right]; \tag{4.19}$$

(ii)

$$\lim_{T\to\infty} \frac{R_{n,C}(T) - \hat{R}_{n,C}(T)}{e^{-\mu T}} = 0. \tag{4.20}$$

Proof. Using Lemma 2.1 and the proof of Theorem 4.1, we see that the partial fraction expansion of

$$D_{n,C}(s) \equiv \tilde{R}_{n,C}(s) - \frac{\theta\lambda\pi_{\rho,C}}{s^2} - \frac{a_{n,C}}{s}$$

is of the form

$$D_{n,C}(s) = \sum_{i=1}^{C} \frac{K_i}{s+\sigma_i}, \tag{4.21}$$

where

$$K_i = \lim_{s\to-\sigma_i} (s+\sigma_i) \left[s\tilde{R}_{n,C}(s) - \frac{\theta\lambda\pi_{\rho,C}}{s} \right]. \tag{4.22}$$

From this we can see that

$$R_{n,C}(T) - \hat{R}_{n,C}(T) = \sum_{i=1}^{C} K_i e^{-\sigma_i T} \tag{4.23}$$

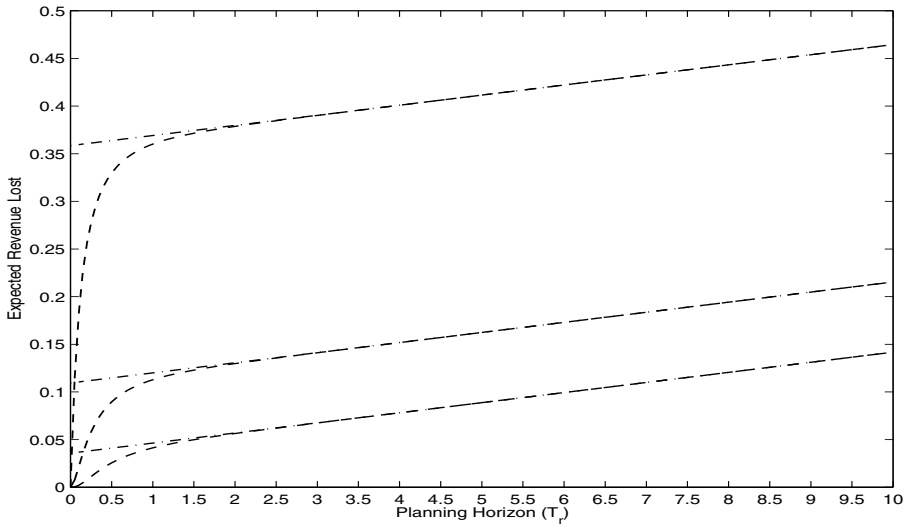


FIG. 2. *$R_{n,C}(T)$ (- -) versus $\hat{R}_{n,C}(T)$ (· -) with $n = 4, 5, 6, C = 6, \lambda = 3$, and $\mu^{-1} = 0.5$.*

and part (i) follows because $-\sigma_1$ is the largest root of $G_c(s)$. Part (iii) of Lemma 2.1 tells us that $-\sigma_1 < -\mu$ and so part (ii) of this lemma is an easy consequence of part (i). □

Lemma 4.2 tells us that the linear approximation $\hat{R}_{n,C}(T)$ approaches the actual function $R_{n,C}(T)$ at a rate that is better than exponential with coefficient $-\mu$. This gives us a useful indication as to the quality of the approximation.

5. A comparison of $R_{n,C}(T)$ and $\hat{R}_{n,C}(T)$. As an illustration of the behavior of our approximation function $\hat{R}_{n,C}(T)$, we revisit the example presented in section 2. Figure 2 depicts the original loss curves $R_{n,C}(T)$ and also $\hat{R}_{n,C}(T)$ for $C = 6, n \in [4, 6], \lambda = 3$, and $\mu^{-1} = 0.5$. Note that $R_{4,6}(T) < R_{5,6}(T) < R_{6,6}(T)$ and that similarly $\hat{R}_{4,6}(T) < \hat{R}_{5,6}(T) < \hat{R}_{6,6}(T)$.

From Figure 2 we immediately see that the curves $\hat{R}_{n,c}(T)$ are very good approximations for $R_{n,C}(T)$ for large T.

Next, we compare $\hat{B}_{n,C}(T)$ against $B_{n,C}(T)$ for a larger network where $C = 100, n = \{50, 100\}, \lambda = 85$, and $\mu^{-1} = 1$. Both functions are displayed in Figure 3. We observe that the original and approximated functions coincide once the system has reached equilibrium.

Of particular interest in this example is the linear function described by $\hat{B}_{50,100}(T)$. We see that in this instance the approximation results in a negative value for the intercept. Moreover, $B_{50,100}(T)$ is very close to zero for T such that $\hat{B}_{50,100}(T)$ is negative. In practice, this suggests that when $\hat{B}_{50,100}(T)$ is negative, we should treat the value of an extra unit of capacity as equal to zero. If we compare the negative values of $\hat{B}_{50,100}(T)$ with the corresponding values produced by $B_{50,100}(T)$, we see that this approximation is likely to be accurate.

Similar observations can be made with respect to Figure 4, where we compare $\hat{S}_{n,C}(T)$ against $S_{n,C}(T)$ for the larger network. In this instance we would once again interpret all negative values as zero in practice.

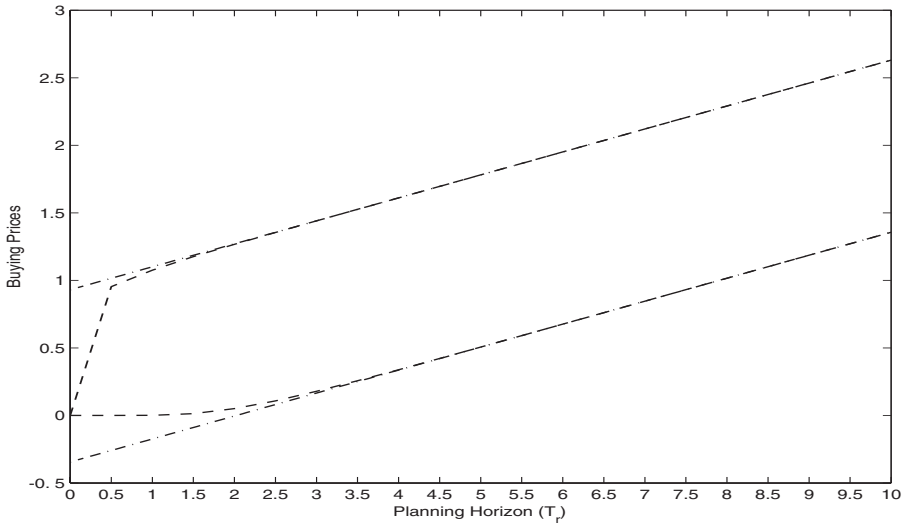


FIG. 3. $B_{n,C}(T)$ *(- -) versus* $\hat{B}_{n,C}(T)$ *(· -) with* $n = 50, 100, \lambda = 85$*, and* $\mu^{-1} = 1$.

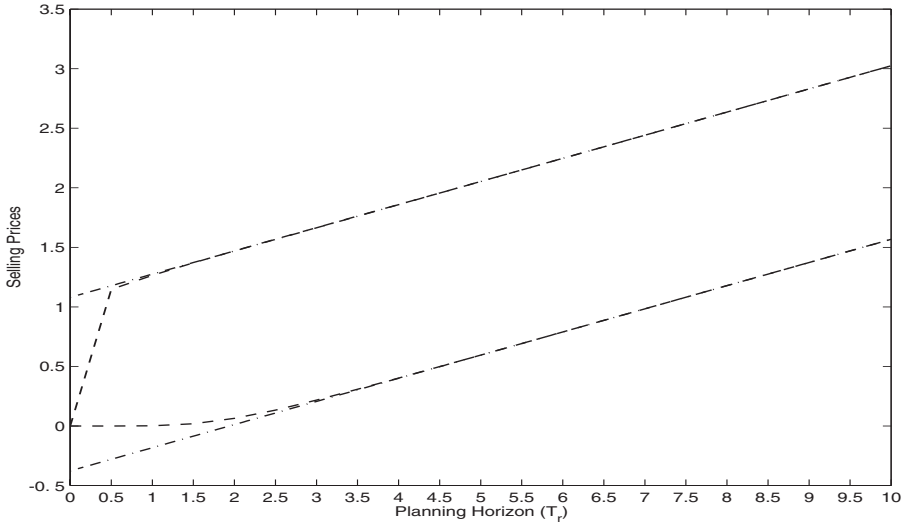


FIG. 4. $S_{n,C}(T)$ *(- -) versus* $\hat{S}_{n,C}(T)$ *(· -) with* $n = 50, 100, \lambda = 85$*, and* $\mu^{-1} = 1$.

6. Conclusions. Of considerable importance in telecommunications is the ability to transfer capacity between network flows according to some scheme that ascribes a value to capacity, in order to alleviate network congestion. The value function must meet three specific criteria: it is decentralized, scalable, and able to be implemented by a simple network switch.

In this paper, we have considered the valuation model presented in [4] which satisfied the first two criteria, and we developed a suitable approximate equivalent. The calculations involved are sufficiently simple to be implemented by a network switch.

Further implementation issues, such as the online estimation of the traffic param-

eters required for these pricing models, are still to be explored. This topic has already attracted considerable interest in the field and will be the subject of future research.

Acknowledgment. The authors would like to thank an anonymous referee for pointing out the recursion (4.17).

REFERENCES

[1] J. Abate and W. Whitt, *Numerical inversion of Laplace transforms of probability distributions*, ORSA J. Comput., 7 (1995), pp. 36–43.

[2] H. Akimaru and K. Kawashima, *Teletraffic: Theory and Application*, Springer-Verlag, London, 1999.

[3] A. Arvidsson, *High level B-ISDN/ATM traffic management in real time*, in ATM Networks, Performance Modelling and Analysis, Vol. 1, D. D. Kouvatsos, ed., IFIP Conference Proc. 17, Chapman & Hall, London, 1995, pp. 177–207.

[4] B. A. Chiera and P. G. Taylor, *What is a unit of capacity worth?*, Probab. Engrg. Inform. Sci., 16 (2002), pp. 513–522.

[5] T. Chihara, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[6] L. Debnath, *Integral Transforms and Their Applications*, CRC Press, Boca Raton, FL, 1995.

[7] E. Fulp and D. Reeves, *QOS rewards and risks: A multi-market approach to resource allocation*, in Proceedings of the IFIP-TC6 Networking 2000 Conference, Paris, France, 2000, pp. 945–956.

[8] M. A. Gibney and N. R. Jennings, *Dynamic Resource Allocation by Market-Based Routing in Telecommunications Networks*, in Proceedings of the Second International Workshop on Intelligent Agents for Telecommunication Applications (IATA'98), Lecture Notes in Artificial Intelligence 1437, Springer, Berlin, 1998, pp. 102–117.

[9] L. Kleinrock, *Queuing Systems. Volume* 1: *Theory*, John Wiley and Sons, New York, 1975.

[10] S. Lanning, W. Massey, B. Rider, and Q. Wang, *Optimal pricing in queueing systems with quality of service constraints*, In Proceedings of the 16th International Teletraffic Congress, Edinburgh, UK, 1999, pp. 747–756.

[11] M. Wellman, *A market-oriented programming environment and its application to distributed multicommodity flow problems*, J. Artificial Intelligence Res., 1 (1993), pp. 1–23.

SIAM J. APPL. MATH.
Vol. 65, No. 4, pp. 1420–1442

DIFFUSIVE AND CHEMOTACTIC CELLULAR MIGRATION: SMOOTH AND DISCONTINUOUS TRAVELING WAVE SOLUTIONS*

K. A. LANDMAN†, M. J. SIMPSON†, J. L. SLATER†, AND D. F. NEWGREEN‡

Abstract. A mathematical model describing cell migration by diffusion and chemotaxis is considered. The model is examined using phase plane, numerical, and perturbation techniques. For a proliferative cell population, traveling wave solutions are observed regardless of whether the migration is driven by diffusion, chemotaxis, or a combination of the two mechanisms. For pure chemotactic migration, both smooth and discontinuous solutions with shocks are shown to exist using phase plane analysis involving a curve of singularities, and identical results are obtained numerically. Alternatively, pure diffusive migration and combinations of diffusive and chemotactic migration yield smooth solutions only. For all cases the wave speed depends on the exponential decay rate of the initial cell density, and it is bounded by a minimum value which is numerically observed whenever the initial cell distribution has compact support. The minimum wave speed c_{min} is proportional to $\sqrt{\chi}$ or $\sqrt{D}$ for pure chemotaxis and pure diffusion cases, respectively. The value of c_{min} for combined diffusion and chemotactic migration is examined numerically. The rate at which the mixed migration system approaches either a diffusion-dominated or chemotaxis-dominated system is investigated as a function of a dimensionless parameter involving D/χ. Finally, a perturbation analysis provides details of the steep critical layer when $D/\chi \ll 1$, and these are confirmed with numerical solutions. This analysis provides a deeper qualitative and quantitative understanding of the interplay between diffusion and chemotaxis for invading cell populations.

Key words. migration, chemotaxis, diffusion, traveling wave, numerical solution, phase plane, shock, wave speed

AMS subject classifications. 34A34, 35L40, 35L67, 92C17, 35K57, 92C15, 65M99

DOI. 10.1137/040604066

1. Introduction.

Cell migration is an essential feature of many important biological systems, including wound healing, tumor invasion, and several developmental biology processes [12, 27, 37]. Typically, to model cell migration, a system of conservation equations is proposed which incorporates the migratory processes in conjunction with kinetic terms to simulate proliferation of the migratory population. Additional kinetic processes (e.g., cell death, cell-receptor binding) can be included in the kinetic terms where required. Diffusion and chemotaxis are two common cell migration mechanisms [5].

Diffusion simulates random walk processes of cells. The Fisher equation [6] is the archetypal pure diffusion model which considers diffusive migration together with proliferation of cells via a logistic process.

Chemotaxis describes the movement of cells in the direction of a spatial gradient of a signaling species called the chemoattractant. The chemoattractant kinetics may be specified in several ways. An early chemotactic model was developed by Keller and Segel [13] describing bacterial motion. Other important contributions have been

*Received by the editors February 10, 2004; accepted for publication (in revised form) October 12, 2004; published electronically April 26, 2005. This research was supported by National Health and Medical Research Council project grant ID237144.

http://www.siam.org/journals/siap/65-4/60406.html

†Department of Mathematics and Statistics, University of Melbourne, Victoria 3010, Australia (k.landman@ms.unimelb.edu.au, m.simpson@ms.unimelb.edu.au, j.slater@ms.unimelb.edu.au). The research of the first author was supported by the Particulate Fluids Processing Center, an Australian Research Council ARC Special Research Center.

‡Embryology Laboratory, Murdoch Childrens Research Institute, Royal Children's Hospital, Parkville, Victoria 3052, Australia (don.newgreen@mcri.edu.au).

made by Tranquillo [39], Tranquillo and Alt [40], Hillen [10], Othmer and Stevens [29], and Horstmann and Stevens [11], as well as those reviewed by Ford and Cummings [7].

The classical Fisher model and several pure chemotaxis models [1, 18, 24, 25, 31, 32] are known to support traveling wave solutions moving with a constant speed. For the Fisher model, the wave speed is bounded by a minimum value [25]. For pure chemotactic migration, Landman, Pettet, and Newgreen [18] recently demonstrated the existence of traveling wave solutions with a minimum wave speed. It should be noted that haptotaxis, which is based on migration along adhesive extracellular matrix gradients, is mathematically equivalent to chemotaxis; hence a pure haptotactic model can also support traveling wave solutions with a minimum wave speed [20, 22, 30]. The focus of these previous analyses has been to examine the characteristics of traveling wave solutions for cell migration in response to a single mechanism. The more complex case of multimechanism migration has received less attention and is therefore poorly understood.

In this article we consider a model of diffusive and chemotactic cell migration. The model is motivated by migration processes during embryological development. The rostral-to-caudal migration of neural crest cells along the developing avian and mammalian intestine is one of the most extensive migration paths known in developmental biology [15]. Neural crest cells show a variety of responses including chemotactic attraction to growth factors, which are thought to be produced uniformly along the intestine mesenchymal tissue (e.g., glial derived neurotrophic factor (GDNF) [43]). Local gradients in the chemoattractant concentration are postulated to arise from the binding of the chemoattractant to receptors on the migrating cells, rather than from diffusion of growth factors from a source. In addition to promoting migration, the chemoattractant also acts as a survival factor for the migrating population [9, 26, 43]. Interest in the migration of enteric neural crest cells stems from hypotheses which have linked neural crest cell migration to a common birth defect in humans called Hirschsprung's disease or aganglionic megacolon. This defect occurs when the caudal part of the gut lacks intrinsic nerve cells. Hirschsprung's disease is thought to occur when the rostral-to-caudal migration of the neural crest cells fails to completely colonize the developing intestine [17, 28].

This paper constructs a mathematical framework for the analysis of the combined diffusive and chemotactic migration, relevant to developmental biology processes. We utilize a holistic approach incorporating both analytical and numerical analyses of traveling wave solutions for the proposed model. For the case of purely chemotactic migration, the results presented here extend the previous work of Landman, Pettet, and Newgreen [18] in two significant ways. First, the relationship between the wave speed and the transition from smooth to discontinuous solutions is examined in detail. Second, an analysis of the functional dependence of the minimum wave speed for pure chemotaxis migration is presented. This analysis provides a useful relationship similar to the well-known expression for the Fisher equation.

For the more complex case of combined diffusion and chemotaxis migration we use a specifically designed numerical algorithm to examine the traveling wave solutions. In particular, the numerical results are used to show how the combined diffusive and chemotactic migration model approaches the limits of diffusion-only and chemotaxis-only cases as the relative contributions of diffusion and chemotaxis are altered. This kind of analysis is unexplored in previous studies [18, 21]. We use the numerically determined wave speeds to conjecture some useful bounds on the minimum wave speed

for the combined diffusion and chemotaxis problem.

The mathematical model for this problem is a coupled system of partial differential equations for cell density and chemoattractant concentration. A traveling wave coordinate system is introduced with an unknown wave speed to convert the system into a coupled system of ordinary differential equations. Phase plane and singular perturbation methods are then used to explore the solutions of the system. The numerical algorithm is applicable to pure hyperbolic problems including the formation of shock-fronted solutions as well as parabolic problems.

2. Diffusive and chemotactic cell migration in one dimension. A system of equations is introduced to describe the diffusive and chemotactic migration of cells in one dimension. Let $n(x,t)$ and $g(x,t)$ denote the cell density and chemoattractant concentration per unit length, respectively; x and t are position and time coordinates. A conservation-of-mass argument for a diffusion and chemotaxis transport of cells gives

$$\frac{\partial n}{\partial t} = D\frac{\partial^2 n}{\partial x^2} - \chi\frac{\partial}{\partial x}\left(n\frac{\partial g}{\partial x}\right) + f(n,g), \tag{2.1}$$

$$\frac{\partial g}{\partial t} = h(n,g), \tag{2.2}$$

where the diffusion coefficient D and the chemotactic factor χ are assumed to be constant [20, 21, 22]. The assumption of a constant chemotactic factor ignores saturation effects. Although alternative forms for $\chi(g)$ that incorporate saturation have been proposed [8], the specific relationship relevant to the system of interest is unknown and therefore a constant value is adopted. Preliminary investigations indicate that the results of this study are qualitatively insensitive to this assumption. The f and h terms in (2.1)–(2.2) represent the kinetic terms. Equation (2.2) reflects our assumption that the distribution of chemoattractant is governed by kinetic processes rather than diffusion. This is particularly relevant for the migration of neural crest cells where the chemoattractant GDNF is produced uniformly within the underlying tissues and not from diffusion from some external source [43]. For this case, the distribution of chemoattractant is governed by a balance between the underlying production of chemoattractant, the natural decay of chemoattractant, and also the uptake of chemoattractant by the migrating cells. Furthermore, care must be taken to ensure that the steady state of (2.2) does not permit a zero solution as the chemoattractant is a trophic factor necessary for the survival of the migratory population. Therefore, the chemoattractant concentration must be strictly positive at all times to sustain the migratory species.

In keeping with these biologically motivated considerations, the kinetic terms are chosen to reflect the following assumptions. The cells n proliferate by mitosis and have a carrying capacity density; these characteristics can be described by a logistic-type term for f. The chemoattractant g is produced uniformly at a constant rate throughout the domain and decays with time. Furthermore, the chemoattractant binds to the cells. Therefore a localized initial distribution of cells creates a gradient of chemoattractant, which produces a chemoattractant migration velocity. These effects are described with the following choice of f and h:

$$f = \lambda_1 n\left(1 - \frac{n}{k_1}\right), \tag{2.3}$$

$$h = \lambda_2 - \lambda_3 g - \lambda_4 n g. \tag{2.4}$$

For simplicity, a constant mitotic index λ_1 is assumed rather than a more complex form with $\lambda_1(g)$.

The system (2.1)–(2.4) reduces to some special cases. When the chemotactic factor χ is zero, the equation describing the cell population n reduces to the Fisher equation [6, 25]. Alternatively, when the diffusivity D is zero, cell migration is driven by chemotaxis alone. Several authors (e.g., [2, 18, 32]) have studied some aspects of simple chemotaxis models, or mathematically equivalent haptotaxis models, with a different choice of kinetic term h. The choice of h here is governed by considerations relevant to developmental biology problems as discussed.

We are interested in cells at their maximum density migrating into a region without such cells, giving rise to an invading profile with a constant shape and moving at a constant speed. The well-studied Fisher equation allows such traveling wave solutions, while purely chemotactic systems also support such solutions [18]. The nature of such solutions, whether they are smooth or discontinuous functions, and their minimum wave speed will be investigated here.

Scaling time with the mitotic index and introducing a length scale L, all the variables can be made dimensionless using the definitions as shown:

$$n = k_1 n^*, \qquad g = \frac{\lambda_2}{\lambda_3} g^*, \qquad t = \frac{1}{\lambda_1} t^*, \qquad x = L x^*, \tag{2.5}$$

$$D^* = \frac{D}{L^2 \lambda_1}, \qquad \chi^* = \frac{\chi \lambda_2}{L^2 \lambda_1 \lambda_3}, \qquad \beta = \frac{\lambda_3}{\lambda_1}, \qquad \gamma = \frac{\lambda_4 k_1}{\lambda_1}. \tag{2.6}$$

In later sections we choose L so that one of the dimensionless parameters D^* or χ^* is equal to unity. Omitting the asterisk notation, the dimensionless system is

$$\frac{\partial n}{\partial t} = D \frac{\partial^2 n}{\partial x^2} - \chi \frac{\partial}{\partial x}\left(n \frac{\partial g}{\partial x} \right) + n(1-n), \tag{2.7}$$

$$\frac{\partial g}{\partial t} = \beta(1-g) - \gamma n g. \tag{2.8}$$

To explore the nature of the dynamics of the system (2.7)–(2.8), we consider numerical solutions in conjunction with phase plane and perturbation analyses.

3. Numerical solution. Numerical solutions to the full system (2.7)–(2.8) are sought. We are interested in obtaining results under a wide range of conditions where diffusion or chemotaxis can either dominate or be absent. Therefore, the numerical scheme must be sufficiently robust to solve either a purely hyperbolic system ($D = 0$) or the simpler diffusion-reaction system ($\chi = 0$). An operator splitting technique is used to overcome this difficulty [16, 38, 36, 41]. Within each time increment, temporal integration of the system (2.7)–(2.8) is split into two steps. First the purely hyperbolic system (2.7)–(2.8) with $D = 0$, namely,

$$\frac{\partial n}{\partial t} = -\chi \frac{\partial}{\partial x}\left(n \frac{\partial g}{\partial x} \right) + n(1-n), \tag{3.1}$$

$$\frac{\partial g}{\partial t} = \beta(1-g) - \gamma n g, \tag{3.2}$$

is solved to yield intermediate solutions. Second, these intermediate solutions are used as initial conditions to solve the remaining parabolic system

$$\frac{\partial n}{\partial t} = D \frac{\partial^2 n}{\partial x^2}, \tag{3.3}$$

$$\frac{\partial g}{\partial t} = 0. \tag{3.4}$$

Obtaining numerical solutions of linear hyperbolic problems using standard numerical schemes has been referred to as an "embarrassingly difficult problem" [42]. Therefore, the greatest of care must be exercised in obtaining numerical solutions of the nonlinear hyperbolic system (3.1)–(3.2). To this end, we spatially discretize (3.1)–(3.2) with the semidiscrete scheme described by Kurganov and Tadmor [14]. The resulting system of ordinary differential equations is explicitly integrated with a fourth-order Runge–Kutta algorithm using a constant time step [33]. The solution to the parabolic system (3.3)–(3.4) is obtained using a linear finite element mesh composed of uniformly spaced elements. Temporal integration of the discretized finite element equations is achieved with a mass-lumped backward Euler scheme [34]. Both spatial and temporal discretizations are uniform. We choose to include the kinetic terms in the hyperbolic step of the splitting scheme since this step is solved using an explicit method convenient for solving nonlinear kinetic terms; alternatively, if the kinetic terms are included in the parabolic step, then the implicit Euler stepping would require further iterations to solve the resulting nonlinear system of equations. From this point of view the splitting regime (3.1)–(3.4) is computationally efficient.

It should be noted that the main limitation with this numerical scheme is imposed through the hyperbolic solution method [14] which requires sufficiently small time steps so that the Courant condition is satisfied,

$$Cr = \max \frac{\mid \lambda_i \mid \Delta t}{\Delta x} \leq M, \tag{3.5}$$

where the λ_i are the eigenvalues associated with the Jacobian of the flux vector [14] and M is some constant. The λ_i relate to the speed of propagation of information for the system. Fortunately, since we are interested in traveling wave solutions which move with a constant wave speed, it is clear that a uniformly optimal time step Δt exists for a particular uniform spatial mesh. The optimal time step can be determined using a straightforward trial-and-error approach. The finite element solution of the parabolic system (3.3)–(3.4) is not subject to any numerical stability limitation since the mass-lumped implicit Euler scheme is known to be unconditionally stable [34].

The numerical scheme outlined here is particularly convenient for analyzing general solutions of the system (2.7)–(2.8). The inclusion of Kurganov and Tadmor's central scheme is necessary so that the nonlinear hyperbolic term associated with chemotactic migration can be solved accurately without incurring any high Peclet number-induced oscillations and numerical diffusion associated with standard numerical techniques [44]. Furthermore, incorporating diffusion through an operator splitting scheme is required to maintain generality of the algorithm. Previous attempts at simulating a combined haptotactic and diffusive migration system discretized the diffusion term explicitly within the central scheme [21]. This previous approach is very restrictive as explicit solutions of the diffusion equation are subject to well-known stability criteria [4] which are satisfied only for small values of the diffusion coefficient. These limitations are completely overcome in this work as the diffusion term is split and solved implicitly thereby yielding an algorithm valid for any value of χ and D.

The problem is modeled on the infinite x domain. However, for numerical computations the finite domain $[0, X]$ is selected with X chosen to be sufficiently large to avoid boundary effects. Zero-flux conditions are specified for both boundaries. Since we are interested in the invasion of cells into the domain, the initial data are chosen to be primarily localized near the left boundary as discussed in section 4.

After a particular time, the numerical solution converges to a fixed profile moving with a constant speed. The traveling wave speed c is computed by selecting a particular contour, say, $n(x,t) = N$, and locating the position of that contour at each time interval using a linear interpolation scheme. Once the position of the contour is known over successive time intervals, the wave speed can be approximated by

$$c_n = \frac{x^{n+1} - x^n}{\Delta t} \tag{3.6}$$

for large n, where x^n and x^{n+1} are the positions of the contour at the n and $n+1$ time step, respectively, and Δt is the time step. The speed of convergence varies with initial conditions and parameter values. Consequently, the domain length X must be chosen sufficiently large if the convergence is slow. When $D = 0$, the chemotactic migration cell profile is expected to develop a shock in the low-concentration region of the profile for some choices of initial condition [22]. It is impractical to use linear interpolation to determine the position of the contour within the shock because of the discontinuity. This complication is circumvented by choosing a concentration away from the shock region to compute the wave speed. Therefore, it is best to use a sufficiently large contour value N.

4. Traveling wave speed and dependence on the initial conditions. Traveling wave solutions with a range of possible wave speeds greater than some minimum value are known to occur for purely diffusive or purely chemotactic migration [6, 25, 18]. We expect the same behavior when both diffusion and chemotaxis are present. Here we investigate how various types of initial data evolve to traveling wave solutions with different wave speeds. To determine the minimum wave speed numerically, it is necessary to know the relationship between the initial conditions and the wave speed so that the appropriate initial conditions are specified.

It is possible to investigate the speed of the traveling waves by examining the leading edge of the wave, assuming it decays exponentially in space [25]. McKean [23] and Marchant [19] determine relationships between exponential decay rates of initial data and the wave speed of solutions for the Fisher equation (purely parabolic) and a haptotactic invasion (purely hyperbolic) system, respectively. We extend this work to our system with both diffusion and chemotaxis.

Consider initial conditions, where for large x

$$n(x,0) = A_1 e^{-\xi_1 x}, \tag{4.1}$$

$$g(x,0) = 1 - A_2 e^{-\xi_2 x} \tag{4.2}$$

for arbitrary positive constant A_1, A_2, ξ_1 and ξ_2. Looking at the evolving wave near the leading edge and writing $n = \tilde{n}, g = 1 - \tilde{g}$, assuming that $\tilde{n}$ and $\tilde{g}$ are small, the system (2.7)–(2.8) simplifies to the linear system

$$\frac{\partial \tilde{n}}{\partial t} = D\frac{\partial^2 \tilde{n}}{\partial x^2} + \tilde{n}, \tag{4.3}$$

$$-\frac{\partial \tilde{g}}{\partial t} = \beta \tilde{g} - \gamma \tilde{n} \tag{4.4}$$

with the initial conditions (for large x)

$$\tilde{n}(x,0) = A_1 e^{-\xi_1 x}, \tag{4.5}$$

$$\tilde{g}(x,0) = A_2 e^{-\xi_2 x}. \tag{4.6}$$

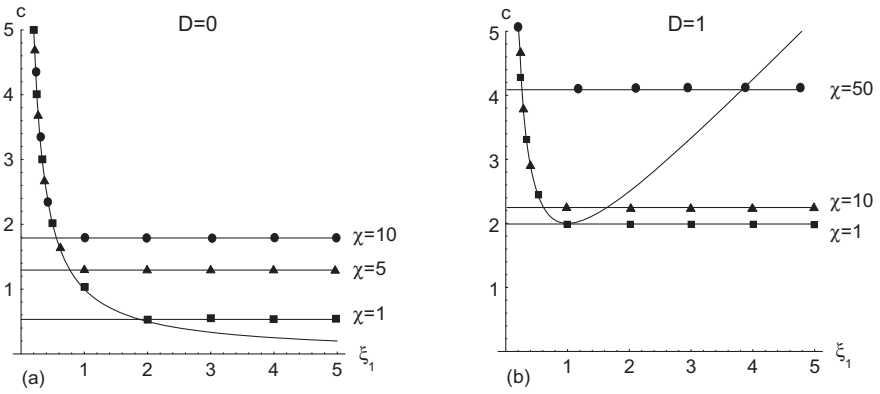


FIG. 4.1. *Numerical wave speed c versus ξ_1 for initial data of the form* (4.1)–(4.2). *Numerical results are shown in squares, circles, and triangles. These results were generated using* $\Delta x = 0.05$ *and* $\Delta t = 0.01$. *The continuous curves are given by* (4.7). *The horizontal lines represent* c_{min}. *Here* $\beta = 1$ *and* $\gamma = 1$. (*a*) *With* $D = 0$ *and increasing values of* χ. (*b*) *With* $D = 1$ *and increasing values of* χ.

A solution of the form $\tilde{n} = A_1 e^{-\xi_1(x-ct)}$ is sought. Substitution into (4.3)–(4.4) requires

$$c = \frac{1}{\xi_1} + D\xi_1 \tag{4.7}$$

for large values of x. Then solving (4.4) with (4.6), the leading edge of chemoattractant concentration is

$$\tilde{g} = e^{-\beta t}\left[A_2 e^{-\xi_2 x} - \frac{\gamma A_1}{\beta + \xi_1 c} e^{-\xi_1 x}\right] + \frac{\gamma A_1}{\beta + \xi_1 c} e^{-\xi_1(x-ct)}. \tag{4.8}$$

For large values of t, both $\tilde{n}$ and $\tilde{g}$ are functions of the traveling wave coordinate $x - ct$, where c is given by (4.7). This condition is independent of ξ_2 and hence independent of the initial conditions imposed on g.

The analytical result for c given by (4.7) is confirmed by the numerical results illustrated in Figure 4.1. Two cases are described. In the first, $D = 0$, so that the cells are purely chemotactically driven. For fixed values of the kinetic parameters β and γ and chemotactic factor χ, Figure 4.1(a) shows that a traveling wave solution with wave speed satisfying (4.7) is realized when $\xi_1 < 1/c_{min}$. Alternatively, when $\xi_1 > 1/c_{min}$ a traveling wave of fixed wave speed develops where $c = c_{min}$. Furthermore Figure 4.1(a) shows that c_{min} increases proportional to $\sqrt{\chi}$. In section 6.2, both smooth and discontinuous solutions will be found numerically using initial data of the form (4.1)–(4.2) (with $\xi_2 = 0$). (Note that (4.7) is also valid when both $D = \chi = 0$; under these conditions a traveling wave results from the initial nonzero cell density distribution in conjunction with the kinetics.) The second case, $D = 1$, $\chi > 0$, illustrated in Figure 4.1(b), has the same qualitative behavior as the case $D = 0$. The solution with $\chi = 0$ corresponds to the Fisher equation ($c_{min} = 2$, [25]) and is not shown here. As χ increases the value of c_{min} again increases, but for the case of nonzero D, it clearly does not scale with $\sqrt{\chi}$.

In summary, numerical computations yield a suite of traveling waves with the wave speed dependent on the exponential decay rate of the initial cell population

$n(x,0)$. There is a maximum exponential decay rate such that for ξ_1 larger than the maximum value, the initial data develops into a traveling wave moving with a minimum wave speed c_{min}. Further discussion of c_{min} will appear in section 6.3.2.

The asymptotic form of the initial conditions given by (4.1)–(4.2) is useful for numerically investigating the dependence of the wave speed on the decay rate ξ_1. However, in the limit $\xi_1 \to \infty$ the initial cell distribution tends towards having semi-compact support, a typical choice being

$$n(x,0) = \begin{cases} 1, & x < x_1, \\ q(x), & x_1 < x < x_2, \\ 0, & x > x_2, \end{cases} \tag{4.9}$$

where $q(x)$ is monotonic and continuous. Since all such functions decay faster than any exponential function, $n(x,t)$ will evolve to a traveling wave with speed $c = c_{min}$. Numerical solutions with such initial data confirm this result.

In light of this discussion, initial conditions used in this study take the form

$$n(x,0) = \begin{cases} 1, & x < 10, \\ e^{-\xi_1(x-10)}, & x \geq 10, \end{cases} \tag{4.10}$$

$$g(x,0) \equiv 1. \tag{4.11}$$

Altering the value of ξ_1 in (4.10) enables the leading front of the cell density distribution to decay exponentially with a variable rate. In the limit $\xi_1 \to \infty$ the initial conditions (4.10) approach a step function at $x = 10$. This is a particular case of the more general initial condition (4.9) with $x_1 = x_2 = 10$. The location of the transition point to exponential decay is arbitrary as identical traveling wave behavior results regardless of the point chosen.

5. Traveling wave solution. Introducing the traveling wave coordinate transformation $z = x - ct$, where c is the dimensionless wave speed, and the variable $v = \frac{\partial n}{\partial x}$, the dimensionless system (2.7)–(2.8) becomes the following first-order system of equations:

$$c\frac{dg}{dz} = -\left[\beta(1-g) - \gamma n g\right], \tag{5.1}$$

$$\frac{dn}{dz} = v, \tag{5.2}$$

$$\begin{aligned} D\frac{dv}{dz} = {} & \frac{\chi n}{c^2}\left[\gamma n + \beta\right]\left[\gamma n g - \beta(1-g)\right] - n(1-n) \\ & - \left[1 + \frac{\chi}{c^2}(\beta(1-g) - 2\gamma n g)\right] cv. \end{aligned} \tag{5.3}$$

There are two steady states of this system, namely, $(g,n,v) = (\frac{\beta}{\beta+\gamma}, 1, 0)$ and $(1,0,0)$. The first state corresponds to cells at their carrying capacity density and therefore can be thought of as the colonized or invaded state, whereas the second is the uncolonized state. We seek traveling wave solutions connecting the colonized to the uncolonized state. Note that $\frac{\beta}{\beta+\gamma}$ is a function which depends only on the ratio γ/β; it is an increasing function of the production rate β, is a decreasing function of binding rate γ, and is always less than unity, the value of the chemoattractant concentration in the absence of cells.

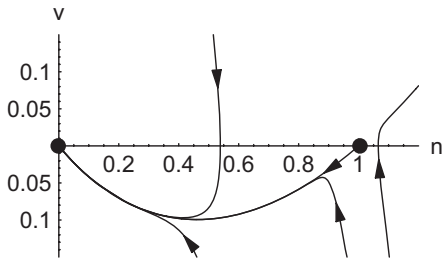


FIG. 6.1. *Phase plane for the Fisher equation. Here* $D = 1$, $c = 2.5$. *The steady states are marked* (•).

6. Phase plane, perturbation analysis, and numerical solutions. We first investigate phase plane and numerical solutions corresponding to the two special cases when one of D or χ is zero and discuss the nature of the solutions and the minimum wave speed c_{min}. For the remaining case, when both migration mechanisms are active, the transition from Fisher type solutions to chemotactic solutions (and vice versa) is investigated, and the solutions and c_{min} are determined numerically. In addition, perturbation analysis provides some insight into any rapid transition zones.

6.1. Diffusion-driven migration, no chemotaxis. If the chemotactic coefficient is zero, then the model equations reduce to the Fisher equation which describes cell migration driven by diffusion and proliferation. Then the system (2.7)–(2.8), in the traveling wave coordinate, reduces to the differential equation system

$$\frac{dn}{dz} = v, \tag{6.1}$$

$$D\frac{dv}{dz} = -n(1-n) - cv. \tag{6.2}$$

It is well known that traveling waves exist and can be found by phase plane analysis in the (n, v) plane, as illustrated in Figure 6.1. The state $(n, v) = (1, 0)$ is a saddle for all values of c. The other steady state (0,0) is a stable node if $c^2 > 4D$ and a stable spiral if $c^2 < 4D$. The population density n is required to be nonnegative and hence cannot be oscillatory around zero; therefore, the wave speed must be restricted to $c^2 \geq 4D$ giving a minimum wave speed $c_{min} = 2\sqrt{D}$. The traveling wave solutions are smooth. Clearly, the cell density is independent of the chemoattractant kinetics.

Numerical solutions to (2.7)–(2.8) with $\chi = 0$ are shown in Figure 6.2 with both rapid and slowly decaying initial conditions. In both cases, the profiles of $n(x, t)$ show clear traveling wave behavior characterized by a constant wave speed. For rapidly decaying initial conditions, Figure 6.2(a) demonstrates a minimum wave speed of $c_{min} = 2.0$, which agrees with the theoretical result. Alternatively for slowly decaying initial conditions, Figure 6.2(b) illustrates an increased wave speed of $c = 2.5$, as given by (4.7). This example confirms the result that the traveling wave speed c depends on the exponential decay rate of the initial distribution of the cell population.

6.2. Chemotactically driven migration, no diffusion. If the diffusion coefficient is zero, then the variable v does not need to be introduced. As discussed in section 2, when there is no diffusion, we can choose the length scale L so that the dimensionless χ is identically equal to unity. Then equations (5.1)–(5.3) reduce to the

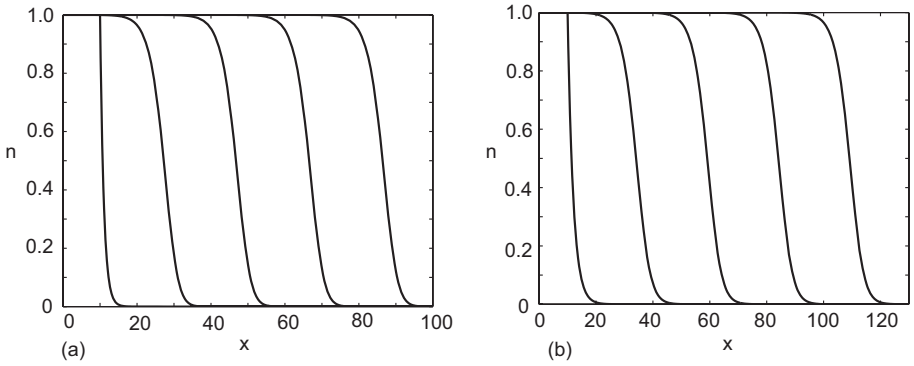


FIG. 6.2. *Numerical solutions of $n(x,t)$ for the Fisher equation with $D = 1$, $\Delta x = 0.05$, and $\Delta t = 0.01$. (a) Solutions at $t = 0, 10, 20, 30$, and 40 left to right with $\xi_1 = 10$. The computed wave speed is $c = c_{min} = 2$. (b) Solutions at $t = 0, 10, 20, 30$, and 40 left to right with $\xi_1 = 0.5$. The computed wave speed is $c = 2.5$.*

following system:

$$c\frac{dg}{dz} = -\left[\beta(1-g) - \gamma n g\right], \tag{6.3}$$

$$c\left[1 + \frac{1}{c^2}(\beta(1-g) - 2\gamma n g)\right]\frac{dn}{dz} = \frac{n}{c^2}\left[\gamma n + \beta\right]\left[\gamma n g - \beta(1-g)\right] - n(1-n). \tag{6.4}$$

The chemoattractant kinetic term h chosen here differs from that in [18], resulting in a different system with different steady states. The steady states of (6.3)–(6.4) are $(g,n) = (\frac{\beta}{\beta+\gamma}, 1)$ and (1,0). The point $(\frac{\beta}{\beta+\gamma}, 1)$ is an unstable focus when $c^2 > \frac{\beta\gamma}{\beta+\gamma}$ and is a saddle when $c^2 < \frac{\beta\gamma}{\beta+\gamma}$, while the point (1,0) is always a saddle. It is worth noting that with no diffusion, the stability of the steady states does not provide a minimum for the wave speed, since the eigenvalues are always real.

When the function premultiplying $\frac{dn}{dz}$ in (6.4) is identically zero, the derivative $\frac{dn}{dz}$ is no longer defined. Pettet, McElwain, and Norbury [32] defined such a curve as a *wall-of-singularities*. Here the wall-of-singularities can be written as

$$n = \frac{1}{2\gamma g}\left(c^2 + \beta(1-g)\right). \tag{6.5}$$

This wall is asymptotic to the n-axis, cutting the positive g-axis at

$$g = 1 + \frac{c^2}{\beta}$$

to the right of the steady state $(1,0)$. Hence when $c^2 > \frac{\beta\gamma}{\beta+\gamma}$ the two steady states (an unstable focus and a saddle) are to the left of the wall. Alternatively, when $c^2 < \frac{\beta\gamma}{\beta+\gamma}$, then the two steady states (both saddles) are on either side of the wall. The wall gets closer to the origin as c^2 decreases, and therefore it is possible for the wall to move below the steady state $(\frac{\beta}{\beta+\gamma}, 1)$.

Pettet, McElwain, and Norbury [32] showed that a solution approaching a wall-of-singularities could not cross the wall unless it passed through a special point called a *hole* in the wall. A hole is defined by both the function premultiplying $\frac{dn}{dz}$ and the

right-hand side of (6.4) being equal to zero simultaneously. Marchant, Norbury, and Perumpanani [20] and Landman, Pettet, and Newgreen [18] showed that for a system of equations (in the class of (2.1)–(2.2)) a trajectory exiting one steady state in the phase plane which passed through a hole in the wall could in fact recross the wall by way of a jump discontinuity to join up with the second steady state.

Similar behavior, where the two steady states are on same side of the wall, occurs for the system considered here. As noted, our system also allows the two steady states to be on the opposite sides of the wall. We will show that for this case the presence of a hole in the wall is irrelevant and all traveling wave solutions exhibit a shock or discontinuity. For this problem, there is at most one hole in the wall in the positive (g, n) quadrant.

In seeking a trajectory connecting $(\frac{\beta}{\beta+\gamma}, 1)$ to $(1, 0)$, two different types of behavior can occur, and these are explained with two examples.

Example 1. In our first example, Figure 6.3 illustrates the (g, n) phase plane with decreasing values of wave speed c, for one choice of the kinetic parameters β and γ. For sufficiently large wave speeds, the two steady states are below the wall as in Figure 6.3(a) and there is a unique trajectory to the left of the wall, connecting the two states; this gives a smooth traveling wave. However, as c is decreased, there is a value $c = c_{crit}$ where the wall begins to interfere with trajectories emanating from the unstable node. At this value the trajectory just touches the hole in the wall as in Figure 6.3(b). For $c < c_{crit}$, we must determine whether a trajectory emanating from $(\frac{\beta}{\beta+\gamma}, 1)$ can cross the wall and connect to the other steady state (1,0).

Marchant [19] and Landman, Pettet, and Newgreen [18] investigated a similar scenario. The arguments in section 4 of [18] for general kinetic terms apply to our system of equations, allowing us to summarize the results here. No smooth connection between the two states can be made; however, there is the possibility for the solution to be nonsmooth by containing a jump discontinuity. The method relies on hyperbolic partial differential equation theory, Lax entropy condition, and the Rankine–Hugoniot jump condition. A solution for n with a shock or discontinuity, traveling of course with the constant wave speed c, is shown to exist. Let the subscripts L and R denote the value of the variable on the left and right side of the shock, respectively. Then from (4.10)–(4.12) in [18], with $h = \beta(1 - g) - \gamma n g$, the shock conditions are

$$g_L = g_R = g, \tag{6.6}$$

$$n_L + n_R = \frac{1}{\gamma g}\left(c^2 + \beta(1 - g)\right), \tag{6.7}$$

$$u_L - u_R = \frac{\gamma g}{c}(n_L - n_R), \tag{6.8}$$

where $u = \frac{\partial g}{\partial x}$. These equations establish that g is continuous, while n and the spatial gradient of the chemoattractant concentration u support a discontinuity. The Lax entropy condition [3] is satisfied only if $n_L > n_R$. Recall that the wall-of-singularities satisfies (6.5). Hence, from (6.7) the geometric center of the jump $\frac{1}{2}(n_L + n_R)$ lies exactly on the wall-of-singularities, and therefore any jump takes the trajectory to the other side of the wall. In this way, it is possible for a trajectory to pass through the hole in the wall and then jump to a trajectory on the other side of the wall, thus connecting the colonized and uncolonized states when $c < c_{crit}$, although the wall prevents a smooth joining trajectory. Such a case is shown in Figure 6.3(c), where the discontinuity corresponds to the vertical portion of the trajectory that joins the colonized and uncolonized steady states. After the jump discontinuity, n will have a

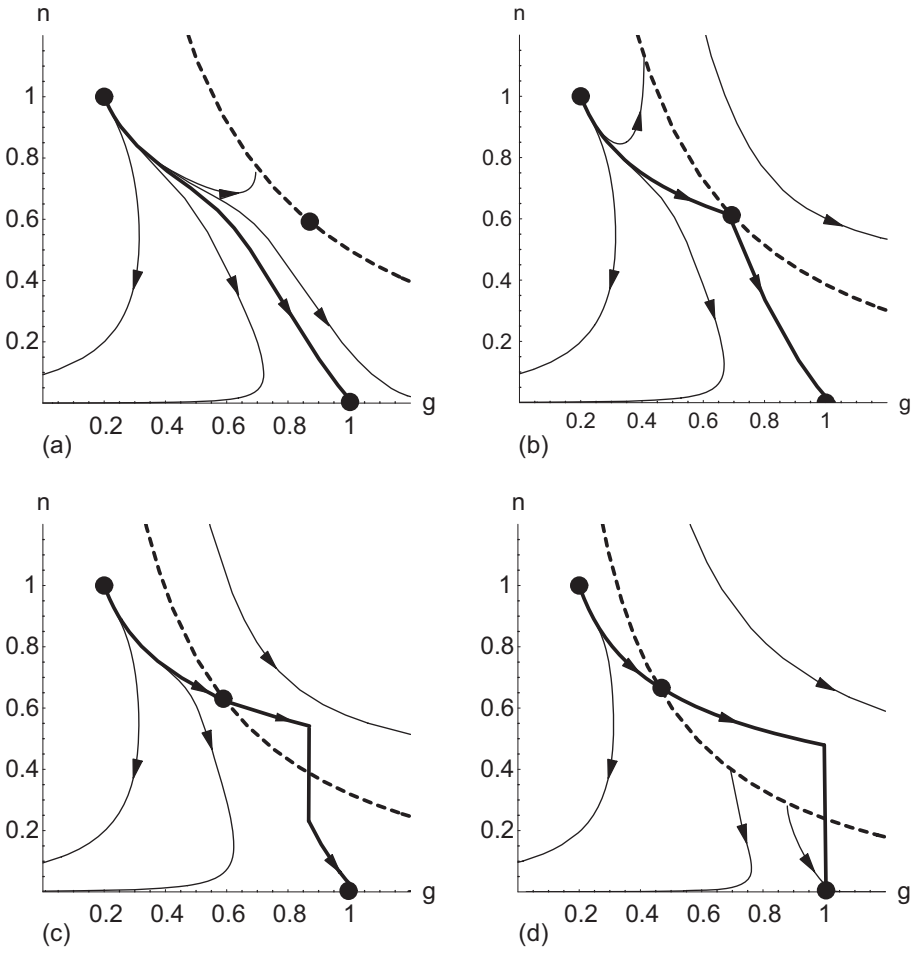


FIG. 6.3. *Phase plane for* (g, n) *for decreasing values of wave speed* c. *Here* $\beta = 0.25$, $\gamma = 1.0$. *The positions of the steady states* (•), *wall-of-singularities* (*dotted line*), *holes in the wall* (•), *and the trajectory joining the colonized and uncolonized steady states* (*thick line*) *are shown. The vertical lines in* (c) *and* (d) *correspond to the jump discontinuity in* n. (a) $c = 1.0$, (b) $c = c_{crit} \approx 0.88$, (c) $c = 0.8$, (d) $c = c_{min} \approx 0.69$.

smooth leading edge which asymptotes to zero.

However, for a realistic solution, $n_R > 0$, so the jump cannot be so large as to take the trajectory across the g-axis. As c decreases, the jump size becomes larger, until at some $c = c_{min}$, the trajectory jumps directly from $(g, n) = (1, c^2/\gamma)$ to $(1, 0)$, as illustrated in Figure 6.3(d). This solution with $c = c_{min}$ is the only solution with a zero leading edge and hence has compact support. If $c < c_{min}$, no smooth or nonsmooth traveling shock wave solution exists.

Therefore, our system supports traveling shock wave solutions with wave speed $c_{crit} > c > c_{min}$. Clearly for this example $c_{min}^2 > \frac{\beta\gamma}{\beta+\gamma}$, since both steady states remain on the same side of the wall. Example 2 considers the alternative case.

Example 2. In our second example, the value of the production rate β is increased sufficiently, so that $c_{min}^2 < \frac{\beta\gamma}{\beta+\gamma}$, allowing the possibility for the two steady states to lie on opposite sides of the wall, as shown in Figure 6.4. For sufficiently large wave

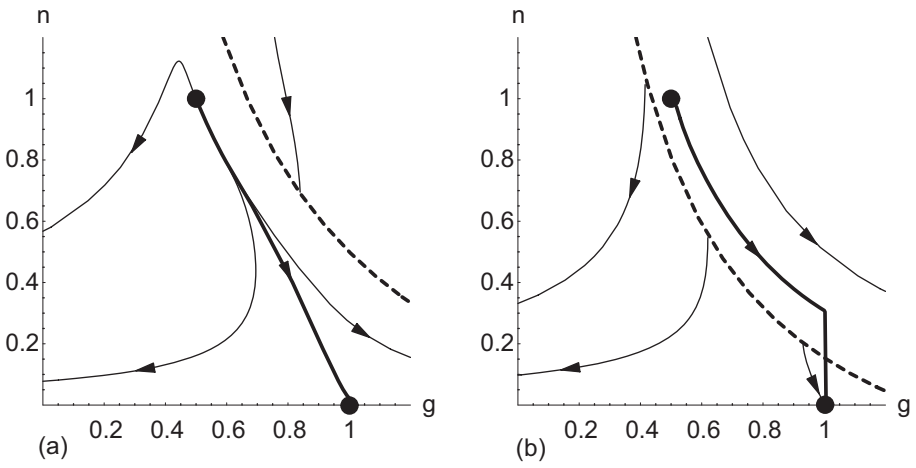


FIG. 6.4. *Phase plane for* (g, n) *for decreasing values of wave speed* c. *Here* $\beta = 1.0$, $\gamma = 1.0$. *The positions of the steady states* (•), *wall-of-singularities* (*dotted line*), *and the trajectory joining the colonized and uncolonized steady states* (*thick line*) *are shown. The vertical line in* (b) *corresponds to the jump discontinuity in* n. (a) $c = 1.0$, (b) $c = c_{min} \approx 0.556$.

speeds, the two steady states are below the wall, as in Figure 6.4(a), and there is a unique trajectory to the left of the wall, connecting the two states; this gives a smooth traveling wave. However, as c is decreased to c_{crit} where

$$c_{crit}^2 = \frac{\beta\gamma}{\beta + \gamma}, \tag{6.9}$$

the steady state lies on the wall and so is also a hole in the wall. For $c < c_{crit}$, the steady state lies on the other side of the wall, as shown in Figure 6.4(b). The jump discontinuity theory can then be applied again, so that the trajectory emanating from $(\frac{\beta}{\beta+\gamma}, 1)$ can cross the wall to join with a trajectory which connects with the saddle at (1,0). Again, the requirement that $n_R > 0$ implies that as c decreases, the jump size becomes larger, until at some $c = c_{min}$, the trajectory jumps directly from $(g, n) = (1, c^2/\gamma)$ to $(1, 0)$. If $c < c_{min}$, no smooth or nonsmooth traveling shock wave solution exists. Note that, for this case, no hole is needed when the two steady states are on opposite sides of the wall, as shown here.

In addition to the phase plane analysis, a numerical solution to the system (2.7)–(2.8) with $D = 0$ illustrates the smooth and discontinuous solutions and their corresponding dependence on the wave speed. As discussed in section 4 the wave speed depends on the exponential decay rate of the initial data for n, and therefore the mechanism for generating the smooth and discontinuous traveling waves is through varying the rate of decay ξ_1.

With the same parameter values as in Figure 6.4, profiles of n and g at a fixed time for three cases where the initial cell density distribution decreases at a rapid, moderate, and slow exponential rate are given in Figure 6.5. The left-most profile corresponds to a rapidly decaying initial condition. The cell density profile shows that the cell front is discontinuous, with the discontinuity extending to $n = 0$, and therefore the profile has compact support. The gradient of the chemoattractant profile is also discontinuous at the same position, namely, the smallest value of x where $g = 1$. This corresponds to the case where the trajectory in the phase plane jumps across the

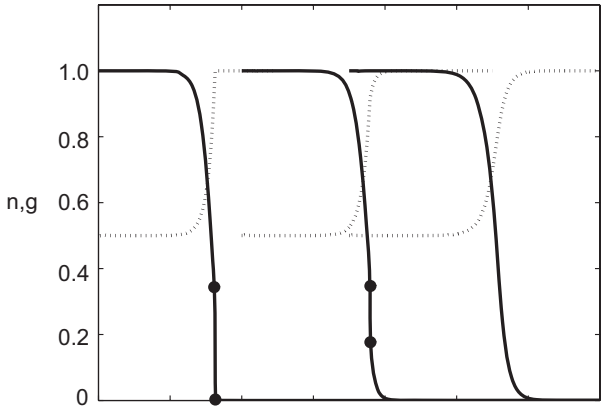


FIG. 6.5. *Numerical profiles for $n(x,t)$ (solid line) and $g(x,t)$ (dotted line) with $\gamma = 1.0$, $\beta = 1.0$. Left to right $n(x,t)$: Discontinuous solution with maximum shock length ($n_R = 0$) with $\xi_1 = 3.0$; discontinuous solution with smaller shock length ($n_R > 0$) with $\xi_1 = 1.5$; continuous solution with $\xi_1 = 1$. The end points of the shocks (•) are shown. Numerical computations were performed with $\Delta x = 0.05$ and $\Delta t = 0.01$.*

wall to the completely colonized steady state, as in Figure 6.4(b). The middle profile in Figure 6.5 corresponds to an initial condition where the decay is moderate. This profile shows a smaller discontinuity in the cell density; however, the discontinuity does not extend to the base of the profile as the toe of the profile is continuous. Again there is a discontinuity in the gradient of the chemoattractant at the same position where the discontinuity in the cell density occurs. Finally, with a slowly decaying initial condition, the distributions of the cell density, chemoattractant concentration, and the gradient of the chemoattractant concentration are continuous, as shown in the rightmost profile. The phase plane corresponding to this final case has the two steady states on the same side of the wall, as in Figure 6.4(a). With the parameter values used in Figure 6.3, the numerical solutions are qualitatively similar.

The profiles in Figure 6.5, together with the phase diagrams in Figures 6.3 and 6.4, give a comprehensive understanding of the behavior of the traveling wave solutions obtained from (2.7)–(2.8) when $D = 0$ and chemotaxis is the only cell migration process. Similar to the alternative diffusion-only case ($\chi = 0$), the existence of traveling wave solutions is established. In contrast, migration by pure diffusion cannot give rise to discontinuous solutions because of the smoothing nature of linear diffusion. However, both these limiting cases show that the speed of the resulting traveling wave solution is determined by the exponential decay rate of the initial distribution of the migrating cell population.

Finally, it follows from our scaling arguments (2.6), (6.3)–(6.4), that the minimum wave speed for the chemotaxis-only migration case scales with $\sqrt{\chi}$ and hence has the form

$$c_{min} = K(\beta,\gamma)\sqrt{\chi}, \tag{6.10}$$

where $K(\beta,\gamma)$ is a constant dependent on the kinetic parameters. This was anticipated in the earlier numerical simulations presented in Figure 4.1(a). Therefore c_{min} has a similar form to the minimum wave speed of $2\sqrt{D}$ for the diffusion driven migration as discussed in section 6.1. The major difference is that the coefficient $K(\beta,\gamma)$ is

not a constant but varies in a complicated way with the kinetic parameters β and γ. The two examples discussed above provide the criterion for determining K. As in Example 1, if $c_{min}^2 > \frac{\beta\gamma}{\beta+\gamma}$, then c_{min} is defined as that value of c such that the trajectory from the hole in the wall passes through $(1, c^2/\gamma)$. Alternatively, as in Example 2, if $c_{min}^2 < \frac{\beta\gamma}{\beta+\gamma}$, then c_{min} is defined as that value of c such that the trajectory from the steady state $(\frac{\beta}{\beta+\gamma}, 1)$ passes through $(1, c^2/\gamma)$.

An analytical solution for $K(\beta, \gamma)$ has been attempted but does not appear possible at the present time. Instead, numerical solutions are used to compute the minimum wave speed $K(\beta, \gamma)$ over a range of kinetic parameters β and γ. The form of $K(\beta, \gamma)$ is shown in Figure 6.6. In general, $K(\beta, \gamma)$ decreases with increasing β and increases with increasing γ, that is, $\frac{\partial K}{\partial \beta} < 0$ and $\frac{\partial K}{\partial \gamma} > 0$. These trends can be understood by considering the biological processes associated with the kinetic terms. The steady state concentration $g = \frac{\beta}{\gamma+\beta}$ increases with increasing β or with decreasing γ. As this steady concentration increases, the chemotactic gradient decreases giving rise to slower traveling wave speeds and a reduced value of $K(\beta, \gamma)$. This intuitive argument agrees with the form of $K(\beta, \gamma)$ deduced with the numerical solutions shown in Figure 6.6. We also investigated whether $K(\beta, \gamma)$ depended on a similarity variable, such as the ratio $\frac{\beta}{\gamma}$ alone, as illustrated in Figure 6.6(c). It appears that $K(\beta, \gamma)$ has a similar shape for the wide range of $\frac{\beta}{\gamma}$ investigated. However, the location of the curve can vary considerably for various choices of β.

Incorporating both numerical and phase plane analyses in this work reveals a remarkable advantage regarding the development and testing of the numerical algorithm. In general, testing numerical schemes for coupled nonlinear migration problems can be very difficult because of a lack of suitable analytical solutions [35]. Using the phase plane for the pure chemotaxis problem quantifies certain properties of the solution, such as the critical wave speed c_{crit}, the minimum wave speed c_{min}, and the size of the discontinuity. This unique information is useful in developing the numerical scheme as these quantitative checks are invoked to ensure that the numerical scheme is accurate.

6.3. Migration with both diffusion and chemotaxis. A three-dimensional phase plane analysis of (5.1)–(5.3) does not provide a productive way for seeking traveling wave solutions. A numerical study is convenient for examining both the shape of the invading profile as well as the minimum wave speeds. In particular, the robust numerical algorithm presented here has no difficulty in generating numerical solutions for any value of the diffusion coefficient and chemotactic factor. Therefore, it is of interest to investigate how this general case of combined chemotaxis and diffusive migration relates to the two limiting cases when either D or χ is zero.

6.3.1. Numerical solution profiles. Various solution profiles showing the influence of increasing the chemotactic factor χ for a fixed value of diffusivity $D = 1$ are shown in Figure 6.7(a). Comparison of these profiles shows that their smooth shape evolves to one with a developing discontinuity as χ increases. Moreover, the gradient of both the cell density and chemoattractant concentration increases with χ. Since the profiles are plotted at a fixed time starting from the same initial data, the wave speed clearly increases with χ from the minimum wave speed associated with the Fisher equation. This increase in wave speed with χ is expected because the inclusion of a second migration process enhances cell migration.

Similarly, the effect of increasing D on the numerical solutions is shown in Figure 6.7(b). Now the steep profiles evolve to smooth, flatter profiles as the diffusivity

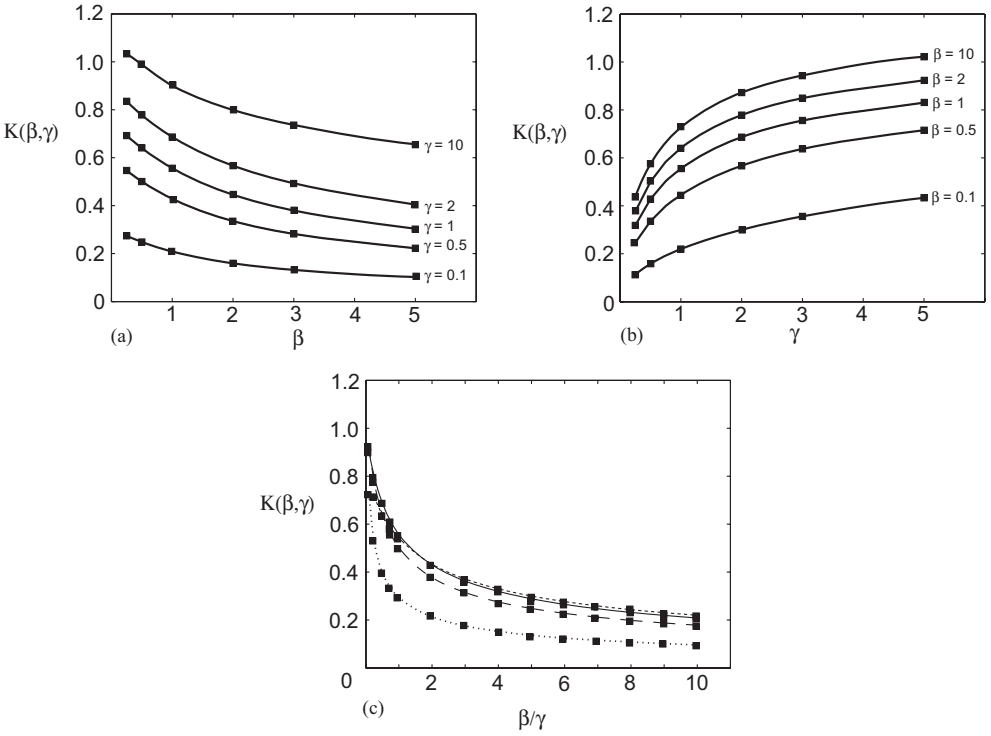


FIG. 6.6. *Dependence of $K(\beta,\gamma)$ on the kinetic parameters.* (*a*) *β for various γ values;* (*b*) *γ for various β values;* (*c*) *β/γ for $\beta = 10$ (short dashed line), $\beta = 1$ (solid line), $\beta = 0.5$ (long dashed line), and $\beta = 0.1$ (dotted line).*

increases, which reflect the smoothing nature of linear diffusion. These flatter profiles travel at a faster rate, as for the Fisher equation [25].

It is interesting to compare the rate at which the added migration processes competes with the underlying migration. In Figure 6.7(a), with the addition of chemotaxis to diffusive migration, the shape of the front steepens with increasing χ; however, the smooth shape is maintained fairly consistently up until $\chi = 50$ and it is not until $\chi = 100$ that the profile begins to tend toward the upper limit of chemotaxis-only migration with a discontinuous front. Conversely, in Figure 6.7(b), with the addition of diffusion to chemotactic migration, the shape of the front is very sensitive to the addition of a small amount of diffusion. The sharp front is smoothed with increasing D and tends toward the limit of diffusion-only migration for $D = 0.5$. These observations show that diffusion masks the influence of chemotaxis more efficiently than chemotaxis masks diffusion. These trends will now be more thoroughly explored in terms of the minimum wave speed c_{min}.

6.3.2. Minimum wave speed. For the system (5.1)–(5.3), a linear stability analysis of the steady state $(g, n, v) = (1, 0, 0)$ gives real eigenvalues if and only if $c \geq 2\sqrt{D}$, ensuring the point is a saddle point, just like for the Fisher equation. This condition provides a lower bound for the minimum wave speed. Numerical computations provide an extended analysis of the influence of mixed migration on the minimum wave speed. We examine the case where one of χ or D is held constant while simultaneously varying the other migration parameter. Numerical computations are

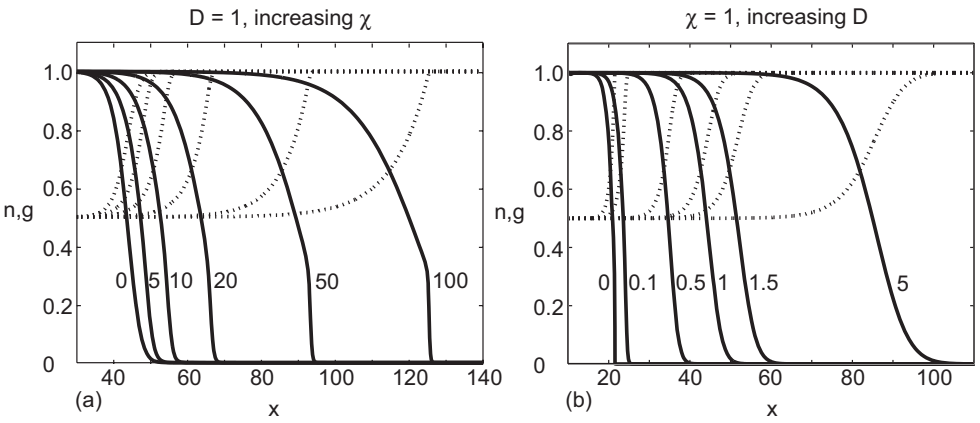


FIG. 6.7. *Numerical profiles of* $n(x, 20)$ (*solid line*) *and* $g(x, 20)$ (*dotted line*). (*a*) *The influence of increasing chemotaxis with fixed* $D = 1$, *with* χ *increasing from left to right with values indicated.* (*b*) *The influence of increasing diffusion with fixed* $\chi = 1$, *with* D *increasing from left to right with values indicated. All results were computed with* $\beta = 1, \gamma = 1$, $\xi_1 = 10$, $\Delta x = 0.05$, *and* Δt *was varied depending on* χ.

conducted to determine the effect on c_{min}.

Fixing the value of the chemotactic factor, namely, $\chi = 1$, the minimum wave speed increases monotonically with the diffusion coefficient D, as shown in Figure 6.8(a). Furthermore, as the diffusion coefficient increases, the minimum wave speed asymptotes to $2\sqrt{D}$. For this choice of kinetic parameters, $2\sqrt{D}$ provides a good approximation to c_{min} when $D/\chi > 0.2$. In general, for sufficiently large D/χ, diffusion dominates over chemotaxis and the minimum wave speed is accurately approximated by the Fisher wave speed $c_{min} = 2\sqrt{D}$, while for smaller values of D/χ, chemotaxis dominates and c_{min} is greater than that associated with the Fisher equation or chemotaxis alone. The numerical results for large D lie a little below $2\sqrt{D}$; this trend was also found for haptotactic invasion with added diffusion [19].

Similarly, setting the diffusion coefficient as $D = 1$, the c_{min} monotonically increases with the chemotactic factor χ and asymptotes to $K(\beta, \gamma)\sqrt{\chi}$, as illustrated in Figure 6.8(b). In this example, $K(\beta, \gamma)\sqrt{\chi}$ gives a good approximation to c_{min} when $\chi/D > 50.0$. In general for sufficiently large χ/D, chemotaxis dominates over diffusion and the minimum wave speed is well approximated by $c_{min} = K(\beta, \gamma)\sqrt{\chi}$ as given in (6.10). Conversely, for smaller values of χ/D, diffusion dominates and c_{min} is greater than that associated with chemotaxis or diffusion alone.

An explicit formula for c_{min} as a function of χ and D (as well as the kinetic parameters) has not been determined at this stage. However, some descriptive comments can be made. The discussion above indicates a natural lower bound for c_{min} as $\max[K(\beta, \gamma)\sqrt{\chi}, 2\sqrt{D}]$. An upper bound can be conjectured, as indicated in Figure 6.8. These can be combined as

$$\max[K(\beta, \gamma)\sqrt{\chi}, 2\sqrt{D}] < c_{min} < \sqrt{4D + K^2(\beta, \gamma)\chi}. \tag{6.11}$$

This expression suggests that diffusion dominates over chemotaxis when $\frac{K^2(\beta,\gamma)\chi}{4D} \ll 1$, and alternatively that chemotaxis dominates over diffusion when $\frac{4D}{K^2(\beta,\gamma)\chi} \ll 1$.

6.3.3. Perturbation analysis. When chemotaxis is small compared to diffusive migration, namely, $\chi/D \ll 1$, a regular perturbation analysis could be undertaken to

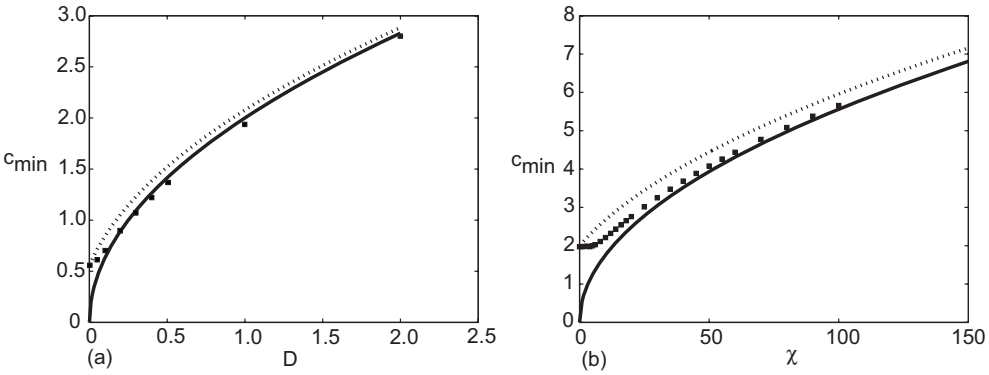


FIG. 6.8. *Numerically calculated minimum wave speed c_{min} shown with squares evolving from initial data with $\xi_1 = 10$. Results computed with $\beta = 1, \gamma = 1$. (a) Dependence on D with $\chi = 1$. The solid curve is $c_{min} = 2\sqrt{D}$. (b) Dependence on χ with $D = 1$. The solid curve is $c_{min} = K(\beta,\gamma)\sqrt{\chi}$; for this example $K(\beta,\gamma) \approx 0.556$. The dotted curves are the conjectured upper bound $c_{min} = \sqrt{4D + K^2(\beta,\gamma)\chi}$.*

give a solution valid for small χ/D. The first-order terms for n would just be the solution to the Fisher equation. This analysis is not very insightful and therefore is not shown here. A more illuminating analysis comes from the alternative case when D/χ is small.

Marchant [19] examined the case where a small amount of diffusion was added to a haptotactic invasion problem, using singular perturbation and phase plane arguments. A similar analysis is performed here but can be taken further and solved exactly. As discussed in section 6.2, when $D = 0$, our model supports discontinuous traveling wave solutions for a range of values of c. We know that a small amount of diffusion added to a purely chemotactic system has the effect of smoothing out any discontinuities. However, the gradients are expected to remain large in a small region. When D/χ is small, a perturbation analysis provides an understanding of the transition region. The analysis determines the evolution from a discontinuous traveling wave solution ($D = 0$) to one which is smooth, but has large derivative, in a small critical layer. Set $\chi = 1$ without any loss of generality. With $D \ll 1$, we seek solutions to (5.1)–(5.3) as an asymptotic expansion in terms of D as

$$g = g_0(z) + Dg_1(z) + D^2 g_2(z) + \cdots, \tag{6.12}$$

$$n = n_0(z) + Dn_1(z) + D^2 n_2(z) + \cdots, \tag{6.13}$$

$$v = v_0(z) + Dv_1(z) + D^2 v_2(z) + \cdots. \tag{6.14}$$

Hence g_0 and n_0 will satisfy (6.3)–(6.4). We choose to consider the traveling wave solution with the minimum wave speed c_{min}. We shift the origin so that the jump occurs at $z = 0$. This solution is the first term in the *outer* solution of the asymptotic expansion of the solution. At $z = 0$, for small D, there will be a narrow region where the rates of change of n are large, since n has to connect the left-hand limit n_L and right-hand limit $n_R = 0$. In this critical layer we seek a solution in the expanded variable $\xi = z/D$ as

$$g = G_0(\xi) + DG_1(\xi) + D^2 G_2(\xi) + \cdots, \tag{6.15}$$

$$n = N_0(\xi) + DN_1(\xi) + D^2 N_2(\xi) + \cdots, \tag{6.16}$$

$$v = \frac{1}{D}V_0(\xi) + V_1(\xi) + DV_2(\xi) + \cdots. \tag{6.17}$$

Substitution into (5.1)–(5.3) yields the highest-order terms satisfying

$$\frac{dG_0}{d\xi} = 0, \tag{6.18}$$

$$\frac{dN_0}{d\xi} = V_0, \tag{6.19}$$

$$\frac{dV_0}{d\xi} = -\left[1 + \frac{1}{c_{min}^2}(\beta(1-G_0) - 2\gamma N_0 G_0)\right] c_{min} V_0. \tag{6.20}$$

For the inner solution to match the outer solution, we require

$$G_0 = 1, \quad \xi \to \pm\infty, \tag{6.21}$$

$$N_0 = n_L = \frac{c_{min}^2}{\gamma}, \quad \xi \to -\infty\,, \qquad N_0 = n_R = 0, \quad \xi \to \infty, \tag{6.22}$$

$$V_0 = 0, \quad \xi \to \pm\infty. \tag{6.23}$$

Note that the value of n_L is obtained using the jump condition (6.7). Equations (6.18) and (6.21) give $G_0(\xi) = 1$ for all ξ. This simplifies the coupled system (6.19)–(6.20) as

$$\frac{dV_0}{d\xi} = -\left[1 - \frac{2\gamma}{c_{min}^2}N_0\right] c_{min}\frac{dN_0}{d\xi} = -c_{min}\left[\frac{dN_0}{d\xi} - \frac{2\gamma}{c_{min}^2}N_0\frac{dN_0}{d\xi}\right], \tag{6.24}$$

which integrates to

$$\frac{dN_0}{d\xi} = V_0 = -c_{min}\left(N_0 - \frac{\gamma}{c_{min}^2}N_0^2\right), \tag{6.25}$$

where the integration constant is zero from the conditions at $\xi \to \infty$. This is a logistic equation with solution

$$N_0 = \frac{c_{min}^2}{\gamma}\frac{e^{-c_{min}\xi}}{1 + e^{-c_{min}\xi}}, \tag{6.26}$$

where $N_0(0) = c_{min}^2/(2\gamma)$ with no loss of generality. Therefore, adding a small amount of diffusion introduces a steep transition region, of width D with exponential behavior depending on $c_{min}z/D$ (having set $\chi = 1$).

Figure 6.9 compares the numerically generated solutions to the perturbation analysis logistic solution (6.26). The region about the sharp front is stretched via the transformation $\xi = z/D$ so that the gradient is $\mathcal{O}(1)$ in the ξ coordinate. The numerical profile is translated so that $n(\xi, t) = c_{min}^2/(2\gamma)$ occurs at $\xi = 0$, as it does for N_0. The profiles of the leading order perturbation analysis and the numerically generated solution compare very well in the leading edge for $\xi > 0$ for small values of D as shown. The perturbation solution does not match as well in the region $\xi < 0$ for two reasons. First, we have matched with the jump density n_L as $\xi \to -\infty$, whereas the full numerical solution goes to the $n = 1$ state. Second, the slope of n as $\xi \to -\infty$ does not match at this dominant order of the approximation. The next order term, V_1, would be required to match the slope of the outer solution $\frac{\partial n_0}{\partial z}$ at the left of the shock as $\xi \to -\infty$.

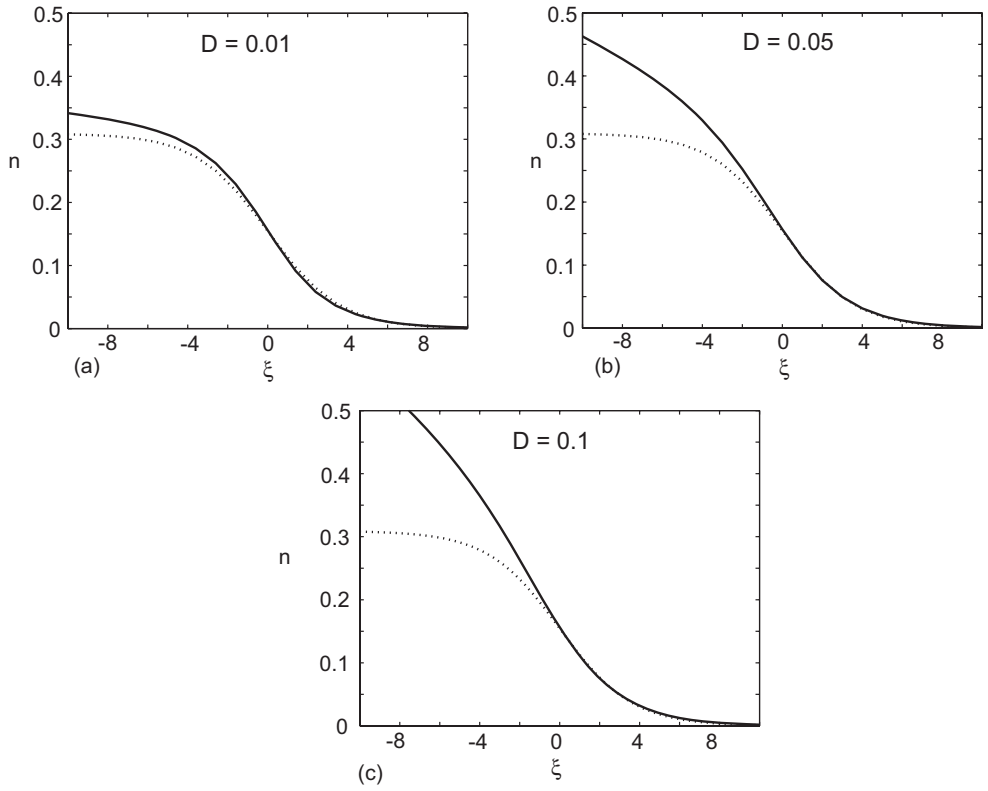


FIG. 6.9. *Critical layer comparison of numerical solution with the dominant perturbation solution* $N_0(\xi)$ *in* (6.26). *The solid curve is the numerical solution* $n(\xi, 20)$ *and the dotted curve is* $N_0(\xi)$. *Results computed with* $\chi = 1$, $\beta = 1, \gamma = 1$ *and using initial data with* $\xi_1 = 10$. (*a*) $D = 0.01$, (*b*) $D = 0.05$, (*c*) $D = 0.1$.

7. Conclusions. This article considers a mathematical model of cell invasion, where both diffusion and chemotaxis are the migration mechanisms. The details of the model were developed such that the results of the analysis are applicable to certain cell migration processes which are known to occur in developmental biology. A suite of traveling wave solutions is shown to exist regardless of whether the migration is pure diffusive, pure chemotaxis, or a combination of diffusive and chemotaxis migration. For all three cases, the traveling wave speed is bounded from below. The minimum wave speed is always observed whenever numerical simulations are performed using initial data where the cell density has compact support. Since the initial distribution of invading cells usually falls to zero for x large enough, this seems to be the most biologically relevant situation. Therefore, in general the most biologically relevant solution for these cell migration models is the solution corresponding to the minimum wave speed.

An understanding of the nature of the minimum wave speed as a function of the migration parameters is important. Phase plane analysis can provide values for the minimum wave speed for the two limiting cases when either the migration is purely diffusive or chemotactic. These values and the explicit shapes of the solutions can also be found using numerical methods. In particular, a robust numerical algorithm is developed which gives stable traveling waves solutions including shocks. The numeri-

cal algorithm combines a high-accuracy explicit central scheme [14] for the nonlinear hyperbolic and reaction terms together with a standard implicit finite element solution of the diffusion term with an operator split approach. The use of operator splitting for this particular problem was critical in combining the numerical solutions of the chemotaxis and diffusion terms together in a way that conveniently minimized numerical stability issues. Therefore, the numerical algorithm presented in this work provides an extremely accurate and versatile means of solving combined chemotaxis and diffusive migration problems.

For the combined diffusion and chemotactic migration case, numerical results were used to determine an upper and a lower bound on the minimum wave speed. Numerical results also demonstrate how the diffusion and chemotaxis mechanisms interact in a combined migration problem. The rate at which the minimum wave speed for the mixed migration case approached the minimum wave speed for the two limiting cases indicated that diffusion dominates over chemotaxis for relatively small values of the ratio of $\frac{D^*}{\chi^*} = \frac{D\lambda_3}{\chi\lambda_2}$.

The results from the combined diffusion and chemotaxis case indicate that adding a small amount of diffusion to a pure chemotaxis problem can result in the chemotactic characteristics of the problem being completely masked by the added diffusion. This observation is particularly relevant for numerical computations, when parabolic solvers are often used for chemotaxis (or haptotaxis) dominated processes. Further, this result also implies that standard numerical solutions of chemotaxis problems might be extremely sensitive to numerical diffusion and so great care should be exercised in obtaining such solutions.

In summary, this analysis provides a deeper qualitative and quantitative understanding of the interplay between diffusion and chemotaxis for invading cell populations. Often, when modeling biological cell migration, parameter values are difficult to estimate. If the wave speed can be determined experimentally, and the diffusion rate estimated, then some reasonable estimates of the chemotactic term may be deduced from the results presented here.

REFERENCES

[1] H. M. Byrne, M. A. J. Chaplain, G. J. Pettet, and D. L. S. McElwain, *A mathematical model of trophoblast invasion,* J. Theoret. Med., 1 (1999), pp. 275–286.

[2] H. M. Byrne, M. A. J. Chaplain, G. J. Pettet, and D. L. S. McElwain, *An analysis of a mathematical model of trophoblast invasion*, Appl. Math. Lett., 14 (2001), pp. 1005–1010.

[3] R. Courant and D. Hilbert, *Methods of mathematical physics. Vol.* II, 2nd ed., Interscience, New York, 1964.

[4] J.Crank, *The Mathematics of Diffusion*, 2nd ed., Oxford University Press, Oxford, UK, 1975.

[5] D. Dormann and C. J. Weijer, *Chemotactic cell movement during development*, Curr. Opin. Genet. Dev., 13 (2003), pp. 358–364.

[6] R. A. Fisher, *The wave of advance of advantageous genes*, Ann. Eugenics, 7 (1937), pp. 353–369.

[7] R. M. Ford and P. T. Cummings, *Mathematical models of bacterial chemotaxis,* in Mathematical Modeling in Microbial Ecology, A. L. Koch, J. A. Robinson, and G. A. Milliken, eds., Chapman and Hall, New York, 1998, pp. 228–269.

[8] R. M. Ford and D. A. Lauffenburger, *Analysis of chemotactic bacterial distributions in population migration assays using a mathematical model applicable to steep or shallow attractant gradients,* Bull. Math. Biol., 53 (1991), pp. 721–749.

[9] C. J. Hearn, M. Murphy, and D. F. Newgreen, *GDNF and ET-3 differentially modulate the numbers of avian enteric neural crest cells and enteric neurons in vitro,* Dev. Biol., 197 (1998), pp. 93–105.

[10] T. Hillen, *Hyperbolic models for chemosensitive movement,* Math. Models Methods Appl. Sci., 12 (2002), pp. 1007–1034.

[11] D. Horstmann and A. Stevens, *A constructive approach to traveling waves in chemotaxis,* J. Nonlinear Sci., 14, (2004), pp. 1–25.
[12] J. Kassis, D. A. Lauffenburger, T. Turner, and A. Wells, *Tumor invasion as dysregulated cell motility,* Sem. Cancer. Biol., 11 (2001), pp. 105–119.
[13] E. F. Keller and L. A. Segel, *Travelling bands of chemotactic bacteria: A theoretical analysis,* J. Theoret. Biol., 30 (1971), pp. 235–248.
[14] A. Kurganov and E. Tadmor, *New high-resolution central schemes for non-linear conservation laws,* J. Comput. Phys., 160 (2000), pp. 241–282.
[15] N. M. Le Douarin and M. A. M. Teillet, *Experimental analysis of the migration and differentiation of the autonomic nervous system and of neurectodermal mesenchymal derivatives using a biological cell marking technique,* Dev. Biol., 41 (1974), pp. 162–184.
[16] R. J. LeVeque and J. Oliger, *Numerical methods based on additive splittings for hyperbolic partial differential equations,* Math. Comput., 40 (1983), pp. 469–497.
[17] K. A. Landman, G. J. Pettet, and D. F. Newgreen, *Mathematical models of cell colonisation of uniformly growing domains,* Bull. Math. Biol., 65 (2003), pp. 235–262.
[18] K. A. Landman, G. J. Pettet, and D. F. Newgreen, *Chemotactic cellular migration: Smooth and discontinuous travelling wave solutions,* SIAM J. Appl. Math., 63 (2003), pp. 1666–1681.
[19] B. P. Marchant, *Modelling Cell Invasion*, Ph.D. thesis, University of Oxford, Oxford, UK, 1999.
[20] B. P. Marchant, J. Norbury, and A. J. Perumpanani, *Travelling shock waves arising in a model of malignant invasion,* SIAM J. Appl. Math., 60 (2000), pp. 463–476.
[21] B. P. Marchant, J. Norbury, and J. A. Sherratt, *Travelling wave solutions to a haptotaxis-dominated model of malignant invasion,* Nonlinearity, 14 (2001), pp. 1653–1671.
[22] B. P. Marchant and J. Norbury, *Discontinuous travelling wave solutions for certain hyperbolic systems,* IMA J. Appl. Math., 67 (2002), pp. 201–224.
[23] H. P. McKean, *Application of Brownian motion to the equation of Kolmogorov-Petrovskii-Piskunov,* Comm. Pure Appl. Math., 28 (1975) pp. 323–331.
[24] J. D. Murray and M. R. Myerscough, *Pigmentation pattern formation on snakes,* J. Theoret. Biol., 149 (1991), pp. 339–360.
[25] J. D. Murray, *Mathematical Biology,* 2nd ed., Springer-Verlag, Heidelberg, 1993.
[26] D. Natarajan, C. Marcos-Gutierrez, V. Pachnis, and E. de Graaff, *Requirement of signalling by receptor tyrosine kinase RET for the directed migration of enteric nervous system progenitor cells during mammalian embryogenesis,* Development, 129 (2002), pp. 5151–5160.
[27] D. F. Newgreen, *Control of the directional migration of mesenchyme cells and neurites,* Sem. Developmental Biol., 1 (1990), pp. 301–311.
[28] D. F. Newgreen, B. Southwell, L. Hartly, and I. J. Allan, *Migration of enteric neural crest cells in relation to growth of the gut in avian embryos,* Acta Anat., 157 (1996), pp. 105–115.
[29] H. G. Othmer and A. Stevens, *Aggregation, blowup, and collapse: The ABC's of taxis in reinforced random walks,* SIAM J. Appl. Math., 57 (1997), pp. 1044–1081.
[30] A. J. Perumpanani, J. A. Sherratt, J. Norbury, and H. M. Byrne, *A two parameter family of travelling waves with a singular barrier arising from the modelling of extracellular matrix mediated cell invasion,* Phys. D, 126 (1999), pp. 145–159.
[31] G. J. Pettet, H. M. Byrne, D. L. S. McElwain, and J. Norbury, *A model of wound-healing angiogenesis in soft tissue,* Math. Biosci., 136 (1996), pp. 35–63.
[32] G. J. Pettet, D. L. S. McElwain, and J. Norbury, *Lotka–Volterra equations with chemotaxis: Walls, barriers and travelling waves,* IMA J. Math. Appl. Med. Biol., 17 (2000), pp. 395–413.
[33] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in FORTRAN: The Art of Scientific Computing,* 2nd ed., Cambridge University Press, Cambridge, UK, 1992.
[34] L. J. Segerlind, *Applied Finite Element Analysis,* 2nd ed., John Wiley and Sons, Singapore, 1984.
[35] M. J. Simpson and T. P. Clement, *A theoretical analysis of the worthiness of Henry and Elder problems as benchmarks of density-dependent groundwater flow models,* Adv. Water Resour., 26 (2003), pp. 17–31.
[36] M. J. Simpson, K. A. Landman, and T. P. Clement, *Assessment of a non-traditional operator split algorithm for simulation of reactive transport*, Math. Comput. Simulation, in press, 2005.

[37] M. Starz-Gaiano and D. J. Montell, *Genes that drive invasion and migration in Drosophila,* Curr. Opin. Genet. Dev., 14 (2004), pp. 86–91.

[38] G. Strang, *On the construction and comparison of difference schemes,* SIAM J. Numer. Anal., 5 (1968), pp. 506–517.

[39] R. T. Tranquillo, *Perspectives and models of gradient perception,* in Biology of the Chemotactic Response, J. P. Armitage and J. M. Lackie, eds., Cambridge University Press, Cambridge, UK, 1991, pp. 35–75.

[40] R. T. Tranquillo and W. Alt, *Receptor-mediated models for leukocyte chemotaxis,* in Dynamics of Cell and Tissue Motion, W. Alt, A. Deutsch, and G. Dunn, eds., Birkhäuser, Berlin, 1997, pp. 141–147.

[41] A. J. Valocchi and M. Malmstead, *Accuracy of operator splitting for advection-dispersion-reaction problems,* Water Resour. Res., 28 (1992), pp. 1471–1476.

[42] G. T. Yeh, *Computational Subsurface Hydrology. Reactions, Transport and Fate,* Kluwer Academic Publishers, Norwell, MA, 2000.

[43] H. M. Young, C. J. Hearn, P. G. Farlie, A. J. Canty, P. Q. Thomas, and D. F. Newgreen, *GDNF is a chemoattractant for enteric neural crest cells,* Dev. Biol., 229 (2001), pp. 503–516.

[44] C. Zheng and G. D. Bennett, *Applied Contaminant Transport Modelling,* John Wiley and Sons, New York, 2002.

SIAM J. APPL. MATH.
Vol. 65, No. 4, pp. 1443–1462

ASYMPTOTIC THEORY OF ELECTROSEISMIC PROSPECTING*

BENJAMIN S. WHITE†

Abstract. In a porous medium such as the earth's subsurface, electromagnetic (EM) waves and mechanical waves are coupled through the phenomenon of electrokinetics, for which a complete set of partial differential equations was derived by S. Pride. In this paper, we derive from Pride's equations an asymptotic theory that enables forward modeling of the seismic response to an EM source in fully three-dimensional geometries on a scale that is relevant to exploration. For simplicity, we consider piecewise homogeneous media separated by interfaces which are curved surfaces in three dimensions. The following physical picture emerges: An EM source excites an EM wave which propagates into the earth, stirring up local mechanical movement. At an interface, EM energy is converted to seismic waves, which may be described by ray theory. Instantly, on the seismic time scale, every interface becomes a wavefront for both compressional and shear waves; that is, seismic P- and S-waves explode from both sides of each interface, at every point on it. The rays for these waves leave the interface in the orthogonal direction and propagate up and down into the homogeneous media on both sides of the surface. We derive formulas for the initial amplitudes of these waves. Conventional seismic ray theory then describes propagation of the P- and S-waves, including reflection, transmission, and mode conversion at any other interfaces that they may encounter. Thus, three-dimensional electroseismic modeling may be accomplished with conventional EM and conventional seismic modeling tools, using the present theory to provide the link between them.

Key words. electrokinetics, Biot theory, ray theory, WKB expansion, electromagnetic waves, seismic waves

AMS subject classifications. 35Q99, 41A60

DOI. 10.1137/040604108

1. Introduction.

In a porous medium such as the earth's subsurface, electromagnetic (EM) waves and mechanical waves are coupled through the phenomenon of electrokinetics [14]. Ions in the pore fluid are attracted to ions of the opposite sign in the solid at the pore walls, so that there is an electrical double layer, called the Debye layer, at the pore boundaries. An electric field acting on this double layer will move the ions relative to each other, creating movement of both the fluid and the solid. Conversely, a mechanical wave which moves the fluid and solid relative to each other will create an EM wave. This phenomenon was proposed as the basis of a hydrocarbon exploration method in 1936 by R. R. Thompson of Humble Oil Company, in volume 1 of *Geophysics* [19]. It was described theoretically 60 years ago [7] and has been demonstrated in the laboratory, where the magnitude of the coupling, called the electrokinetic mobility, has been measured (see [13] and references therein).

Starting with the microscopic description, Pride [14] derived from first principles a complete set of macroscopic equations describing electrokinetics. These are 19 scalar partial differential equations in which Maxwell's equations for EM are coupled with Biot's equations [1, 2, 3, 15] for movement of fluid and solid in a porous medium. Some general properties of these equations are known, including uniqueness, source/receiver reciprocity, energy conservation, and point source and plane wave responses in homogeneous media [17]. Also, computer codes have been written to compute the solution of Pride's equations in plane layered media [8], that is, media which are piecewise

*Received by the editors February 11, 2004; accepted for publication (in revised form) October 18, 2004; published electronically April 26, 2005.
http://www.siam.org/journals/siap/65-4/60410.html
†ExxonMobil Corporate Strategic Research, Route 22 East, Annandale, NJ 08801 (benjamin.s.white@exxonmobil.com).

homogeneous and which have material properties that vary only in one dimension, which is depth.

Techniques based on electrokinetics have been proposed for use in exploration of the earth's subsurface, and several groups have conducted field experiments [19, 18, 4, 11, 12]. All of these experiments were seismoelectric, i.e., a mechanical (seismic) source was used, and an EM wave was detected. In this paper, we will analyze the electroseismic method, where an EM source is used and a seismic wave is detected.

Electrokinetic prospecting methods are especially promising for detecting materials whose electrical resistivity varies markedly from that of the background, e.g., hydrocarbons. EM methods [20, 6] may also be used for this, but their spatial resolution is usually less than desirable because of the long wavelengths of EM waves in the earth. Seismic methods [6] have much better spatial resolution, but they respond to much smaller contrasts in material properties. It may be hoped that electrokinetic methods can combine the virtues of both approaches.

In this paper, we will derive an asymptotic theory that enables forward modeling of the electroseismic response in fully three-dimensional geometries on a scale that is relevant to exploration. For these problems, the cost of direct calculation, for instance with finite differences, is prohibitive. In deriving the theory, we use three basic assumptions: first, that the EM and mechanical coupling is weak; second, that, as is common in seismology, conventional elasticity theory may be used in place of Biot's theory; and third, that the scales are such that seismic ray theory is valid. Each of these assumptions is specified in terms of a single parameter that is assumed to be large or small. Our asymptotic theory is then derived from a systematic perturbation expansion of Pride's full system of equations for electrokinetics. For simplicity, we consider only piecewise homogeneous media separated by interfaces which are curved surfaces in three dimensions.

The physical picture that emerges from the perturbation expansion is as follows: An EM source excites an EM wave which propagates into the earth, stirring up local mechanical movement as it passes. At an interface, EM energy is converted to seismic waves, which may be described by ray theory. Instantly, on the seismic time scale, every interface becomes a wavefront for both compressional and shear waves; that is, seismic P- and S-waves explode from both sides of each interface, at every point on it. The rays for these waves leave the interface in the orthogonal direction, and propagate up and down into the homogeneous media on both sides of the surface. Conventional seismic ray theory [5, 10] then describes propagation of the P- and S-waves, including reflection, transmission, and mode conversion at any other interfaces that they may encounter.

All that is necessary to complete this picture are the initial amplitudes of the P- and S-waves when they originate at the interfaces. Formulas for these initial amplitudes are given below in section 8. These formulas depend on the material parameters and on values of the electric field on both sides of the interface where the seismic waves originate.

With this theory, a general computer program for three-dimensional electroseismic modeling can be constructed, using commercially available software for the major computational tasks. First, the EM field is computed using, say, a finite element EM solver. A small subroutine then uses the formulas of section 8, combined with the computed values of the electric field, to get the initial P- and S-wave amplitudes on every interface. Finally, a ray trace program may be used to follow the rays that are created. In this way, three-dimensional electroseismic modeling may be accomplished

with conventional EM and conventional seismic modeling tools, using the present theory to provide the link between them.

This paper is organized as follows.

In section 2, we nondimensionalize Pride's equations, and introduce the small and large parameters used for the perturbation analysis. In section 3 we show how the problem can be reduced to solving the usual homogeneous Biot's equations, but with inhomogeneous boundary conditions at all the interfaces. For each interface, we derive eight scalar jump conditions for Biot's equations, conditions that depend on the material parameters and on the computed values of the electric field on both sides of the interface.

In sections 4–7 we derive a two-parameter asymptotic expansion for Biot's equations. This approximation is applicable when seismic ray theory is valid, and when Biot's equations give rise to waves that look like the conventional seismic P- and S-waves which are derived from the theory of an elastic solid. Besides the P- and S-waves, we obtain the form, in our approximation, of the Biot slow wave.

The Biot slow wave is a diffusive wave, which decays rapidly to zero with propagation distance. It is therefore difficult to observe, and laboratory observation of it in real rocks has only been accomplished recently [9]. However, as is shown in section 8, EM energy is converted to Biot slow waves at an interface, and slow waves must be considered in order to calculate the amplitudes of the P- and S-waves which also originate there. The necessity of including the Biot slow waves for energy conversion at an interface is consistent with the results of Pride and Garambois [16] for conversions of seismic to EM energy.

In section 8, we combine our asymptotic Biot theory with the interface conditions of section 3 to derive the initial amplitudes of all the waves created at an interface by an EM source.

In section 9, we compare the asymptotic theory to the results of a computer program which is designed to compute the electroseismic response in plane layered media.

Concluding remarks are in section 10.

2. An EM source in the seismic band. For an EM source current $\mathbf{j_s}$ in a porous medium, Pride's equations for the electric and magnetic field vectors, $\mathbf{E}$ and $\mathbf{H}$, respectively, are

$$\nabla \mathbf{x} \mathbf{E} = i\omega\mu \mathbf{H}, \tag{1}$$

$$\nabla \mathbf{x} \mathbf{H} = (\sigma - i\epsilon\omega)\mathbf{E} + L(-\nabla p + \omega^2 \rho_f \mathbf{u}) + \mathbf{j_s}, \tag{2}$$

where ω is frequency, σ, ϵ, and μ are, respectively, conductivity, dielectric constant, and magnetic permeability, L is the electrokinetic mobility parameter, p is the pore pressure, ρ_f is the density of the pore fluid, and $\mathbf{u}$ is the solid displacement.

Pride's equations for $\mathbf{u}$ and the relative fluid displacement $\mathbf{w}$ are

$$-\omega^2(\rho \mathbf{u} + \rho_f \mathbf{w}) = \nabla \cdot \tau, \tag{3}$$

$$-i\omega \mathbf{w} = L\mathbf{E} + (\kappa/\eta)(-\nabla p + \omega^2 \rho_f \mathbf{u}), \tag{4}$$

$$\tau = (\lambda \nabla \cdot \mathbf{u} + C \nabla \cdot \mathbf{w})\mathbf{I} + G(\nabla \mathbf{u} + \nabla \mathbf{u}^{\mathbf{T}}), \tag{5}$$

$$-p = C\nabla \cdot \mathbf{u} + M\nabla \cdot \mathbf{w}, \tag{6}$$

where τ is the stress tensor, $\mathbf{I}$ is the 3×3 identity, κ is the permeability, η is the viscosity of the pore fluid, λ and G are the Lamé parameters of elasticity, and C and M are the Biot moduli parameters.

Note that in this theory, the electrokinetic mobility L provides coupling between the EM system (1), (2) and the mechanical system (3)–(6). That is, when $L = 0$ these systems are decoupled. We assume weak coupling so that to leading (i.e., zeroth) order in L, the EM field satisfies (1), (2) with $L = 0$. These equations are the conventional Maxwell equations, and can be solved independently of the mechanical system. Then to leading (i.e., first) order in L the mechanical system satisfies (3)–(6), which are Biot's equations with the EM field as a source.

We consider these equations in a homogeneous region of space, i.e., where all parameters are constant. Then putting (5) into (3) yields

$$-\omega^2(\rho\mathbf{u} + \rho_f\mathbf{w}) = (\lambda + G)\nabla(\nabla \cdot \mathbf{u}) + G\nabla^{\mathbf{2}}\mathbf{u} + C\nabla(\nabla \cdot \mathbf{w}). \tag{7}$$

Putting (6) into (4) yields

$$-\omega^2(\rho_f\mathbf{u} + \tilde{\rho}\mathbf{w}) = C\nabla(\nabla \cdot \mathbf{u}) + M\nabla(\nabla \cdot \mathbf{w}) - i\omega\tilde{\rho}L\mathbf{E}, \tag{8}$$

where the pure imaginary parameter

$$\tilde{\rho} = \frac{i\eta}{\omega\kappa} \tag{9}$$

has units of density.

The equations may be nondimensionalized by introducing typical values $\bar{\rho}$, $\bar{\lambda}$, $\bar{G}$, $\bar{L}$, $\bar{E}$, and a typical length scale of the geometry, $\bar{l}$. Define

$$\begin{aligned} \bar{v} &= \sqrt{\frac{(\bar{\lambda} + 2\bar{G})}{\bar{\rho}}}, \\ \bar{u} &= \frac{\bar{l}}{\bar{v}}\bar{L}\bar{E}. \end{aligned} \tag{10}$$

Note that $\bar{v}$ is a typical compressional wave speed in elasticity theory, i.e., a typical seismic P-wave speed.

Define the dimensionless variables

$$\begin{aligned} &\mathbf{x}' = \mathbf{x}/\bar{l}, \quad \omega' = \omega\bar{l}/\bar{v}, \quad L' = L/\bar{L}, \\ &\rho' = \rho/\bar{\rho}, \quad \rho_f' = \rho_f/\bar{\rho}, \quad \tilde{\rho}' = \tilde{\rho}/\bar{\rho}, \\ &\lambda' = \lambda/(\bar{\lambda} + 2\bar{G}), \quad G' = G/(\bar{\lambda} + 2\bar{G}), \quad M' = M/(\bar{\lambda} + 2\bar{G}), \quad C' = C/(\bar{\lambda} + 2\bar{G}), \\ &\mathbf{u}' = \mathbf{u}/\bar{u}, \quad \mathbf{w}' = \mathbf{w}/\bar{u}, \quad p' = \frac{\bar{l}}{\bar{u}(\bar{\lambda} + 2\bar{G})}p, \quad \tau' = \frac{\bar{l}}{\bar{u}(\bar{\lambda} + 2\bar{G})}\tau, \quad \mathbf{E}' = \mathbf{E}/\bar{E}. \end{aligned} \tag{11}$$

Use of (10) and (11) shows that (5)–(8) are also satisfied by the primed variables.

We consider the seismic ray theory regime when a typical seismic wavelength is much smaller than a typical dimension $\bar{l}$. Thus the dimensionless frequency ω' satisfies the high frequency condition

$$\omega' >> 1. \tag{12}$$

However, the frequency is assumed to be subcritical for the porous medium, i.e.,

$$|\tilde{\rho}'| >> 1. \tag{13}$$

As is shown in sections 4–7, condition (13) assures that the dynamics of the porous medium may in most regions of space be approximated by the dynamics of elasticity theory, as is commonly assumed in seismology. In particular, this condition is necessary to obtain wave modes that approximate those of seismic P- and S-waves. Condition (13) also plays a role in microscopic theory, where it guarantees that the frequency is much less than the transition frequency separating low frequency viscous flow in the pores from high frequency inertial flow [14]. Thus the present theory corresponds to that in [1] rather than that in [2].

Because a typical EM wavelength is much larger than a seismic wavelength, $\mathbf{E}$ is not rapidly varying on the length scale $\bar{l}$. Therefore, ray theory is not appropriate for the calculation of the EM field, and the full Maxwell equations (1), (2) with $L = 0$ must be solved.

Equations (7) and (8) are satisfied in each homogeneous region of space. Let $\mathcal{S}$ be an interface, i.e., a surface separating two homogeneous regions, and let $\mathbf{n}$ be a normal to $\mathcal{S}$. Then Pride's interface conditions [17] are

$$\text{Continuity of} \quad \mathbf{u},\ p,\ \mathbf{w}\cdot\mathbf{n},\ \tau\cdot\mathbf{n} \quad \text{across } \mathcal{S}. \tag{14}$$

Note that (14) comprises eight scalar boundary conditions that must be satisfied at any interface between two homogeneous regions. Of course Maxwell's equations also require continuity of the tangential components of $\mathbf{E}$ and $\mathbf{H}$ across $\mathcal{S}$.

For notational convenience, we drop primes in what follows and use nondimensional units.

3. Interface conditions for the converted waves. Let

$$\mathbf{u} = \mathbf{u}^{(\mathbf{p})} + \hat{\mathbf{u}} \qquad \mathbf{w} = \mathbf{w}^{(\mathbf{p})} + \hat{\mathbf{w}}, \tag{15}$$

where $\mathbf{u} = \mathbf{u}^{(\mathbf{p})}$ and $\mathbf{w} = \mathbf{w}^{(\mathbf{p})}$ are particular solutions of (7) and (8) that do not necessarily satisfy the interface conditions (14), and $\hat{\mathbf{u}}, \hat{\mathbf{w}}$ are solutions of the homogeneous equations, i.e., (7) and (8) with $\mathbf{E} = \mathbf{0}$.

Particular solutions are easy to find asymptotically by seeking $\mathbf{u} = \mathbf{u}^{(\mathbf{p})}$ and $\mathbf{w} = \mathbf{w}^{(\mathbf{p})}$ that are not rapidly varying in space. From (7) and (8) we obtain

$$\begin{aligned} -i\omega\mathbf{u}^{(\mathbf{p})} &= -\left(\frac{\rho_f}{\rho}\right) L\mathbf{E} + O\left(\frac{1}{\omega^2}\right) + O\left(\frac{1}{|\tilde{\rho}|}\right), \\ -i\omega\mathbf{w}^{(\mathbf{p})} &= L\mathbf{E} + O\left(\frac{1}{|\tilde{\rho}|}\right). \end{aligned} \tag{16}$$

In what follows, we neglect the small terms in this equation. Note, however, that this approximation is only valid in the far field of the source, where the interfaces are assumed to lie. Near the source, gradients of the electric field may be large, invalidating (16). This raises the possibility that some EM energy is converted directly to seismic waves where the EM source contacts the ground, but we will not investigate these near-source seismic waves in the present theory.

The particular solution given by (16) represents a mechanical disturbance that propagates along with the exciting EM wave. In a completely homogeneous space,

this represents the asymptotic solution. However, if there are interfaces, new waves must be generated at those interfaces in order for $\mathbf{u}$ and $\mathbf{w}$ to satisfy the conditions (14). Let $\mathbf{n}$ be a normal to the interface $\mathcal{S}$. Continuity of $\mathbf{u}$ across $\mathcal{S}$ yields the jump condition

$$[|\mathbf{u}|] = \left[\left|\mathbf{u^{(p)}} + \hat{\mathbf{u}}\right|\right] = \mathbf{0}, \tag{17}$$

where $[|\cdot|]$ represents the jump in the value of a quantity across the interface $\mathcal{S}$. That is, let the normal $\mathbf{n}$ point to the side of the interface $\mathcal{S}^+$ and let $\mathcal{S}^-$ be the other side. The jump in any quantity A is

$$[|\ A\ |] \equiv A\ |_{\mathcal{S}^+} - A\ |_{\mathcal{S}^-}\ . \tag{18}$$

From (16) and (17) we obtain the jump condition for $\hat{\mathbf{u}}$

$$[|-i\omega\hat{\mathbf{u}}|] = \left[\left|\left(\frac{\rho_f L}{\rho}\right)\mathbf{E}\right|\right]. \tag{19}$$

Similarly, from continuity of $\mathbf{w}\cdot\mathbf{n}$ across $\mathcal{S}$ and (15), (16), we obtain

$$[|-i\omega\hat{\mathbf{w}}\cdot\mathbf{n}\,|] = -[|\,L\mathbf{E}\cdot\mathbf{n}\,|]. \tag{20}$$

Next, pressure p is decomposed into parts corresponding to the particular and homogeneous solutions

$$p = p^{(p)} + \hat{p}, \tag{21}$$

where $p^{(p)}$ satisfies (6) with $\mathbf{u},\mathbf{w}$ replaced by $\mathbf{u^{(p)}},\mathbf{w^{(p)}}$ and $\hat{p}$ satisfies (6) with $\mathbf{u},\mathbf{w}$ replaced by $\hat{\mathbf{u}},\hat{\mathbf{w}}$. However, in a homogeneous region of space

$$\nabla\cdot\mathbf{u^{(p)}} = \nabla\cdot\mathbf{w^{(p)}} = \mathbf{0}. \tag{22}$$

Equation (22) follows from taking the divergence of (16) and using the fact that the divergence of the electric field vanishes in a source-free region of homogeneous space. From (22) and (6) it follows that $p^{(p)} = 0$ and so $p = \hat{p}$. Continuity of p across $\mathcal{S}$ then yields the condition

$$[|\hat{p}|] = 0. \tag{23}$$

Finally, the stress tensor τ is decomposed into parts corresponding to the particular and homogeneous solutions

$$\tau = \tau^{(\mathbf{p})} + \hat{\tau}, \tag{24}$$

where $\tau^{(\mathbf{p})},\hat{\tau}$ each satisfy (5) for $\mathbf{u},\mathbf{w}$ replaced by, respectively, the particular or homogeneous quantities. Substituting (16) into (5) and using (22) yields

$$\tau^{(\mathbf{p})} = \left(\frac{-iG\rho_f L}{\rho\omega}\right)\left(\nabla\mathbf{E} + \nabla\mathbf{E}^{\mathbf{T}}\right). \tag{25}$$

Now use of (24), (25) in the condition (14) that $\tau\cdot\mathbf{n}$ be continuous across $\mathcal{S}$ yields

$$[|-i\omega\hat{\tau}\cdot\mathbf{n}|] = \left[\left|\left(\frac{G\rho_f L}{\rho}\right)(\nabla\mathbf{E} + \nabla\mathbf{E}^{\mathbf{T}})\cdot\mathbf{n}\right|\right]. \tag{26}$$

In summary, new waves will be generated at an interface $\mathcal{S}$. These waves will satisfy the homogeneous equations (7) and (8) with the electric field set equal to zero. However, they will satisfy the inhomogeneous interface conditions (19), (20), (23), and (26). These jump conditions comprise eight scalar equations, with inhomogeneous terms that depend on the electric field and on the values of the material parameters on both sides of the interface.

For notational convenience, we will drop hats in what follows.

4. WKB expansion. The WKB ansatz [10] for a high frequency wave with phase ϕ is an asymptotic expansion, as $\omega \to \infty$, of the form

$$\begin{aligned} \mathbf{u} &\sim e^{i\omega\phi}\left(\mathbf{u_0} + \frac{1}{i\omega}\mathbf{u_1} + \frac{1}{(i\omega)^2}\mathbf{u_2} + \cdots\right), \\ \mathbf{w} &\sim e^{i\omega\phi}\left(\mathbf{w_0} + \frac{1}{i\omega}\mathbf{w_1} + \frac{1}{(i\omega)^2}\mathbf{w_2} + \cdots\right). \end{aligned} \tag{27}$$

From the form of the interface conditions derived in section 3, it may be anticipated that $-i\omega\mathbf{u}, -i\omega\mathbf{w}$ are of order $O(1)$, i.e., $\mathbf{u_j}, \mathbf{w_j}$ are of order $O(1/\omega)$. This is of no consequence in the expansion (27), since factors of ω may be divided through the homogeneous Biot equations. It will, however, insert an extra factor of $1/\omega$ in the error estimates of sections 5–7.

Substituting (27) into (7) and equating coefficients of $(i\omega)^k$ to zero yields a series of equations. Similarly, substituting (27) into (8) with $\mathbf{E} = \mathbf{0}$ yields a second sequence of equations. The results are summarized below in matrix-vector form, where $\mathcal{L}$ is a 6×6 matrix defined below in terms of its 3×3 blocks. Let

$$\mathcal{L} = \mathcal{L}(\nabla\phi) = \begin{bmatrix} \left(\rho - G(\nabla\phi)^2\right)\mathbf{I} - (\lambda+G)\nabla\phi\nabla\phi^T & \rho_f\mathbf{I} - C\nabla\phi\nabla\phi^T \\ \rho_f\mathbf{I} - C\nabla\phi\nabla\phi^T & \tilde{\rho}\mathbf{I} - M\nabla\phi\nabla\phi^T \end{bmatrix}, \tag{28}$$

$$\begin{aligned} \mathbf{R_1}(\nabla\phi, \mathbf{u}, \mathbf{w}) &= (\lambda+G)\left\{(\nabla\cdot\mathbf{u})\nabla\phi + \nabla(\nabla\phi\cdot\mathbf{u})\right\} \\ &\quad + G\left\{2(\nabla\phi\cdot\nabla)\mathbf{u} + (\nabla^2\phi)\mathbf{u}\right\} + C\left\{(\nabla\cdot\mathbf{w})\nabla\phi + \nabla(\nabla\phi\cdot\mathbf{w})\right\}, \end{aligned} \tag{29}$$

$$\mathbf{R_2}(\nabla\phi, \mathbf{u}, \mathbf{w}) = C\left\{(\nabla\cdot\mathbf{u})\nabla\phi + \nabla(\nabla\phi\cdot\mathbf{u})\right\} + M\left\{(\nabla\cdot\mathbf{w})\nabla\phi + \nabla(\nabla\phi\cdot\mathbf{w})\right\}. \tag{30}$$

Then we obtain

$$\mathcal{L}\begin{bmatrix} \mathbf{u_0} \\ \mathbf{w_0} \end{bmatrix} = \mathbf{0}, \tag{31}$$

$$\mathcal{L}\begin{bmatrix} \mathbf{u_1} \\ \mathbf{w_1} \end{bmatrix} = \begin{bmatrix} \mathbf{R_1}(\nabla\phi, \mathbf{u_0}, \mathbf{w_0}) \\ \mathbf{R_2}(\nabla\phi, \mathbf{u_0}, \mathbf{w_0}) \end{bmatrix}, \tag{32}$$

and for $j = 2, 3, 4, \ldots,$

$$\begin{aligned} \mathcal{L}\begin{bmatrix} \mathbf{u_j} \\ \mathbf{w_j} \end{bmatrix} &= \begin{bmatrix} \mathbf{R_1}(\nabla\phi, \mathbf{u_{j-1}}, \mathbf{w_{j-1}}) \\ \mathbf{R_2}(\nabla\phi, \mathbf{u_{j-1}}, \mathbf{w_{j-1}}) \end{bmatrix} \\ &\quad + \begin{bmatrix} (\lambda+G)\nabla(\nabla\cdot\mathbf{u_{j-2}}) + G\nabla^2\mathbf{u_{j-2}} + C\nabla(\nabla\cdot\mathbf{w_{j-2}}) \\ C\nabla(\nabla\cdot\mathbf{u_{j-2}}) + M\nabla(\nabla\cdot\mathbf{w_{j-2}}) \end{bmatrix}. \end{aligned} \tag{33}$$

From (31), $\mathcal{L}$ has a nontrivial null space, which contains the vector $[\mathbf{u_0}, \mathbf{w_0}]^{\mathbf{T}}$. Because $\mathcal{L}$ is complex symmetric, the complex conjugate $[\mathbf{u_0^*}, \mathbf{w_0^*}]^{\mathbf{T}}$ is in the null space of the adjoint of $\mathcal{L}$. The Fredholm alternative then implies that $[\mathbf{u_0^*}, \mathbf{w_0^*}]^{\mathbf{T}}$ is orthogonal to the right-hand side of (32), i.e.,

$$\mathbf{u_0} \cdot \mathbf{R_1}(\nabla\phi, \mathbf{u_0}, \mathbf{w_0}) + \mathbf{w_0} \cdot \mathbf{R_2}(\nabla\phi, \mathbf{u_0}, \mathbf{w_0}) = 0. \tag{34}$$

Similarly, $[\mathbf{u_0^*}, \mathbf{w_0^*}]^{\mathbf{T}}$ is orthogonal to the right-hand side of (33) for all $j = 2, 3, 4, \ldots$.

Equation (31) reduces to the Biot theory dispersion relations if $\nabla\phi$ is identified as the slowness vector. This is because a sinusoidal wave is a special case of the form (27), when ϕ is linear in $\mathbf{x}$. Of course, wave fields that are much more complicated than sinusoidal can be constructed from the general WKB expansion given here. However, analogous to the classification of wave modes, we can classify three distinct types of waves, which, for large $|\tilde{\rho}|$, correspond to seismic S-waves, seismic P-waves, and the Biot slow wave. This is done in the next three sections.

5. S-waves. First consider transverse waves $\mathbf{u_0} = \mathbf{u_S}, \mathbf{w_0} = \mathbf{w_S}$, $\phi = \phi_S$, satisfying

$$\mathbf{u_S} \cdot \nabla\phi_S = \mathbf{w_S} \cdot \nabla\phi_S = \mathbf{0}. \tag{35}$$

Substitution of (35) into (31) gives

$$\left(\rho - G(\nabla\phi_S)^2\right) \mathbf{u_S} + \rho_f \mathbf{w_S} = \mathbf{0}, \tag{36}$$

$$\rho_f \mathbf{u_S} + \tilde{\rho} \mathbf{w_S} = \mathbf{0}. \tag{37}$$

From (36) and (37) we get the dispersion relation for transverse waves

$$\rho - G(\nabla\phi_S)^2 = \rho_f^2/\tilde{\rho}, \tag{38}$$

which for $|\tilde{\rho}|$ large can be written as

$$(\nabla\phi_S)^2 = \frac{1}{V_S^2} + O\left(\frac{1}{|\tilde{\rho}|}\right), \tag{39}$$

where

$$V_S = \sqrt{\frac{G}{\rho}} \tag{40}$$

is the elastic shear wave (seismic S-wave) speed. The eikonal equation [5, 10], (39) with the small terms neglected, identifies the phase as approximately that of an S-wave. The S-wave ray system is derived as the subcharacteristic curves of this equation, which in a homogeneous medium are straight lines in the directions of $\nabla\phi_S$.

From (39), corrections to ϕ_S are of order $O(1/|\tilde{\rho}|)$. Because of the exponential factor in (27) these terms are negligible only if $\omega/|\tilde{\rho}| << 1$. Thus, this condition is necessary for elasticity theory without significant modification, e.g., with attenuation, to be valid.

From (37)

$$\mathbf{w_S} = -\frac{\rho_f}{\tilde{\rho}} \mathbf{u_S}. \tag{41}$$

From (41) note that for an S-wave, the fluid displacement relative to the solid is small, of order $O(1/|\tilde{\rho}|)$. Thus the fluid motion does not depart substantially from the solid motion, as would be expected if the porous medium is to approximate an elastic solid.

To get a transport equation for the S-wave amplitudes, assume that $\mathbf{w_S}$ is small compared to $\mathbf{u_S}$, so that the second term in (34) may be neglected. Then simplifying, using that $\mathbf{u_S}$ is orthogonal to $\nabla\phi_S$ and $\mathbf{w_S}$ is small yields

$$\mathbf{u_S}\cdot\mathbf{R_1} = G\mathbf{u_S}\cdot\left\{2(\nabla\phi_S\cdot\nabla)\mathbf{u_S} + (\nabla\phi_S)^2\mathbf{u_S}\right\} = 0. \tag{42}$$

This equation reduces to the usual transport equation in which energy is conserved in a ray tube [10]:

$$\nabla\cdot\left((\mathbf{u_S})^2\nabla\phi_S\right) = 0. \tag{43}$$

To determine the direction of $\mathbf{u_S}$, note that $\mathbf{u_0'} = \nabla\phi_S\times\mathbf{u_S}, \mathbf{w_0'} = -(\rho_f/\tilde{\rho})\mathbf{u_0'}$ is also a solution of (31). That is, $[\nabla\phi_S\times\mathbf{u_S}, -(\rho_f/\tilde{\rho})\nabla\phi_S\times\mathbf{u_S}]^T$ is in the null space of $\mathcal{L}$, and so the complex conjugate of this vector is in the null space of the adjoint of $\mathcal{L}$. Therefore, the Fredholm alternative implies that $[(\nabla\phi_S\times\mathbf{u_S})^*, (-(\rho_f/\tilde{\rho})\nabla\phi_S\times\mathbf{u_S})^*]^T$ is orthogonal to $[\mathbf{R_1}(\nabla\phi_S, \mathbf{u_S}, \mathbf{w_S}), \mathbf{R_2}(\nabla\phi_S, \mathbf{u_S}, \mathbf{w_S})]^T$. To leading order,

$$(\nabla\phi_S\times\mathbf{u_S})\cdot\{(\nabla\phi_S\cdot\nabla)\mathbf{u_S}\} = 0. \tag{44}$$

Thus changes in $\mathbf{u_S}$ in the direction of $\nabla\phi_S$ remain in the plane determined by $\nabla\phi_S$ and $\mathbf{u_S}$. Since, in a homogeneous medium, the rays are straight lines in the direction of $\nabla\phi_\mathbf{S}$, the direction of $\mathbf{u_S}$ remains constant along a ray.

To summarize, we will use the following approximation for S-waves when $-i\omega\mathbf{u_S}$ is of order $O(1)$:

$$\begin{aligned}\mathbf{u} &= e^{i\omega\phi_S}\left[\mathbf{u_S} + O\left(\frac{1}{\omega|\tilde{\rho}|}\right) + O\left(\frac{1}{\omega^2}\right)\right],\\ \mathbf{w} &= e^{i\omega\phi_S}\left[-\frac{\rho_f}{\tilde{\rho}}\mathbf{u_S} + O\left(\frac{1}{\omega|\tilde{\rho}|^2}\right) + O\left(\frac{1}{\omega^2}\right)\right].\end{aligned} \tag{45}$$

Here $\mathbf{u_S}$, with $\mathbf{u_S}\cdot\nabla\phi_S = 0$, is determined by (43) and (44), and ϕ_S is determined, with error of order $O(1/|\tilde{\rho}|)$, by (39).

6. P-waves. Next, consider longitudinal waves

$$\mathbf{u_0} = u_L\nabla\phi, \quad \mathbf{w_0} = w_L\nabla\phi\,, \tag{46}$$

where u_L, w_L are scalars. Putting (46) into (31) gives

$$\left(\rho - (\lambda+2G)(\nabla\phi)^2\right)u_L + \left(\rho_f - C(\nabla\phi)^2\right)w_L = 0, \tag{47}$$

$$\left(\rho_f - C(\nabla\phi)^2\right)u_L + \left(\tilde{\rho} - M(\nabla\phi)^2\right)w_L = 0. \tag{48}$$

From (47) and (48) we obtain the longitudinal dispersion relation

$$\left(\rho_f - C(\nabla\phi)^2\right)^2 = \left(\tilde{\rho} - M(\nabla\phi)^2\right)\left(\rho - (\lambda+2G)(\nabla\phi)^2\right). \tag{49}$$

There are two roots of the quadratic equation (49) for $(\nabla\phi)^2$. One root ϕ_P is obtained asymptotically by a power series expansion in $(\tilde{\rho})^{-1}$:

$$(\nabla\phi_P)^2 = \frac{1}{V_P^2} + O\left(\frac{1}{|\tilde{\rho}|}\right), \tag{50}$$

where

$$V_P = \sqrt{\frac{(\lambda + 2G)}{\rho}} \tag{51}$$

is the elastic compressional wave (seismic P-wave) speed. Neglecting the small terms in the eikonal equation (50) identifies these waves as, approximately, seismic P-waves. For this root, the solution is written as

$$\mathbf{u_0} = u_P V_P \nabla\phi_P, \quad \mathbf{w_0} = w_P V_P \nabla\phi_P. \tag{52}$$

The constant factor V_P is inserted in (52) because $V_p\nabla\phi_P$ is a unit vector.

The ray system for P-waves are the subcharacteristic curves of equation (50), i.e., straight lines in the direction of $\nabla\phi_P$.

As for S-waves, the condition $\omega/|\tilde{\rho}| << 1$ must be assumed for accuracy of $\exp\{i\omega\phi_P\}$, when terms of order $O(1/|\tilde{\rho}|)$ are dropped in determining ϕ_P. Also, as for S-waves, the fluid motion does not depart substantially from that of the solid. This can be seen by combining (48), (52), and (50) to get

$$w_P \sim \frac{1}{\tilde{\rho}}\left(\frac{C}{V_P^2} - \rho_f\right) u_P. \tag{53}$$

To derive the transport equation, let w_P be small compared to u_P, and substitute into (34). Then use of (50) yields again the transport equation for energy conservation in a ray tube:

$$\nabla \cdot (u_P^2 \nabla\phi_P) = 0. \tag{54}$$

To summarize, we will use the following approximation for P-waves when $-i\omega u_P$ is of order $O(1)$:

$$\begin{aligned} \mathbf{u} &= e^{i\omega\phi_P}\left[u_P \frac{\nabla\phi_P}{|\nabla\phi_P|} + O\left(\frac{1}{\omega|\tilde{\rho}|}\right) + O\left(\frac{1}{\omega^2}\right)\right], \\ \mathbf{w} &= e^{i\omega\phi_P}\left[\frac{1}{\tilde{\rho}}\left(\frac{C}{V_P^2} - \rho_f\right) u_P \frac{\nabla\phi_P}{|\nabla\phi_P|} + O\left(\frac{1}{\omega|\tilde{\rho}|^2}\right) + O\left(\frac{1}{\omega^2}\right)\right]. \end{aligned} \tag{55}$$

Here u_P is determined by (54) and ϕ_P is determined, with error of order $O(1/|\tilde{\rho}|)$, by (50).

7. Biot slow waves. To obtain asymptotically the second root of (49) let $\phi = \phi_B$ be written as

$$\phi_B = \sqrt{\tilde{\rho}}\,\bar{\phi}_B, \tag{56}$$

where the square root is in the first quadrant. Then asymptotically

$$(\nabla\bar{\phi}_B)^2 = \frac{1}{\tilde{M}^2} + O\left(\frac{1}{|\tilde{\rho}|}\right), \tag{57}$$

where

$$\tilde{M} = \sqrt{M - \frac{C^2}{(\lambda + 2G)}} \tag{58}$$

is real. The eikonal equation (57), with (56), identifies this wave as the Biot slow wave, which is slow because its speed, $V_B = Re\{\tilde{M}/\sqrt{\tilde{\rho}}\}$, is small. It is a diffusive wave, which decays rapidly as it propagates away from where it originates. This is because the exponential factor in the WKB. expansion is

$$e^{i\omega\phi_B} = e^{\frac{(-1+i)}{\sqrt{2}}\omega\sqrt{|\tilde{\rho}|}\bar{\phi}_B}, \tag{59}$$

which is transcendentally small unless $|\omega\sqrt{|\tilde{\rho}|}\bar{\phi}_B|$ is $O(1)$.

If $\bar{\phi}_B$ is determined from (57) with the error terms dropped, then $\bar{\phi}_B$ is accurate up to an error of order $O(1/|\tilde{\rho}|)$. In this case corrections to the argument of the exponential in (59) have an error of $O(\omega/\sqrt{|\tilde{\rho}|})$, and so higher order terms in $\tilde{\rho}^{-1}$ might be significant depending on the size of this ratio. Nevertheless, in what follows, we will not need to consider corrections to $\bar{\phi}_B$, because the leading order term is sufficient to determine when the slow wave is transcendentally small such that it need not be calculated at all; and because we will use the exponential only where $\bar{\phi}_B \equiv 0$ to all orders in $\tilde{\rho}$.

For the phase $\bar{\phi}_B$ the solution is written as

$$\mathbf{u_0} = u_B\tilde{M}\nabla\bar{\phi}_B, \quad \mathbf{w_0} = w_B\tilde{M}\nabla\bar{\phi}_B. \tag{60}$$

The constant factor $\tilde{M}$ is inserted in (60) because $\tilde{M}\nabla\bar{\phi}_B$ is a unit vector.

From (47), (57), and (58) we obtain that, dropping terms of order $O\,(1/|\tilde{\rho}|)$,

$$w_B \sim -\frac{(\lambda+2G)}{C}u_B. \tag{61}$$

Note that w_B is not small compared to u_B unlike the corresponding expressions for P- and S-waves. This is consistent with the fact that these waves have no analogue in elasticity theory. Thus differences from elasticity theory can occur, but only locally near where a Biot slow wave originates. We will show in the next section that at interfaces excited by an EM wave, Biot slow waves can play an important role in determining other wave amplitudes, even though the slow waves themselves are undetectably small away from the interfaces.

To obtain the transport equation for the slow wave amplitudes, substitute (60) (61) into (34) and use (56) and (57). After simplification, the result is again conservation of energy in a ray tube:

$$\nabla\cdot(u_B^2\nabla\bar{\phi}_B) = 0. \tag{62}$$

The above equations determine the amplitudes u_B and w_B to an error of order $O(1/|\tilde{\rho}|)$.

In the next section, we will need to differentiate $\mathbf{u}$ and $\mathbf{w}$ in order to calculate the pressure and stress tensor from (6) and (5). A subtlety then arises in the case of the Biot slow wave, since differentiation of the exponential in (59) brings down a large factor of $\sqrt{\tilde{\rho}}$. In this case the higher order terms such as $\mathbf{u_1}, \mathbf{w_1}$ appear to be significant if $\sqrt{|\tilde{\rho}|}/\omega$ is not small. To obtain the form of these higher order terms in ω, but to leading order in $|\tilde{\rho}|$, substitute (56) into (31)–(33) and write the result in matrix form. Let

$$\mathcal{L}^{(0)} = \mathcal{L}^{(0)}(\nabla\bar{\phi}_B) = \begin{bmatrix} -G(\bar{\phi}_B)^2\mathbf{I} - (\lambda+G)\nabla\bar{\phi}_B\nabla\bar{\phi}_B^T & -C\nabla\bar{\phi}_B\nabla\bar{\phi}_B^T \\ -C\nabla\bar{\phi}_B\nabla\bar{\phi}_B^T & \mathbf{I} - M\bar{\phi}_B\nabla\bar{\phi}_B^T \end{bmatrix}. \tag{63}$$

Then the equations can be written as

$$\mathcal{L}^{(0)} \begin{bmatrix} \mathbf{u_0} \\ \mathbf{w_0} \end{bmatrix} = O\left(\frac{1}{|\tilde{\rho}|}\right), \tag{64}$$

and for $j = 1, 2, \ldots,$

$$\mathcal{L}^{(0)} \begin{bmatrix} \mathbf{u_j} \\ \mathbf{w_j} \end{bmatrix} = \frac{1}{\sqrt{\tilde{\rho}}} \begin{bmatrix} \mathbf{R_1}(\nabla\bar{\phi}_B, \mathbf{u_{j-1}}, \mathbf{w_{j-1}}) \\ \mathbf{R_2}(\nabla\bar{\phi}_B, \mathbf{u_{j-1}}, \mathbf{w_{j-1}}) \end{bmatrix} + O\left(\frac{1}{|\tilde{\rho}|}\right). \tag{65}$$

The solution of (64) and (65) may be expanded as a power series in $(\tilde{\rho})^{-1/2}$:

$$\begin{aligned} \mathbf{u_j} &= \sum_{m=0}^{\infty} (\tilde{\rho})^{(-m/2)} \mathbf{u_{j,m}}, \\ \mathbf{w_j} &= \sum_{m=0}^{\infty} (\tilde{\rho})^{(-m/2)} \mathbf{w_{j,m}}. \end{aligned} \tag{66}$$

Letting $|\tilde{\rho}| \to \infty$, the leading order terms, $\mathbf{u_{j,0}}, \mathbf{w_{j,0}}$, satisfy, for $j = 0, 1, 2, \ldots,$

$$\mathcal{L}^{(0)} \begin{bmatrix} \mathbf{u_{j,0}} \\ \mathbf{w_{j,0}} \end{bmatrix} = \mathbf{0}. \tag{67}$$

Consideration of the null space of $\mathcal{L}^{(0)}$ shows that for $j = 0, 1, 2, \ldots$

$$\mathbf{u_{j,0}} = u_{B,j}\tilde{M}\nabla\bar{\phi}_B, \quad \mathbf{w_{j,0}} = w_{B,j}\tilde{M}\nabla\bar{\phi}_B, \tag{68}$$

where

$$w_{B,j} = -\frac{(\lambda + 2G)}{C} u_{B,j}. \tag{69}$$

Finally, the Fredholm condition for the jth term is that the vector $[\nabla\bar{\phi}_B, -C^{-1}(\lambda+2G)\nabla\bar{\phi}_B]^{*T}$ is orthogonal to $[\mathbf{R_1}(\nabla\bar{\phi}_B, \mathbf{u_{j,0}}, \mathbf{w_{j,0}}), \mathbf{R_2}(\nabla\bar{\phi}_B, \mathbf{u_{j,0}}, \mathbf{w_{j,0}})]$. As in the case of $j = 0$, which was first done separately above, we derive the expression for energy conservation in a ray tube:

$$\nabla \cdot (u_{B,j}^2 \nabla\bar{\phi}_B) = 0. \tag{70}$$

Therefore the same relations (68), (69), and (70) govern the slow wave amplitudes to leading order in $\tilde{\rho}^{-1/2}$ to all orders $j = 0, 1, 2, \ldots$ in $(i\omega)^{-j}$. These are the same relations (60), (61), and (62) that were first derived for $j = 0$. However, corrections to $\bar{\phi}_B$ and to the $j = 0$ amplitude terms are of order $O(1/|\tilde{\rho}|)$, while corrections to the amplitudes for $j \geq 1$ are larger, of order $O(1/\sqrt{|\tilde{\rho}|})$.

To summarize, we will use the following approximation for Biot slow waves when $-i\omega u_B$ and $-i\omega u_{B,j}$ are of order $O(1)$:

(71)

$$\mathbf{u} \sim e^{i\omega\sqrt{\bar\rho}\bar\phi_B}\left[u_B\frac{\nabla\bar\phi_B}{|\nabla\bar\phi_B|}+O\left(\frac{1}{\omega|\tilde\rho|}\right)+\sum_{j=1}^{\infty}\left(\frac{u_{B,j}}{(i\omega)^j}\frac{\nabla\bar\phi_B}{|\nabla\bar\phi_B|}+O\left(\frac{1}{\omega^{j+1}\sqrt{|\tilde\rho|}}\right)\right)\right],$$

$$\mathbf{w} \sim -\frac{(\lambda+2G)}{C}e^{i\omega\sqrt{\bar\rho}\bar\phi_B}\left[u_B\frac{\nabla\bar\phi_B}{|\nabla\bar\phi_B|}+O\left(\frac{1}{\omega|\tilde\rho|}\right)\right.$$
$$\left.+\sum_{j=1}^{\infty}\left(\frac{u_{B,j}}{(i\omega)^j}\frac{\nabla\bar\phi_B}{|\nabla\bar\phi_B|}+O\left(\frac{1}{\omega^{j+1}\sqrt{|\tilde\rho|}}\right)\right)\right].$$

Here $u_B, u_{B,j}$ are determined by (62) and (70), and $\tilde{\phi_B}$ is determined, with error of order $O(1/|\tilde\rho|)$, by (57).

8. Initial amplitudes of the seismic waves. Combining (45), (55), and (71), a wavefield with the three types of waves is written as

(72)

$$\mathbf{u} \sim e^{i\omega\phi_S}\left[u_{S1}\mathbf{e_1}+u_{S2}\mathbf{e_2}+O\left(\frac{1}{\omega|\tilde\rho|}\right)+O\left(\frac{1}{\omega^2}\right)\right]$$
$$+\ e^{i\omega\phi_P}\left[u_P\frac{\nabla\phi_P}{|\nabla\phi_P|}+O\left(\frac{1}{\omega|\tilde\rho|}\right)+O\left(\frac{1}{\omega^2}\right)\right]$$
$$+\ e^{i\omega\sqrt{\bar\rho}\bar\phi_B}\left[u_B\frac{\nabla\bar\phi_B}{|\nabla\bar\phi_B|}+O\left(\frac{1}{\omega|\tilde\rho|}\right)+\sum_{j=1}^{\infty}\left(\frac{u_{B,j}}{(i\omega)^j}\frac{\nabla\bar\phi_B}{|\nabla\bar\phi_B|}+O\left(\frac{1}{\omega^{j+1}\sqrt{|\tilde\rho|}}\right)\right)\right],$$

$$\mathbf{w} \sim e^{i\omega\phi_S}\left[-\frac{\rho_f}{\tilde\rho}(u_{S1}\mathbf{e_1}+u_{S2}\mathbf{e_2})+O\left(\frac{1}{\omega|\tilde\rho|^2}\right)+O\left(\frac{1}{\omega^2}\right)\right]$$
$$+\ e^{i\omega\phi_P}\left[\frac{1}{\tilde\rho}\left(\frac{C}{V_P^2}-\rho_f\right)u_P\frac{\nabla\phi_P}{|\nabla\phi_P|}+O\left(\frac{1}{\omega|\tilde\rho|^2}\right)+O\left(\frac{1}{\omega^2}\right)\right]\quad-\frac{(\lambda+2G)}{C}$$
$$\times\, e^{i\omega\sqrt{\bar\rho}\bar\phi_B}\left[u_B\frac{\nabla\bar\phi_B}{|\nabla\bar\phi_B|}+O\left(\frac{1}{\omega|\tilde\rho|}\right)+\sum_{j=1}^{\infty}\left(\frac{u_{B,j}}{(i\omega)^j}\frac{\nabla\bar\phi_B}{|\nabla\bar\phi_B|}+O\left(\frac{1}{\omega^{j+1}\sqrt{|\tilde\rho|}}\right)\right)\right].$$

Here $\mathbf{e_1}$, $\mathbf{e_2}$ are orthogonal unit vectors, orthogonal to $\nabla\phi_S$, and u_{S1}, u_{S2}, u_P, u_B, $u_{B,j}$ are scalar wave amplitudes. As noted previously, each of these amplitudes is anticipated to be of order $O(1/\omega)$, because of the form of interface conditions (19), (20), (23), and (26).

In the expression for $\mathbf{w}$ in equation (72) the small terms of order $O(1/|\tilde\rho|)$ will dominate the exponentially small slow wave term, except when the slow wave phase $\bar\phi_B$ is near zero. However, as shown below, $\bar\phi_B = 0$ at an interface where waves originate, and it is then the slow wave term that dominates the expression for $\mathbf{w}$.

Consider an interface $\mathcal{S}$ separating two homogeneous media, and at each point of $\mathcal{S}$ let $\mathbf{n}$ be a unit normal pointing into the side $\mathcal{S}^+$, i.e., away from $\mathcal{S}^-$. We will show that the interface conditions of section 3 can be satisfied by introducing two waves, $\mathbf{u}^\pm, \mathbf{w}^\pm$, each of the form of (72), originating at $\mathcal{S}^\pm$ and subsequently propagating up and down into the two homogeneous media separated by $\mathcal{S}$.

First note that none of the interface conditions from section 3 have rapid spatial oscillations. Therefore all the phases must vanish on $\mathcal{S}$:

$$\phi_S^\pm = \phi_P^\pm = \bar{\phi}_B^\pm = 0 \quad \text{on}\, \mathcal{S}. \tag{73}$$

From (73), every interface is a wavefront for every type of wave. In particular, $\nabla\phi_S^\pm, \nabla\phi_P^\pm, \nabla\bar{\phi}_B^\pm$ are all orthogonal to $\mathcal{S}$. Since rays propagate in the direction of the phase gradients, it is apparent that all P-, S-, and Biot slow wave rays leave $\mathcal{S}$ in the orthogonal direction. That is, the rays for $\mathbf{u}^+, \mathbf{w}^+$ originating at $\mathcal{S}^+$ are in the direction $\mathbf{n}$, while those for $\mathbf{u}^-, \mathbf{w}^-$ originating at $\mathcal{S}^-$ are in the direction $-\mathbf{n}$. These observations are sufficient for determination of the phases to leading order in $\tilde{\rho}$, via conventional ray tracing.

Similarly, the transport (43), (54), and (62) are sufficient to determine the eight wave amplitudes, provided that their initial values $u_{S1}^\pm$, $u_{S2}^\pm$, $u_P^\pm$, $u_B^\pm$ are known on $\mathcal{S}^\pm$. We next determine those initial values by using the eight scalar interface conditions derived in section 3.

First, we obtain from (72) and (73) that to leading order

$$\mathbf{u}^\pm \sim u_{S1}^\pm \mathbf{e_1} + u_{S2}^\pm \mathbf{e_2} \pm u_P^\pm \mathbf{n} \pm u_B^\pm \mathbf{n} \quad \text{on}\, \mathcal{S}^\pm, \tag{74}$$

$$\mathbf{w}^\pm \cdot \mathbf{n} \sim \mp \frac{(\lambda^\pm + 2G^\pm)}{C^\pm} u_B^\pm \quad \text{on}\, \mathcal{S}^\pm, \tag{75}$$

since the slow wave dominates the expression for $\mathbf{w}$ on $\mathcal{S}$. In these equations λ^+ denotes the value of this parameter on $\mathcal{S}^+$, λ^- denotes the value of this parameter on $\mathcal{S}^-$, with a similar notation for the other parameters.

Next, differentiation of (72) yields

$$\nabla \cdot \mathbf{u}^\pm = \sqrt{\tilde{\rho}^\pm} \frac{(i\omega u_B^\pm)}{\tilde{M}^\pm} + \frac{\sqrt{\tilde{\rho}^\pm}}{\tilde{M}^\pm} \sum_{j=1}^{\infty} \frac{(i\omega u_{B,j}^\pm)}{(i\omega)^j} + \frac{(i\omega u_P^\pm)}{V_P} + o(1) \quad \text{on}\, \mathcal{S}^\pm,$$

$$\nabla \cdot \mathbf{w}^\pm = -\frac{\sqrt{\tilde{\rho}^\pm}}{\tilde{M}^\pm} \frac{(\lambda^\pm + 2G^\pm)}{C^\pm} \left[(i\omega u_B^\pm) + \sum_{j=1}^{\infty} \frac{(i\omega u_{B,j}^\pm)}{(i\omega)^j} \right] + o(1) \quad \text{on}\, \mathcal{S}^\pm. \tag{76}$$

Inserting (76) into (6) and use of (57) yields that to leading order

$$p^\pm \sim \sqrt{\tilde{\rho}^\pm} \tilde{M}^\pm \frac{(\lambda^\pm + 2G^\pm)}{C^\pm} (i\omega u_B^\pm) \quad \text{on}\, \mathcal{S}^\pm. \tag{77}$$

From (77) the Biot slow wave determines the fluid pressure on the boundary.

We next compute the normal stress on the interface. To do this, compute from (72),

(78)

$$\left(\nabla\mathbf{u}+\nabla\mathbf{u}^{\mathbf{T}}\right)^{\pm}=\pm\frac{(i\omega u_{S1}^{\pm})}{V_S^{\pm}}\left(\mathbf{e_1n^T}+\mathbf{ne_1^T}\right)\pm\frac{(i\omega u_{S2}^{\pm})}{V_S^{\pm}}\left(\mathbf{e_2n^T}+\mathbf{ne_2^T}\right)$$
$$+2\frac{(i\omega u_P^{\pm})}{V_P^{\pm}}\mathbf{nn^T}\quad+\quad2\frac{\sqrt{\tilde{\rho}^{\pm}}}{\tilde{M}^{\pm}}\left((i\omega u_B^{\pm})+\sum_{j=1}^{\infty}\frac{\left(i\omega u_{B,j}^{\pm}\right)}{(i\omega)^j}\right)\mathbf{nn^T}+o(1).$$

Again, the Biot slow wave term dominates this expression, because of the factor of $\sqrt{\tilde{\rho}^{\pm}}$. However, on substitution of (76) and (78) into (5) and taking the dot product with $\mathbf{n}$ it is found that the slow wave contributions cancel out to all orders in ω in the expression for the normal stress. Because of this remarkable cancellation, it is not necessary to make assumptions about the size of the ratio of large parameters $\sqrt{|\tilde{\rho}|}/\omega^j$. The final expression for the normal stress is

$$\tau^{\pm}\cdot\mathbf{n}=\frac{(\lambda^{\pm}+2G^{\pm})}{V_P^{\pm}}(i\omega u_P^{\pm})\mathbf{n}\pm\frac{G^{\pm}}{V_S^{\pm}}(i\omega u_{S1}^{\pm})\mathbf{e_1}\pm\frac{G^{\pm}}{V_S^{\pm}}(i\omega u_{S2}^{\pm})\mathbf{e_2}+o(1). \tag{79}$$

Now substitution of (74) into (19), (75) into (20), (77) into (23), and (79) into (26) yields eight scalar equations for the eight amplitudes $u_{S1}^{\pm},u_{S2}^{\pm},u_P^{\pm},u_B^{\pm}$. After some algebra, and discarding higher order terms in $(i\omega)^{-1}$ the following formulas are obtained. Let

$$\gamma_1=\frac{\sqrt{\tilde{\rho}^{+}}\tilde{M}^{+}(\lambda^{+}+2G^{+})C^{-}}{\sqrt{\tilde{\rho}^{-}}\tilde{M}^{-}(\lambda^{-}+2G^{-})C^{+}}, \tag{80}$$

$$\gamma_2=\left[1+\frac{(\lambda^{+}+2G^{+})V_P^{-}}{(\lambda^{-}+2G^{-})V_P^{+}}\right]^{-1}, \tag{81}$$

$$\gamma_3=\left[1+\frac{V_S^{-}G^{+}}{V_S^{+}G^{-}}\right]^{-1}. \tag{82}$$

Then

$$-i\omega u_B^{+}=\left[\frac{(\lambda^{+}+2G^{+})}{C^{+}}+\frac{(\lambda^{-}+2G^{-})}{C^{-}}\gamma_1\right]^{-1}[\,|\,L\mathbf{E}\cdot\mathbf{n}\,|\,], \tag{83}$$

$$-i\omega u_P^{+}=\gamma_2\left[\left|\left(\frac{\rho_f}{\rho}\right)L\mathbf{E}\cdot\mathbf{n}\right|\right]-\gamma_2(1+\gamma_1)(-i\omega u_B^{+}), \tag{84}$$

$$-i\omega u_{Sj}^{+}=\gamma_3\left[\left|\left(\frac{\rho_f}{\rho}\right)L\mathbf{E}\cdot\mathbf{e_j}\right|\right]\quad\text{for}\quad j=1,2. \tag{85}$$

The corresponding formulas for $u_B^{-},u_P^{-},u_{Sj}^{-}$ may be obtained from those above by replacing $\mathbf{n}$ by $-\mathbf{n}$, and reversing the roles of + and –. Alternatively, the following

formulas may be used:

$$
\begin{aligned}
u_B^- &= \gamma_1 u_B^+, \\
u_P^- &= \frac{V_P^-}{V_P^+}\frac{(\lambda^+ + 2G^+)}{(\lambda^- + 2G^-)} u_P^+, \\
u_{Sj}^- &= -\frac{V_S^-}{V_S^+}\frac{G^+}{G^-} u_{Sj}^+ \quad \text{for} \quad j = 1, 2.
\end{aligned}
\tag{86}
$$

9. Comparison with an exact solution. To assess the accuracy of the asymptotic solution, we compared it to the results of a computer program designed to solve electroseismic problems in piecewise homogeneous media that are plane layered. For such media, where the material properties vary only with the depth dimension, special methods exist which are based on two-dimensional Fourier transformation of the lateral spatial dimensions. Consequently, plane layered media are the only class of piecewise homogeneous problems in electroseismic prospecting for which solutions that are exact, to within roundoff error and the numerical accuracy of the Fourier transforms, can be computed. One example of such a computer code for layered media is described in [8]. Details of the computer code used here will be described in a separate publication.

For simplicity, we ignored the earth-air boundary, considering a medium consisting of just two homogeneous half spaces joined at an interface at depth D. The source is a single, positive 1 amp point electrode at the origin $x = y = z = 0$, with the corresponding negative electrode at ∞. For comparison, conversions at the EM source were removed. For a negative electrode at a finite location on the surface $z = 0$, the solution computed here is translated spatially to the location of the electrode and the sign reversed. If both electrodes are at finite locations, i.e., the source is a bipole, then the solutions for the two electrodes are summed. Finally, from linearity of the equations, the results can be scaled for any current.

We computed a number of cases that fell well within the regime for which seismic ray tracing is valid. Figures 1 and 2 show the predicted geophone response, that is, the vertical velocity of the solid caused by seismic waves converted from EM at the interface, for one of these cases. Plotted are the amplitude vs. offset and phase vs. offset for geophones located on the surface $z = 0$ at a distance from the source corresponding to the offset coordinate. Note that the phase plotted here is just the phase of the complex number representing the geophone response at a point, not one of the WKB phases, as defined in previous sections. For many of the parameter values we investigated there were no discernible differences between the asymptotic solution and that of the layer code when the results were displayed graphically. However, for the example plotted in Figures 1 and 2 the amplitude vs. offset and phase vs. offset plots do show some small differences near zero offset. Still, the comparison is excellent, as would be expected if we were testing the accuracy of ray tracing for this example.

The parameter values used for Figures 1 and 2 were as follows: $D = 1000$ m; frequency is 30 Hz; $\sigma = 0.1\,(\text{ohm-m})^{-1}$ in the upper layer and $\sigma = 0.001\ (\text{ohm-m})^{-1}$ in the lower layer; $\epsilon = 0$ throughout, i.e., displacement currents have been neglected, as is common in geoelectric prospecting [20]; $\mu = 4\pi \times 10^{-7}\,\text{H/m}$ throughout, i.e., the magnetic permeability of free space; $\eta = 1.0 \times 10^{-3}\,\text{Pa-sec}$ in both layers; $\kappa = 1.0 \times 10^{-16}\,\text{m}^2$ in the upper layer and $\kappa = 1.0 \times 10^{-13}\,\text{m}^2$ in the lower layer; $L = 1.0 \times 10^{-9}\,\text{m}^2/\text{V-sec}$ in both layers.

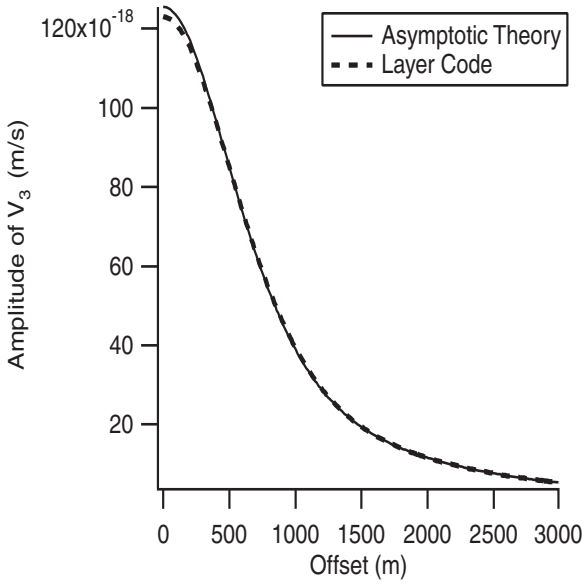


FIG. 1. *Vertical velocity, amplitude vs. offset.*

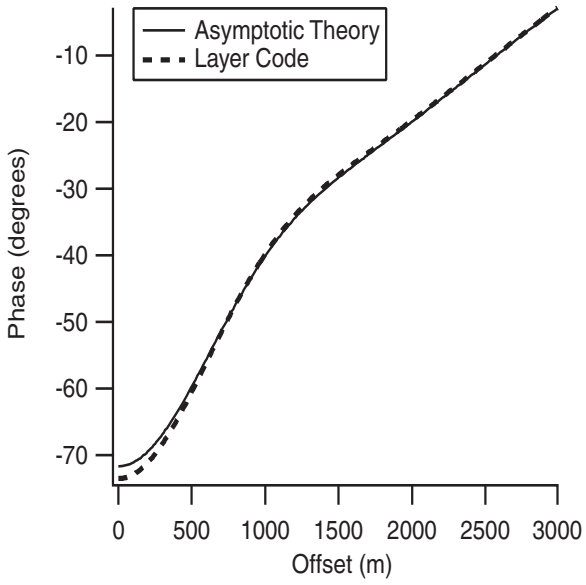


FIG. 2. *Vertical velocity, phase vs. offset.*

For both layers the solid density was taken as $\rho_s = 2650\,\mathrm{kg/m^2}$ and the porosity as 0.3. The fluid density was taken as $\rho_f = 980\,\mathrm{kg/m^2}$ in the upper layer and $\rho_f = 784\,\mathrm{kg/m^2}$. The density in each layer was then calculated as the porosity weighted average of the fluid and solid densities.

The Biot theory parameters were calculated using the theory of Pride, Gangi, and Morgan [15] and reasonable numbers for the variables. Values were $V_P = 1800\,\mathrm{m/sec}$

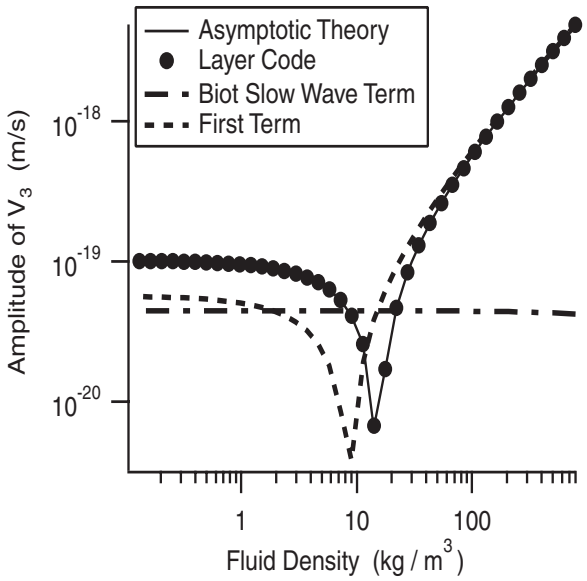


FIG. 3. *Vertical velocity at zero offset. Amplitude vs. density of reservoir fluid.*

and $V_S = 1080\,\mathrm{m/sec}$ for both layers. Then calculated values in $\mathrm{N/m^2}$ were $\lambda = 1,949,572,601$, $G = 2,506,593,700$, $M = 22,436,559$, and $C = 21,603,901$ in the upper layer and $\lambda = 1,896,229,249$, $G = 2,438,009,381$, $M = 22,436,554$, and $C = 21,626,820$ in the lower layer.

Comparison of the asymptotic theory with the geophone response for plane layered media is really only a test of the P-wave amplitude theory. This is because the Biot slow wave does not propagate back to the surface; moreover, since all rays are orthogonal to the plane of the layering, shear waves cannot contribute to the vertical velocity of the solid. However, from the derivation of the theory, it is expected that the S-wave accuracy will be comparable to that of the P-waves.

In Figure 3 the amplitude of the vertical velocity at zero offset is plotted as a function of fluid density in the lower layer, with excellent agreement of the asymptotic theory and the layer code over four orders of magnitude in fluid density. From (84) the vertical velocity, which is determined by the P-waves, is the sum of two terms: the first term in (84) and the second, i.e., the Biot slow wave term. As seen in Figure 3, for small fluid densities both terms contribute about equally, while for large fluid densities the first term dominates, and the slow wave term is unimportant. There is, however, a narrow range of fluid densities for which the Biot slow wave determines the response, and the first term is unimportant. Altogether, both terms are needed to cover the full range of possibilities.

10. Conclusions. As stated in section 1, the theory developed here can be used to link conventional EM modeling software with conventional seismic ray tracing software to model the electroseismic response of three-dimensional subsurface structures in the earth. The theory can also be used to obtain simple estimates of the electroseismic response if the electric field can be estimated at depth; then use can be made of the fact that the tangential component of the electric field is continuous across the interface, while the normal components on both sides of the interface are in inverse

proportion to the conductivities. For example, since conductivities of a hydrocarbon reservoir are often orders of magnitude less than that of other kinds of rock, good P-wave responses are to be expected from reservoir boundaries, because of the large discontinuities in the normal electric field there.

In addition, the amplitude formulas (80)–(85) can be used to estimate other parametric dependencies of the electroseismic response. For example, note from (84) and (85) that all permeability dependencies in both the P- and S-waves are contained in the Biot slow wave term, through the dependence of γ_1 on $\tilde{\rho}$. Thus there is no permeability information in the S-waves, and the P-waves will also lack permeability information in cases where the slow wave term is unimportant compared to the other term in (84). In general, determination of permeability using the type of electroseismic prospecting described here is expected to be problematic. This contrasts with the permeability estimates obtained from EM conversions in the near field of a Stonely wave, as described in [12].

The theory derived here is expected to be of about the same order of accuracy as seismic ray theory. It also suffers from naive ray theory's defects, which include, for instance, edge diffraction. While this phenomenon must surely occur in electroseismics, it and other diffraction phenomena have not been included in the present theory.

Acknowledgment. I would like to thank Minyao Zhou for his help with the numerical comparisons of section 9, and for many helpful conversations.

REFERENCES

[1] M. A. Biot, *Theory of propagation of elastic waves in a fluid-saturated solid* I. *Low frequency range*, J. Acoustic Soc. Am., 28 (1956), pp. 168–178.

[2] M. A. Biot, *Theory of propagation of elastic waves in a fluid-saturated solid* II. *Higher frequency range*, J. Acoustic Soc. Am., 28 (1956), pp. 179–191.

[3] R. Burridge and J. Keller, *Poroelasticity equations derived from microstructure*, J. Acoustic Soc. Am., 70 (1981), pp. 1140–1146.

[4] K. E. Butler, R. D. Russell, A. W. Kepic, and M. Maxwell, *Measurement of the seismoelectric response from a shallow boundary*, Geophysics, 61 (1996), pp. 1769–1778.

[5] V. Cerveny, I. A. Molotkov, and I. Psencik, *The Ray Method in Seismology*, Univerzita Karlova, Praha, 1977.

[6] M. B. Dobrin, *Introduction to Geophysical Prospecting*, 4th ed., McGraw-Hill, New York, 1988.

[7] Ya. Frenkel, *On the theory of seismic and seismoelectric phenomena in moist soil*, J. Physics, 8 (1944), pp. 230–241.

[8] M. W. Haartsen and S. R. Pride, *Electroseismic waves from point sources in layered media*, J. Geophys. Res., 102 (1997), pp. 24745–24796.

[9] O. Kelder and D. M. J. Smeulders, *Observation of the Biot slow wave in water-saturated Nivelsteiner sandstone*, Geophysics, 62 (1997), pp. 1794–1796.

[10] J. B. Keller and R. M. Lewis, *Asymptotic methods for partial differential equations: The reduced wave equation and Maxwell's equations*, in Surveys in Applied Mathematics, J. B. Keller, G. Papanicolaou, and D. McLaughlin, eds. Plenum, NY, 1995, pp. 1–82.

[11] O. V. Mikhailov, M. W. Haartsen, and N. Toksoz, *Electroseismic investigation of the shallow subsurface: Field measurements and numerical modeling*, Geophysics, 62 (1997), pp. 97–105.

[12] O. V. Mikhailov, J. Queen, and N. Toksoz, *Using borehole electroseismic measurements to detect and characterize fractured (permeable) zones*, Geophysics, 65 (2000), pp. 1098–1112.

[13] D. B. Pengra, S. X. Li, and P.-z. Wong, *Determination of rock properties by low-frequency AC electrokinetics*, J. Geophys. Res., 104 (1999), pp. 29485–29508.

[14] S. R. Pride, *Governing equations for the coupled electromagnetics and acoustics of porous media*, Phys. Rev. B, 50 (1994), pp. 15678–15696.

[15] S. R. Pride, A. F. Gangi, and F. D. Morgan, *Deriving the equations of motion for porous isotropic media*, J. Acoustic Soc. Am., 6 (1992), pp. 3278–3290.
[16] S. R. Pride and S. Garambois, *The role of Biot slow waves in electroseismic wave phenomena*, J. Acoustic Soc. Am., 111 (2002), pp. 697–706.
[17] S. R. Pride and M. W. Haartsen, *Electroseismic wave properties*, J. Acoustic Soc. Am., 100 (1996), pp. 1301–1315.
[18] A. H. Thompson and G. A. Gist, *Geophysical applications of electrokinetic conversion*, Leading Edge, 12 (1993), pp. 1169–1173.
[19] R. R. Thompson, *The seismic electric effect*, Geophysics, 1 (1936), pp. 327–335.
[20] M. S. Zhdanov and G. V. Keller, *The Geoelectrical Methods in Geophysical Exploration*, Methods in Geochemistry and Geophysics 31, Elsevier, New York, 1994.

SIAM J. APPL. MATH.
Vol. 65, No. 5, pp. 1463–1484

CURRENT-VOLTAGE RELATIONS FOR ELECTROCHEMICAL THIN FILMS*

MARTIN Z. BAZANT†, KEVIN T. CHU†, AND B. J. BAYLY‡

Abstract. The DC response of an electrochemical thin film, such as the separator in a microbattery, is analyzed by solving the Poisson–Nernst–Planck equations, subject to boundary conditions appropriate for an electrolytic/galvanic cell. The model system consists of a binary electrolyte between parallel-plate electrodes, each possessing a compact Stern layer, which mediates Faradaic reactions with nonlinear Butler–Volmer kinetics. Analytical results are obtained by matched asymptotic expansions in the limit of thin double layers and compared with full numerical solutions. The analysis shows that (i) decreasing the system size relative to the Debye screening length decreases the voltage of the cell and allows currents higher than the classical diffusion-limited current; (ii) finite reaction rates lead to the important possibility of a reaction-limited current; (iii) the Stern-layer capacitance is critical for allowing the cell to achieve currents above the reaction-limited current; and (iv) all polarographic (current-voltage) curves tend to the same limit as reaction kinetics become fast. Dimensional analysis, however, shows that "fast" reactions tend to become "slow" with decreasing system size, so the nonlinear effects of surface polarization may dominate the DC response of thin films.

Key words. Poisson–Nernst–Planck equations, electrochemical systems, thin films, polarographic curves, Butler–Volmer reaction kinetics, Stern layer, surface capacitance

AMS subject classifications. 34B08, 34B16, 34B60, 34E05, 80A30

DOI. 10.1137/040609938

Introduction. Microelectrochemical systems pose interesting problems for applied mathematics because traditional "macroscopic" approximations of electroneutrality and thermal equilibrium [1], which make the classical transport equations more tractable [2], break down at small scales, approaching the Debye screening length. Of course, the relative importance of surface phenomena also increases with miniaturization. Microelectrochemical systems of current interest include ion channels in biological membranes [3, 4, 5] and thin-film batteries [6, 7, 8, 9, 10], which could revolutionize the design of modern electronics with distributed on-chip power sources. In the latter context, the internal resistance of the battery is related to the nonlinear current-voltage characteristics of the separator, consisting of a thin-film electrolyte (solid, liquid, or gel) sandwiched between flat electrodes and interfacial layers where Faradaic electron-transfer reactions occur [11]. Under such conditions, the internal resistance is unlikely to be simply constant, as is usually assumed.

Motivated by the application to thin-film batteries, here we revisit the classical problem of steady conduction between parallel, flat electrodes, studied by Nernst [12] and Brunner [13, 14] a century ago. As in subsequent studies of liquid [15, 16] and solid [17, 18] electrolytes, we do not make Nernst's assumption of bulk electroneutrality and work instead with the Poisson–Nernst–Planck (PNP) equations, allowing

*Received by the editors June 10, 2004; accepted for publication (in revised form) November 1, 2004; published electronically May 12, 2005. This work was supported in part by the MRSEC program of the National Science Foundation under award DMR 02-13282 and in part by the Department of Energy through the Computational Science Graduate Fellowship (CSGF) program provided under grant DE-FG02-97ER25308.

http://www.siam.org/journals/siap/65-5/60993.html

†Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139 (bazant@mit.edu, kchu@mit.edu).

‡Department of Mathematics, University of Arizona, Tucson, AZ 85271 (bjb@math.arizona.edu).

for diffuse charge in solution [1, 2]. What distinguishes our analysis from previous work on current-voltage relations (or "polarographic curves") is the use of more realistic nonlinear boundary conditions describing (i) Butler–Volmer reaction kinetics and (ii) the surface capacitance of the compact Stern layer, as in the recent paper of Bonnefont, Argoul, and Bazant [19]. Such boundary conditions, although complicating mathematical analysis, generally cannot be ignored in microelectrochemical cells, where interfaces play a crucial role. Diffuse-charge dynamics, which can be important for high-power applications, also complicates analysis [20], but in most cases it is reasonable to assume that electrochemical thin films are in a steady state, due to the short distances for electrodiffusion.

We stress, however, that steady state does not imply thermal equilibrium when the system sustains a current, driven by an applied voltage. In this paper, we focus on applied voltages small enough to justify the standard boundary-layer analysis of the PNP equations, which yields charge densities in thermal equilibrium at leading order in the limit of thin double layers [21]. In a companion paper [22], we extend the analysis to larger voltages, at [23] and above [24] the Nernst's diffusion-limited current, and show how more realistic boundary conditions affect diffuse charge, far from thermal equilibrium. In both cases, we obtain novel formulae for polarographic curves by asymptotic analysis in the limit of thin double layers, which we compare with numerical solutions, and we focus on a variety of dimensionless physical parameters: the reaction-rate constants (scaled to the typical diffusive flux) and the ratios of the Stern length to the Debye length to the electrode separation.

1. Mathematical model. Let us consider uniform conduction through a dilute, binary electrolyte between parallel-plate electrodes separated by a distance, L. Our goal is to determine the steady-state response of the electrochemical cell to either an applied voltage, V, or an applied current, I. Specifically, we seek the electric potential $\Phi(X)$ and the concentrations $C_+(X)$ and $C_-(X)$ of cations and anions in the region $0 \le X \le L$. The Faradaic current is driven by redox reactions occurring at the electrodes, and we neglect any other chemical reactions, such as dissociation/recombination in the bulk solution or hydrogen production. Since we do not assume electroneutrality, the region of integration extends to the point where the continuum approximation breaks down near each electrode, roughly a few molecules away. In other words, our integration region includes the "diffuse part" but not the "compact part" of the double layer [1, 11, 25, 26].

1.1. Transport equations. In the context of dilute solution theory [1, 2], the governing equations for this situation are the steady PNP equations:

$$\frac{d}{dX}\left(D_+\frac{dC_+}{dX} + \mu_+ z_+ F C_+ \frac{d\Phi}{dX}\right) = 0, \tag{1}$$

$$\frac{d}{dX}\left(D_-\frac{dC_-}{dX} + \mu_- z_- F C_- \frac{d\Phi}{dX}\right) = 0, \tag{2}$$

$$-\frac{d}{dX}\left(\epsilon_s \frac{d\Phi}{dX}\right) = (z_+ C_+ - z_- C_-)F, \tag{3}$$

where F is Faraday's constant (a mole of charge), $z_\pm$, $\mu_\pm$, and $D_\pm$ are the charge numbers, mobilities, and diffusivities of each ionic species, respectively, and ϵ_s is the permittivity of the solvent, all taken to be constant in the limit of infinite dilution. The first two equations set divergences of the ionic fluxes to zero in order to maintain the steady state, and the third is Poisson's equation relating the electric potential to

the charge density. In each flux expression, the first term represents diffusion and the second electromigration. The Einstein relation, $\mu_\pm = D_\pm/RT$, relates mobilities and diffusion coefficients via the absolute temperature, T, and the universal gas constant, R.

Due to the potentially large electric fields in thin films, as in interfacial double layers, these classical approximations could break down [1, 25, 26]. For example, the polarization of solvent molecules in large electric fields can lower the solvent dielectric permittivity by an order of magnitude, while also affecting diffusivity and mobility. The solution can also become locally so concentrated or depleted of certain ions as to make finite sizes and/or interactions (and thus ionic activities) important. In spite of these concerns, however, it is reasonable to analyze the PNP equations before considering more complicated transport models, especially because our focus is on the effect of boundary conditions.

1.2. Electrode boundary conditions. Although PNP equations constitute a well-understood and widely accepted approximation, appropriate boundary conditions for them are not so clear, and drastic approximations, such as constant concentration, potential or surface charge (or zeta potential), are usually made, largely out of mathematical convenience. On the other hand, in the context of electric circuit models for electrochemical cells [20, 27, 28], much effort has been made to describe the nonlinear response of the electrode-electrolyte interface, while describing the bulk solution as a simple circuit element, such as a resistor. Here, we formulate general boundary conditions based on classical models of the double layer [11, 25, 26], with a unified description of ion transport by the PNP equations.

Our first pair of boundary conditions sets the normal anion flux to zero at each electrode,

$$D_- \frac{dC_-}{dX}(0) + \mu_- z_- F C_-(0) \frac{d\Phi}{dX}(0) = 0, \tag{4}$$

$$D_- \frac{dC_-}{dX}(L) + \mu_- z_- F C_-(L) \frac{d\Phi}{dX}(L) = 0, \tag{5}$$

on the assumption that anions do not specifically adsorb onto the surfaces, which holds for many anions at typical metal surfaces (e.g., SO_4^{-2}, OH^-, F^-). The second pair relates the normal cation flux to the net deposition (or dissolution) flux, or reaction-rate density, $R(C_+, \Delta\Phi_S)$, which in the dilute limit is assumed to depend only on cation concentration and potential drop, $\Delta\Phi_S$, across the compact part of the double layer, originally proposed by Stern [29]. Following the convention in electrochemistry, we take $\Delta\Phi_S$ to be the potential of the electrode surface measured relative to the solution. The reference potential is chosen so that the cathode, located at $X = 0$, is at zero potential and the anode, located at $X = 1$, is at the applied cell voltage V. Therefore, we have the following two boundary conditions:

$$D_+ \frac{dC_+}{dX}(0) + \mu_+ z_+ F C_+(0) \frac{d\Phi}{dX}(0) = R\left(C_+(0), \Phi(0)\right), \tag{6}$$

$$-D_+ \frac{dC_+}{dX}(L) - \mu_+ z_+ F C_+(L) \frac{d\Phi}{dX}(L) = R\left(C_+(L), \Phi(L) - V\right). \tag{7}$$

For electrodes, it is typical to assume a balance of forward (deposition) and backward (dissolution) reaction rates biased by the Stern voltage with an Arrhenius temperature dependence,

$$R(C_+, \Delta\Phi_S) = K_c C_+ \exp\left(\frac{-\alpha_c z F \Delta\Phi_S}{RT}\right) - K_a C_M \exp\left(\frac{\alpha_a z F \Delta\Phi_S}{RT}\right), \tag{8}$$

where C_M is the (constant) density of electrode metal and K_c and K_a are rate constants for the cathodic and anodic reactions [1]. (Expressing the reaction rate in terms of the surface overpotential, $\eta_S = \Delta\Phi_S - \Delta\Phi_S^{eq}$, where $\Delta\Phi_S^{eq}$ is the Stern-layer voltage in the absence of current, $R = 0$, yields the more common form of the Butler–Volmer equation [1, 26].) The Stern-layer voltage contributes $-zF\Delta\Phi_S$ to the activation energy barriers multiplied by transfer coefficients α_c and α_a for the cathodic and anodic reactions, respectively, where $\alpha_c \approx \alpha_a \approx \frac{1}{2}$, for single electron transfer reactions [1, 26, 30].

Following Frumkin [31], we apply (8) just outside the Stern layer, in contrast to "macroscopic" models which postulate the Butler–Volmer equation as a purely empirical description of reactions between the electrode surface and the electrically neutral bulk solution. Physically, the Frumkin approach makes more sense since the activation energy barrier described by the Butler–Volmer equation actually exists at the atomic scale in the Stern layer, not across the entire "interface" including diffuse charge in solution. We are not aware of any prior analysis with the full, nonlinear Butler–Volmer equation as a boundary condition on the PNP equations other than that of Bonnefont, Argoul, and Bazant [19]. Earlier analyses by Chang and Jaffé [15], Jaffé and LeMay [16], and Itskovich, Kornyshev, and Vorotyntsev [17] also include electrode reactions, but only for small perturbations around equilibrium.

The final pair of boundary conditions determines the electric potential by specifying the voltage drop across the Stern layer in terms of the local electric field and concentrations,

$$0 - \Phi(0) = \Delta\Phi_S\left(\frac{d\Phi}{dX}(0), C_+(0), C_-(0)\right), \tag{9}$$

$$V - \Phi(L) = \Delta\Phi_S\left(-\frac{d\Phi}{dX}(L), C_+(L), C_-(L)\right). \tag{10}$$

In macroscopic electrochemistry, these boundary conditions are usually replaced by the assumption of electroneutrality (which eliminates the need to solve Poisson's equation) or by simple Dirichlet boundary conditions on the potential [1]. In colloidal science, they are likewise replaced by simple boundary conditions of constant surface charge (or zeta potential) [32, 33]. Here, we incorporate more realistic properties of the interface as follows: Neglecting the specific adsorption of anions, the Stern layer acts as a nonlinear capacitor in series with the diffuse layer. Grahame's celebrated electrocapillary measurements [34, 35] suggest that (i) the Stern-layer capacitance, C_S, is roughly independent of concentration, depending mainly on the (variable) total charge, σ,

$$\frac{d(\Delta\Phi_S)}{d\sigma} = \frac{1}{C_S(\sigma)}, \tag{11}$$

and (ii) dilute solution theory accurately describes the capacitance, $C_D(\sigma, C_+)$, of the diffuse layer, at least when the charge and current are small enough to be well described by the Poisson–Boltzmann theory (as derived below). Using Gauss's law, the surface charge density can be expressed in terms of the normal electric field, $\sigma = -\epsilon_S d\Phi/dX$, where ϵ_S is an effective permittivity of the compact layer. Therefore,

integrating (11), Grahame's model corresponds to the assumption

$$\Delta\Phi_S = \int_0^{-\epsilon_S d\Phi/dx} \frac{d\sigma}{C_S(\sigma)}, \tag{12}$$

which determines how the voltage across the compact layer (relative to the point of zero charge for which $\Delta\Phi_S = 0$) varies as the two capacitors become charged. The function, $C_S(\sigma)$, should be fit to experimental or theoretical electrocapillary curves at large concentrations (since $1/C_{total} = 1/C_D + 1/C_S \approx 1/C_S$ in that case).

The simplest model that captures this interplay between the compact and diffuse layers is the Stern model [11, 26], which assumes the capacitance of the compact layer, C_S, to be constant [29]. While more complicated models for the compact layer have been proposed [36, 37, 25], the Stern model suffices for our purposes, because it allows us to describe surface capacitance easily in the context of our model of Faradaic reactions. Following Itskovich, Kornyshev, and Vorotyntsev [17] and Bonnefont, Argoul, and Bazant [19], let us introduce an effective width, λ_S, for the compact layer, $\lambda_S = \epsilon_S/C_S$, so that (12) reduces to a linear extrapolation of the potential across the compact layer, $-\Delta\Phi_S = \lambda_S d\Phi/dx$. Substituting this expression into (9) and (10) yields two Robin boundary conditions,

$$\Phi(0) - \lambda_S \frac{d\Phi}{dX}(0) = 0, \tag{13}$$

$$\Phi(L) + \lambda_S \frac{d\Phi}{dX}(L) = V, \tag{14}$$

completing a set of six boundary conditions for our three second-order differential equations. Physically, the Stern layer, as an effective solvation shell for the electrode, is only a few molecules wide, so it is best to think of λ_S as simply a measure of the capacitance of the Stern layer. More generally, the same boundary condition could also describe a thin dielectric layer on the electrode [38, 39, 40], e.g., arising from surface contamination or a passivating monolayer.

Note that since the anion flux is zero, the current passing through the cell is proportional to the cation flux (everywhere in the cell, since it is constant),

$$I = z_+ FA \left(D_+ \frac{dC_+}{dX} + \mu_+ z_+ F C_+ \frac{d\Phi}{dX} \right), \tag{15}$$

where A is the electrode area and a current flow towards the cathode ($x = 0$) is taken to be positive. Under potentiostatic conditions, the cell voltage V is given, and the steady-state polarization curve $I(V)$ is determined by solving the equations. Conversely, under galvanostatic conditions, I is fixed, and we solve for $V(I)$.

1.3. An integral constraint. As formulated above, the boundary value problem is not well posed. Since the anion flux is constant throughout the cell according to (2), the two anion flux boundary conditions are degenerate, leaving one constant of integration undetermined. This is not surprising, as we have omitted one crucial physical parameter: the total number of anions. More precisely, because anions do not react, we must specify their total number, which remains constant as the steady state is reached. This corresponds to the constraint

$$\frac{1}{L} \int_0^L C_-(X) dX = C_{ref}, \tag{16}$$

where C_{ref} is the initial concentration of anions.

Note that the total number of cations (and hence the total charge) is not known a priori because the removal of cations at the cathode and the injection of cations at the anode may significantly alter their total number. This may seem counterintuitive since we are accustomed to assuming that we know the total cation concentration at all times based on the original molarity of the solution, but this "macroscopic" thinking does not apply when the physics at the microscopic level are explicitly being studied (e.g., diffuse charge layers or microelectrochemical systems). Mathematically, the reaction boundary conditions at the electrodes (6) and (7) are sufficient to determine the total cation concentration (and total charge), as long as the total anion concentration is specified.

1.4. Dimensionless formulation. To facilitate our analysis, we formulate the problem in dimensionless form. For simplicity we also assume that the electrolyte is symmetric, $z_+ = -z_- \equiv z$, which does not qualitatively affect any of our conclusions as long as $z_+/|z_-|$ is not too different from 1 (which holds for most simple, aqueous electrolytes). Scaling the basic variables

$$x \equiv X/L, \qquad c_\pm(x) \equiv C_\pm(xL)/C_{ref}, \qquad \phi(x) \equiv \frac{zF\Phi(xL)}{RT}, \tag{17}$$

equations (1)–(3) become

$$\frac{d^2c_+}{dx^2} + \frac{d}{dx}\left(c_+\frac{d\phi}{dx}\right) = 0, \tag{18}$$

$$\frac{d^2c_-}{dx^2} - \frac{d}{dx}\left(c_-\frac{d\phi}{dx}\right) = 0, \tag{19}$$

$$-\epsilon^2\frac{d^2\phi}{dx^2} = \frac{1}{2}(c_+ - c_-), \tag{20}$$

where $\epsilon \equiv \lambda_D/L$ is the ratio of the Debye screening length $\lambda_D \equiv \sqrt{\frac{\epsilon_s RT}{2z^2F^2C_{ref}}}$ to the distance between electrodes. The Debye length is typically on the order of nanometers, so ϵ is always extremely small for macroscopic electrochemical cells. This situation changes, however, as L or C_{ref} is decreased, and in the case of nanoelectrochemical systems ϵ could be as large as 10.

The two flux equations are easily integrated, using (4) to evaluate one constant and leaving the other constant expressed in terms of the current via (15),

$$\frac{dc_+}{dx} + c_+\frac{d\phi}{dx} = 4j, \tag{21}$$

$$\frac{dc_-}{dx} - c_-\frac{d\phi}{dx} = 0, \tag{22}$$

where we have defined a dimensionless current density, $j \equiv I/I_d$, scaled to the Nernst's diffusion-limited current density (see section 2.1), $I_d \equiv 4zFD_+C_{ref}A/L$. Since diffuse charge is of primary interest here, it is convenient to introduce

$$c = \frac{1}{2}(c_+ + c_-) \quad \text{and} \quad \rho = \frac{1}{2}(c_+ - c_-), \tag{23}$$

the average concentration of ions and (half) the charge density, respectively, which leaves us with a coupled set of one second-order and two first-order differential

equations,

$$\frac{dc}{dx} + \rho\frac{d\phi}{dx} = 2j, \tag{24}$$

$$\frac{d\rho}{dx} + c\frac{d\phi}{dx} = 2j, \tag{25}$$

$$-\epsilon^2\frac{d^2\phi}{dx^2} = \rho. \tag{26}$$

Nondimensionalizing the boundary conditions (remembering that there is an extra condition needed to determine the current-voltage relation, $j(v)$ or $v(j)$), we obtain

$$\phi(0) - \delta\epsilon\frac{d\phi}{dx}(0) = 0, \tag{27}$$

$$\phi(1) + \delta\epsilon\frac{d\phi}{dx}(1) = v, \tag{28}$$

$$k_c[c(0) + \rho(0)]e^{\alpha_c\phi(0)} - j_r e^{-\alpha_a\phi(0)} = j, \tag{29}$$

$$-k_c[c(1) + \rho(1)]e^{\alpha_c(\phi(1)-v)} + j_r e^{-\alpha_a(\phi(1)-v)} = j, \tag{30}$$

$$\int_0^1 [c(x) - \rho(x)]dx = 1, \tag{31}$$

where

$$k_c \equiv \frac{K_c L}{4D_+}, \quad j_r \equiv \frac{K_a L C_M}{4D_+ C_{ref}}, \quad v \equiv \frac{zFV}{RT}, \quad \text{and} \quad \delta \equiv \frac{\lambda_S}{\lambda_D}. \tag{32}$$

Keep in mind that the dimensionless rate constants decrease with system size, so that "fast reactions" ($k_c, j_r \gg 1$) may become "slow reactions" ($k_c, j_r = O(1)$) as L is reduced to the micron or submicron scale.

It is important to note that we have scaled the effective Stern-layer width, λ_S, with the Debye screening length, λ_D, rather than the electrode separation L, thus introducing the factor $\epsilon = \lambda_D/L$ in (27) and (28). This choice is important for our asymptotic analysis of the limit $\epsilon \to 0$ at fixed δ, which is intended to describe situations in which L is much larger than *both* λ_S and λ_D. Without it, our analysis would assume that as $\epsilon \to 0$ the Stern layer becomes infinitely wide compared to the diffuse layer, even though it is mainly the macroscopic electrode separation which varies. The limit of very small Stern layer capacitance, which amounts to the Helmholtz model of the double layer [41], is best studied by letting $\delta \to \infty$ *after* $\epsilon \to 0$. In contrast, because ϵ and δ would both be small, the limit of very large Stern-layer capacitance can be studied by simply letting $\delta = 0$, yielding the Dirichlet boundary conditions

$$\phi(0) = 0, \qquad \phi(1) = 1 \tag{33}$$

of the Gouy–Chapman model of the double layer [42, 43]. In our work, we shall consider both limits, starting with the assumption that $\delta = O(1)$, which corresponds to the Gouy–Chapman–Stern model of the double layer [11, 26].

For one-dimensional problems, galvanostatic conditions are more mathematically convenient than potentiostatic conditions. In the former case, j is given, and $v(j)$ is easily obtained from the Stern boundary condition at the anode (28). In the latter case, however, v is specified and $j(v)$ must be determined self-consistently to satisfy (28). Therefore, we shall assume that the current j is specified and solve (24)–(26) subject to the boundary conditions (29) and (30) and the integral constraint (31).

2. Boundary-layer analysis. In this section, we briefly review the classical asymptotic analysis of the PNP equations, pioneered independently by Chernenko [44], Newman [21], and MacGillivray [45], which involves boundary layers of width ϵ (corresponding to diffuse-charge layers of dimensional width, λ_D). As discussed below, the classical asymptotics breaks down at large currents approaching diffusion limitation. Unlike most previous authors, who assume either a fixed potential [42, 43] or fixed interfacial charge [21] at an isolated electrode or fixed concentrations at cell boundaries with ion-permeable membranes [4, 5, 24], we solve for the response of a complete, two-electrode galvanic cell with boundary conditions for Faradaic reactions and Stern-layer capacitance.

Throughout this section, the reader may refer to Figure 1, which compares the uniform asymptotic solutions derived below to numerical solutions at several values of ϵ. These figures illustrate the structure of the field variables in the cell as well as give an indication of the quality of the asymptotic solutions. The numerical solutions are obtained by a straightforward iterative spectral method, described in a companion paper [22].

2.1. Electroneutrality in the bulk solution. The most fundamental approximation in electrochemistry is that of bulk electroneutrality [1]. As first emphasized by Newman [21], however, this does not mean that the charge density is vanishing or unimportant, but rather that over macroscopic distances the charge density is small compared to the total concentration, $|C_+ - C_-| \ll C_+ + C_-$, or, in our dimensionless notation, $|\rho| \ll c$. Mathematically, the "macroscopic limit" corresponds to the limit $\epsilon = \lambda_D/L \to 0$. The electroneutral solution is just the leading order solution when asymptotic series of the form $f(x) = f^{(0)}(x) + \epsilon f^{(1)}(x) + \epsilon^2 f^{(2)}(x) + \cdots$ are substituted for the field variables in (24)–(26).

Carrying out these substitutions and collecting terms with like powers of ϵ, we obtain a hierarchy of differential equations for the expansion functions. At $O(1)$ we have

$$\frac{d\bar{c}^{(0)}}{dx} = 2j, \quad -\bar{c}^{(0)}\bar{E}^{(0)} = 2j, \quad \bar{\rho}^{(0)} = 0, \tag{34}$$

where the bar accent indicates that these expansions are valid in the "bulk region" $\epsilon \ll x \ll 1 - \epsilon$ (or $\lambda_D \ll X \ll L - \lambda_D$). Integrating these equations, we obtain the leading order bulk solution:

$$\bar{c}^{(0)}(x) = c_o + 2jx, \tag{35}$$

$$\bar{E}^{(0)}(x) = \frac{-1}{x + c_o/2j}, \tag{36}$$

$$\bar{\phi}^{(0)}(x) = \phi_o + \log\left(1 + \frac{2jx}{c_o}\right), \tag{37}$$

where the constants of integration c_o and ϕ_o are the values of the bulk concentration and potential extrapolated to the cathode surface at $x = 0$. Note that, despite quasi-electroneutrality, the electrostatic potential does not satisfy Laplace's equation at leading order in the bulk, as emphasized by Levich [46] and Newman [1]. Noting the presence of ϵ^2 in (26) for the dimensionless potential, it is clear that a negligible charge density, $\rho = O(\epsilon^2)$, is perfectly consistent with a nonvanishing Laplacian of the potential. More precisely, we have

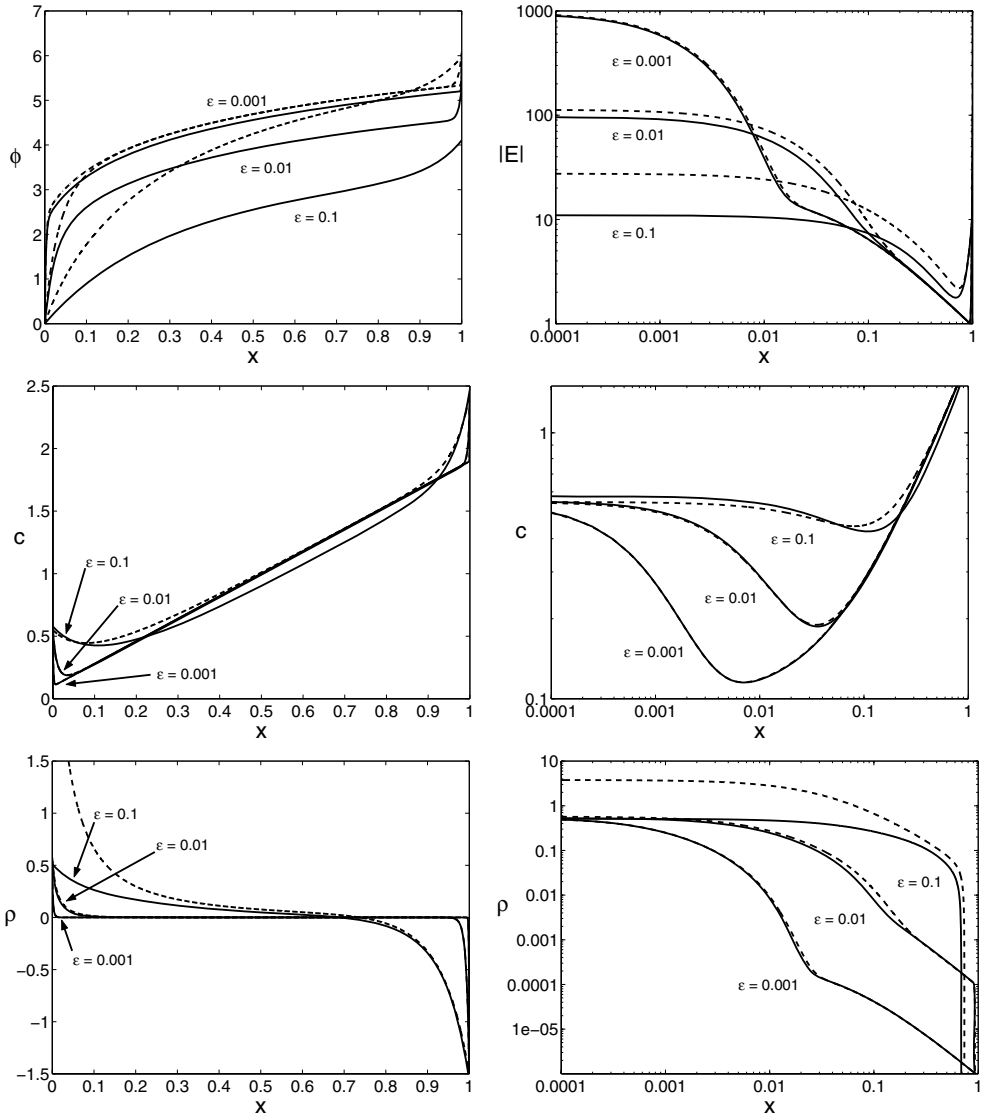


FIG. 1. *Numerical solutions (solid lines) compared with the leading order uniformly valid approximations (dashed lines) given by* (58)–(61) *for the dimensionless potential* $\phi(x)$*, electric field* $E(x)$*, concentration* $c(x)$*, and charge density* $\rho(x)$ *for the case* $j = 0.9$*,* $k_c = 10$*,* $j_r = 10$*,* $\delta = 0$*, and* $\epsilon = 0.001, 0.01, 0.1$*. Linear scales on the left show the entire cell, while log scales on the left zoom in on the cathodic region. Note that in the concentration and charge density profiles, the numerical and asymptotic solutions are barely distinguishable for* $\epsilon \leq 0.01$*.*

$$\bar{\rho}^{(2)}(x) = \frac{d^2\bar{\phi}^{(0)}}{dx^2}(x) = \frac{1}{(x + c_o/2j)^2} \tag{38}$$

at $O(\epsilon^2)$ in (26).

The integral constraint, (31), can be used to evaluate the constant c_o. If we assume that c_- in the boundary layers does not diverge as $\epsilon \to 0$, then they contribute only $O(\epsilon)$ to the total anion number. Therefore,

$$1 = \int_0^1 c_-(x)dx = \int_0^1 \bar{c}^{(0)}(x)dx + O(\epsilon) = c_o + j + O(\epsilon), \tag{39}$$

which implies that $c_o = 1 - j$.

At leading order in the bulk, we have recovered the classical theory dating back a century to Nernst [12, 13, 14]. The solution is electrically neutral with a linear concentration profile whose slope is proportional to the current. This approximation leads to one of the fundamental concepts in electrochemistry, that there exists a "limiting current," $j = 1$, or

$$I = I_d = \frac{4zFD_+C_{ref}A}{L}, \tag{40}$$

corresponding to zero concentration at the cathode, $c_o = 1 - j = 0$. The current is limited by the maximum rate of mass transfer allowed by diffusion, and larger currents would lead to unphysical and mathematically inconsistent negative concentrations (see the appendix).

Examination of the electric field exposes the same limitation on the current: a singularity exists at $j = 1$ that blocks larger currents from being attained. The leading-order bulk approximation to the electric field, $\bar{E}^{(0)} = 1/[x + (1 - j)/2j]$, diverges near the cathode like $1/x$ in the limit $j \to 1$. This would imply that the cell voltage v (calculated below) diverges as $j \to 1$, thus providing a satisfactory theory of the limiting current since an infinite voltage would be necessary to exceed (or even attain) it. Unfortunately, this classical picture due to Nernst [12], based on passing to the *singular* limit $\epsilon = 0$, is not valid for any *finite* value of ϵ because the solution is, in general, unable to satisfy all of the boundary conditions.

2.2. Diffuse charge layers in thermal equilibrium. We now derive the leading order description of the boundary layers in the standard way [21, 5, 26, 1], using (24)–(25). The singular perturbation in (26) can be eliminated with the rescaling $y = x/\epsilon$ indicating that the boundary layer at $x = 0$ has a width $O(\epsilon)$. In terms of this inner variable, the governing equations in the cathode boundary layer are

$$\frac{dc}{dy} + \rho\frac{d\phi}{dy} = 2j\epsilon, \tag{41}$$

$$\frac{d\rho}{dy} + c\frac{d\phi}{dy} = 2j\epsilon, \tag{42}$$

$$-\frac{d^2\phi}{dy^2} = \rho, \tag{43}$$

where ϵ now appears as a regular perturbation since solutions satisfying the cathode boundary conditions and the matching conditions still exist when $\epsilon = 0$. At the anode, the appropriate inner variable is $y = (1 - x)/\epsilon$, and the equations are the same as above except that j is replaced with $-j$, since current is leaving the anode layer, while it is entering the cathode boundary layer.

Expanding the cathode boundary-layer fields (indicated by the check accent) as asymptotic series in powers of ϵ, we obtain at leading order

$$\frac{d\check{c}^{(0)}}{dy} + \check{\rho}^{(0)}\frac{d\check{\phi}^{(0)}}{dy} = 0, \tag{44}$$

$$\frac{d\check{\rho}^{(0)}}{dy} + \check{c}^{(0)}\frac{d\check{\phi}^{(0)}}{dy} = 0. \tag{45}$$

Using (23) to rewrite these equations in terms of c_+ and c_-, we find that the flux of each ionic species in the boundary layer is zero at leading order:

$$\frac{d\check{c}_\pm^{(0)}}{dy} \pm \check{c}_\pm^{(0)} \frac{d\check{\phi}^{(0)}}{dy} = 0. \tag{46}$$

While this equation appears to contradict the fact that the current is nonzero, the paradox is resolved in the same way as electroneutrality is reconciled with a nonharmonic potential in the bulk region: tiny fluctuations about the boundary-layer equilibrium concentration profiles at $O(\epsilon)$ are amplified by a scaling factor of $1/\epsilon$ to sustain the $O(1)$ current. Thus, the leading order contribution to the current in the boundary layer is

$$\epsilon\left(\frac{d\check{c}_+^{(1)}}{dx} + \check{c}_+^{(1)}\frac{d\check{\phi}^{(0)}}{dx} + \check{c}_+^{(0)}\frac{d\check{\phi}^{(1)}}{dx}\right) = \frac{d\check{c}_+^{(1)}}{dy} + \check{c}_+^{(1)}\frac{d\check{\phi}^{(0)}}{dy} + \check{c}_+^{(0)}\frac{d\check{\phi}^{(1)}}{dy} = 4j. \tag{47}$$

Integrating (46) and matching with the bulk, we find that the leading order ionic concentrations are Boltzmann equilibrium distributions:[1]

$$\check{c}_\pm^{(0)}(y) = c_o e^{\pm[\phi_o - \check{\phi}^{(0)}(y)]}, \tag{48}$$

where $c_o = 1 - j$ and $\phi_o = \bar{\phi}^{(0)}(0)$ are obtained by matching with the solution in the bulk. Note that the Boltzmann distribution arises not from an assumption of thermal equilibrium in the boundary layer but as the leading order concentration distribution, even in the presence of a nonnegligible $O(1)$ current.

The general leading-order solution was first derived by Gouy [42] and Chapman [43] and appears in numerous books [1, 26, 32, 33] and recent papers [5, 19]:

$$\check{c}^{(0)}(y) = c_o \cosh[\phi_o - \check{\phi}^{(0)}(y)], \tag{49}$$

$$\check{\rho}^{(0)}(y) = c_o \sinh[\phi_o - \check{\phi}^{(0)}(y)], \tag{50}$$

$$\frac{d\check{\phi}^{(0)}}{dy} = 2\sqrt{c_o}\sinh\left(\frac{\phi_o - \check{\phi}^{(0)}(y)}{2}\right), \tag{51}$$

$$\check{\phi}^{(0)}(y) = \phi_o + 4\tanh^{-1}\left(\gamma_o e^{-\sqrt{c_o}y}\right), \tag{52}$$

where $\gamma_o \equiv \tanh(\zeta_o/4)$ and $\zeta_o \equiv \check{\phi}^{(0)}(0) - \phi_o$ is the leading-order "zeta potential" across the cathodic diffuse layer, which plays a central role in electrokinetic phenomena [32, 33]. Note that the magnitude of the diffuse layer electric field scales as $1/\epsilon$, as illustrated in Figure 1.

The value of $\check{\phi}^{(0)}(0)$ hidden in the zeta potential ζ_o is determined by the Stern boundary condition, (27). If $\delta = 0$ (Gouy–Chapman model), then $\check{\phi}^{(0)}(0) = 0$, or $\zeta_o = -\phi_o$, which means that the entire voltage drop ϕ_o across the cathodic double layer occurs in the diffuse layer. If $\delta = \infty$ (Helmholtz model), then $\check{\phi}^{(0)}(0) = \phi_o$, or $\zeta_o = 0$, in which case the Stern layer carries all the double-layer voltage. For finite $\delta > 0$ (Stern model), ζ_o is obtained in terms of ϕ_o by solving a transcendental algebraic equation,

$$-\zeta_o = 2\delta\sqrt{c_o}\sinh(\zeta_o/2) + \phi_o, \tag{53}$$

[1] The expression for energy in the Boltzmann equilibrium distribution includes only the energy due to electrostatic interactions. "Chemical" contributions to the energy are neglected.

which can be linearized about the two limiting cases and solved for ζ_o,

$$-\zeta_o \sim \begin{cases} \phi_o - 2\delta\sqrt{c_o}\sinh(\phi_o/2) & \text{if } \delta \ll \phi_o/2\sqrt{c_o}\sinh(\phi_o/2), \\ \phi_o/\delta\sqrt{c_o} & \text{if } \delta \gg \phi_o/2\sqrt{c_o}. \end{cases} \tag{54}$$

Note that if $\phi_o \ll 1$, then $-\zeta_o \approx \frac{\phi_o}{1+\delta\sqrt{c_o}}$ is a reasonable approximation for any value of $\delta \geq 0$. Finally, we solve for ϕ_o by applying the Butler–Volmer rate equation, (29), which yields a transcendental algebraic equation for ϕ_o:

$$k_c c_o e^{-\zeta_o + \alpha_c(\zeta_o+\phi_o)} - j_r e^{-\alpha_a(\zeta_o+\phi_o)} = j. \tag{55}$$

Simultaneously solving the pair of (53) and (55) exactly is not possible in general, but below we will analyze various limiting cases.

In the anodic boundary layer, we find the same set of equations as (41)–(43) except that j is replaced by $-j$. Therefore, since the fields do not depend on j at leading order, the anodic boundary layer has the same structure but with different constants of integration. Thus, we find that the leading order description of the anodic boundary layer is given by (49)–(52) with c_o, ϕ_o, γ_o, and ζ_o replaced by different constants c_1, ϕ_1, γ_1, and ζ_1, respectively. Moreover, it is straightforward to show that $c_1 = \bar{c}^{(0)}(1) = 1 + j$ and $\phi_1 = \phi_o + \log(\frac{1+j}{1-j})$.

The leading-order anodic zeta potential, ζ_1, and the potential drop across the entire anodic double layer, $v - \phi_1$, are found by solving another pair of transcendental algebraic equations resulting from the anode Stern and Butler–Volmer boundary conditions, (28) and (30),

$$-\zeta_1 = 2\delta\sqrt{c_1}\sinh(\zeta_1/2) + \phi_1 - v, \tag{56}$$

$$j = -k_c c_1 e^{-\zeta_1 + \alpha_c(\zeta_1+\phi_1-v)} + j_r e^{-\alpha_a(\zeta_1+\phi_1-v)}. \tag{57}$$

As before, the Gouy–Chapman and Helmholtz limits are $\zeta_1 = v - \phi_1$ and $\zeta_1 = 0$, respectively, and for small voltages (or currents) the approximation $\zeta_1 \approx (v-\phi_1)/(1+\delta\sqrt{c_1})$ is valid for all $\delta \geq 0$.

2.3. Leading order uniformly valid approximations. We obtain asymptotic approximations that are uniformly valid across the cell by adding the bulk and boundary-layer approximations and subtracting the overlapping parts:

$$c(x) = [\check{c}^{(0)}(x/\epsilon) - c_o] + \bar{c}^{(0)}(x) + [\hat{c}^{(0)}((1-x)/\epsilon) - c_1] + O(\epsilon), \tag{58}$$

$$\rho(x) = \check{\rho}^{(0)}(x/\epsilon) + \epsilon^2\bar{\rho}^{(2)}(x) + \hat{\rho}^{(0)}((1-x)/\epsilon) + O(\epsilon), \tag{59}$$

$$E(x) = \frac{1}{\epsilon}\frac{d\check{\phi}^{(0)}}{dy}(x/\epsilon) + \frac{d\bar{\phi}^{(0)}}{dx}(x) - \frac{1}{\epsilon}\frac{d\hat{\phi}^{(0)}}{dy}((1-x)/\epsilon) + O(\epsilon), \tag{60}$$

$$\phi(x) = [\check{\phi}^{(0)}(x/\epsilon) - \phi_o] + \bar{\phi}^{(0)}(x) + [\hat{\phi}^{(0)}((1-x)/\epsilon) - \phi_1] + O(\epsilon). \tag{61}$$

Note that we have kept the $O(\epsilon^2)$ term in the charge density since it is the leading-order contribution in the bulk region and is easily computed from (38). As shown in Figure 1, the leading-order uniformly valid solutions are very accurate for $\epsilon \leq 0.01$ (or $L \geq 100\lambda_D$) and reasonably good for $\epsilon = 0.1$. Since higher-order terms are not analytically tractable, it seems numerical solutions must suffice for nanolayers, where $\epsilon \approx 1$, or else other limits of various parameters must be considered, as below.

The discrepancy in electric potential profile at large ϵ in Figure 1 is particularly interesting because it arises from a constraint on the total potential drop across the cell.

To understand the origin of this voltage constraint, recall that the total cell voltage is determined by the current density flowing through the cell (via the voltage-current relationship). While ϵ is technically a parameter in the voltage-current relationship, a leading-order analysis does not capture the ϵ dependence. Thus, the leading-order cell voltage must be the same for all ϵ, which is what we observe in Figure 1. A close examination of the potential and electric field profiles reveals that most of the error in the asymptotic solution for the potential comes from an over-prediction of the electric field strength (and therefore the potential drop) in the cathode region.

3. Polarographic curves for thin double layers, $\epsilon \to 0$. The relationship between current and cell voltage is of primary importance in the study of any electrochemical system, so we now use the results from the previous section to calculate theoretical polarographic curves in several physically relevant regimes. We focus on the effects of the Stern capacitance and the reaction-rate constants through the dimensionless parameters, δ, k_c, and j_r, with $\alpha_c = \alpha_a = 1/2$. For a fixed voltage, the mathematical results are valid in the asymptotic limit of thin double layers, $\epsilon \to 0$.

3.1. Exact results at leading order. Using the uniformly valid approximation (61), we can write the leading-order approximation for the cell voltage as

$$v = \phi_o + 2\tanh^{-1}(j) + (v - \phi_1). \tag{62}$$

We can interpret this expression as a decomposition of the cell voltage into the potential drop across the cathode, bulk, and anode layers, respectively. Note the divergence in the bulk contribution to the cell voltage as $j \to 1$, which we expect from our earlier analysis. In the next section, we explore analytic solutions for several limiting cases and compare them to exact solutions given by (62) with the leading-order cathode and anode diffuse layer potential drops determined implicitly by (53), (55), (56). To make plots in our figures, we use Newton iteration to solve for ϕ_o and $v - \phi_1$ in this algebraic system.

3.2. Cell resistance at low current. Given the common practice of using linear circuit models to describe electrochemical systems [27, 28, 20], it is important to consider the low-current regime, where the cell acts as a simple resistance, $R = V/I$. First, we compute the potential drop across the double layers. Since the procedure is almost identical for the two boundary layers, we focus on the calculation for the cathode. By writing the boundary conditions (29) in the standard Butler–Volmer form involving the exchange current and surface overpotential [26, 1],

$$j = j_o^c\big(e^{-\alpha_c \eta_s^c} - e^{\alpha_a \eta_s^c}\big), \tag{63}$$

where $j_o^c = (k_c c_o e^{-\zeta_o})^{\alpha_a} j_r^{\alpha_c}$ and $\eta_s^c = \Delta\phi_S - \Delta\phi_S^{eq}$ are the cathode exchange current and surface overpotential, respectively. Note that the exchange current contains the Frumkin correction through the factor $e^{-\zeta_o}$ [26]. For low-current densities, we expect the surface overpotential to be small, so we can linearize this equation to obtain

$$j \sim -j_o^c \eta_s^c, \tag{64}$$

where we have used the fact that $\alpha_c + \alpha_a = 1$. Rewriting this equation in terms of $\check{\phi}(0)$, we find that

$$\check{\phi}(0) \sim \frac{j}{j_o^c} + \check{\phi}_{eq}(0), \tag{65}$$

where $\check{\phi}_{eq}(0)$ is the value of $\check{\phi}(0)$ calculated from the cathode Butler–Volmer rate equation when there is no current flowing through the electrode. The zeta potential ζ_o in the formula for the exchange current is determined by combining (65) with (53) to obtain a single equation for ζ_o:

$$-2\delta\sqrt{c_o}\sinh(\zeta_o/2) \sim \frac{j}{(k_c c_o e^{-\zeta_o})^{\alpha_a} j_r^{\alpha_c}} + \log\left(\frac{j_r}{k_c c_o}\right) + \zeta_o. \tag{66}$$

Finally, to compute the total double-layer potential drop, we add the potential drop across the diffuse layer to $\check{\phi}(0)$:

$$\phi_o = \check{\phi}(0) - \zeta_o \sim \frac{j}{j_o^c} + \log\left(\frac{j_r}{k_c c_o}\right). \tag{67}$$

A similar calculation at the anode results in

$$v - \phi_1 \sim \frac{j}{j_o^a} + \log\left(\frac{k_c c_1}{j_r}\right), \tag{68}$$

where $j_o^a = (k_c c_1 e^{-\zeta_1})^{\alpha_a} j_r^{\alpha_c}$ and ζ_1 is determined by the anode equivalent of (66).

Combining these results with the potential drop across the bulk solution, we find that the total cell voltage is given by

$$\begin{aligned} v(j) &\sim 4\tanh^{-1}(j) + \frac{j}{j_o^c} + \frac{j}{j_o^a} \\ &\approx j\left(4 + \frac{1}{j_o^c} + \frac{1}{j_o^a}\right) \\ &= j\ r. \end{aligned} \tag{69}$$

This result gives the dimensionless resistance, r, of the electrochemical thin film as a function of the physical properties of the electrodes and the electrolyte. Note that the Stern-layer capacitance is accounted for implicitly via the calculation of the electrode zeta potentials.

3.3. Simple analytical formulae. The exact leading-order current-voltage relation simplifies considerably in a variety of physically relevant limits. These approximate formulae provide insight into the basic physics and may be useful in interpreting experimental data.

3.3.1. The Gouy–Chapman limit ($\delta \to 0$). In this limit, the capacitance of the diffuse layer of the charged double layer is negligible compared to the capacitance of the compact layer. As a result, the voltage drop across the diffuse layer accounts for the entire potential drop across the charged double layer. Physically, this limit corresponds to the limits of low ionic concentration or zero ionic volume [26]. Since $\zeta_o = -\phi_o$ and $\zeta_1 = v - \phi_1$ when $\delta = 0$, the Butler–Volmer rate equations, (55) and (57), reduce to

$$k_c(1-j)e^{-\zeta_o} - j_r = j \quad \text{and} \quad -k_c(1+j)e^{-\zeta_1} + j_r = j. \tag{70}$$

Solving for ζ_o and ζ_1, we find that

$$\zeta_o = \ln(1-j) - \ln\left(\frac{j_r + j}{k_c}\right) \quad \text{and} \quad \zeta_1 = \ln(1+j) + \ln\left(\frac{k_c}{j_r - j}\right), \tag{71}$$

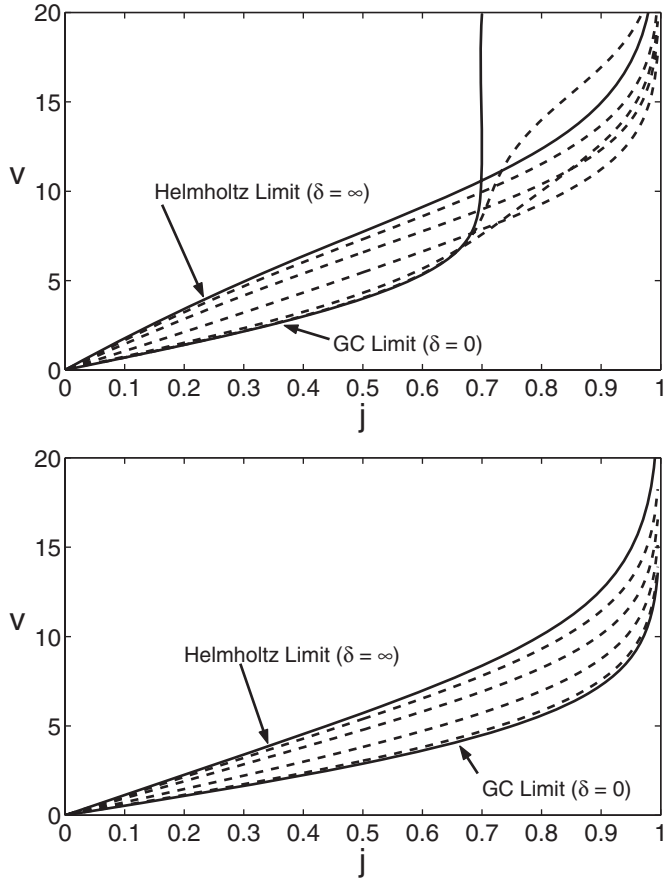


FIG. 2. *Exact polarographic curves (dashed lines) for varying δ values compared to polarographic curves for the Gouy–Chapman ($\delta = 0$) and Helmholtz ($\delta = \infty$) limits (solid lines). Top: a reaction-limited cell ($j_r < 1$) with physical parameters, $k_c = 0.03$, $j_r = 0.7$. Notice that above the reaction-limited current density, j_r, the highest cell voltages occur for δ values near 0. Bottom: a diffusion-limited cell ($j_r > 1$) with physical parameters, $k_c = 0.05$, $j_r = 1.5$. In both cases, δ increases as the curves move upwards.*

which can be substituted into (62) to obtain

$$v(j) = 4\tanh^{-1}(j) + 2\tanh^{-1}(j/j_r). \tag{72}$$

Notice that the boundary layers make a nontrivial contribution to the leading-order cell voltage. The $2\tanh^{-1}(j/j_r)$ term is especially interesting because it indicates the existence of a reaction-limited current when $j_r < 1$. In hindsight, it is obvious that reaction-limited currents exist in the Gouy–Chapman limit because the reaction kinetics at the anode do not permit a current greater than j_r. We emphasize, however, that the Gouy–Chapman limit is singular because there is no problem achieving current densities above j_r for any $\delta > 0$ (see Figure 2). For any finite $\delta > 0$, the shift of the anode double-layer potential drop to the Stern layer helps the dissolution reaction while suppressing the deposition reaction, which permits the current density to rise greater than j_r.

Note that the cathodic and anodic boundary layers do not evenly contribute to

the cell voltage near the limiting currents. In a diffusion-limited cell, the cathodic layer makes the greater contribution because as $j \to 1$, ζ_o diverges while ζ_1 approaches a finite limit. We expect this behavior because as $j \to 1$, the electric field diverges only at $x = 0$. However, when the cell is reaction-limited, the division of cell voltage between the boundary layers is reversed as j approaches the limiting current j_r. Even the voltage drop in the bulk becomes negligible compared to ζ_1 in the reaction-limited case. In this situation, the cell voltage diverges as $j \to j_r$ because the only way to achieve a current near j_r is to drastically reduce the deposition reaction at the anode. In other words, the cation concentration at the anode must be made extremely small, which requires a huge anodic zeta potential.

3.3.2. The Helmholtz limit ($\delta \to \infty$). This is the reverse of the Gouy–Chapman limit. Here, the capacitance of the compact layer is negligible, so the potential drop across the double layer resides completely in the compact layer. The Helmholtz limit holds for concentrated solutions or solvents with low dielectric constants and other situations where the Debye screening length becomes negligible [26]. It also describes a thick dielectric or insulating layer on an electrode [38, 20].

In the Helmholtz limit, $\zeta_o = 0 = \zeta_1$, so the Butler–Volmer rate equations take the form

$$k_c(1-j)e^{\alpha_c \phi_o} - j_r e^{-\alpha_a \phi_o} = j, \tag{73}$$

$$-k_c(1+j)e^{\alpha_c(\phi_1 - v)} + j_r e^{-\alpha_a(\phi_1 - v)} = j. \tag{74}$$

Solving these equations for ϕ_o and $v - \phi_1$ under the assumption of a symmetric electron-transfer reaction (i.e., $\alpha_c = 1/2 = \alpha_a$) and substituting into the formula for the cell voltage, we find that

$$v(j) = 6\tanh^{-1}(j) + 2\ln\left(\frac{j + \sqrt{j^2 + 4j_r k_c(1-j)}}{-j + \sqrt{j^2 + 4j_r k_c(1+j)}}\right). \tag{75}$$

While this expression appears to be more complicated than the one obtained for the Gouy–Chapman model, it is not very different when $j_r > 1$, as can be seen in Figures 2 and 3. In fact, the wide spread in the polarographic curves observed in Figure 2 requires that $k_c \ll j_r$; otherwise, all of the curves would be difficult to distinguish. Moreover, as we shall see in the next section, in the limit of fast reactions, both models lead to the same expression for the cell voltage for $j_r > 1$. On the other hand, when $j_r < 1$, the two models are qualitatively very different. While the Gouy–Chapman model gives rise to a reaction-limited current, the Helmholtz model does not.

3.3.3. The fast-reaction limit ($j_r \gg 1$, $(j_r)^{\alpha_a}(k_c)^{\alpha_c} \gg j/(1-j)^{\alpha_a}$). The polarographic curves for all δ values collapse onto each other in the limit of fast-reaction kinetics (Figure 3). Even the assumption of symmetry in the electron-transfer reaction is not required. When reaction rates are much larger than the current, the two reaction-rate terms in the Butler–Volmer equations, (55) and (57), must balance each other at leading order:

$$k_c(1-j)e^{-\zeta_o + \alpha_c(\zeta_o + \phi_o)} - j_r e^{-\alpha_a(\zeta_o + \phi_o)} \approx 0, \tag{76}$$

$$-k_c(1+j)e^{-\zeta_1 + \alpha_c(\zeta_1 + \phi_1 - v)} + j_r e^{-\alpha_a(\zeta_1 + \phi_1 - v)} \approx 0. \tag{77}$$

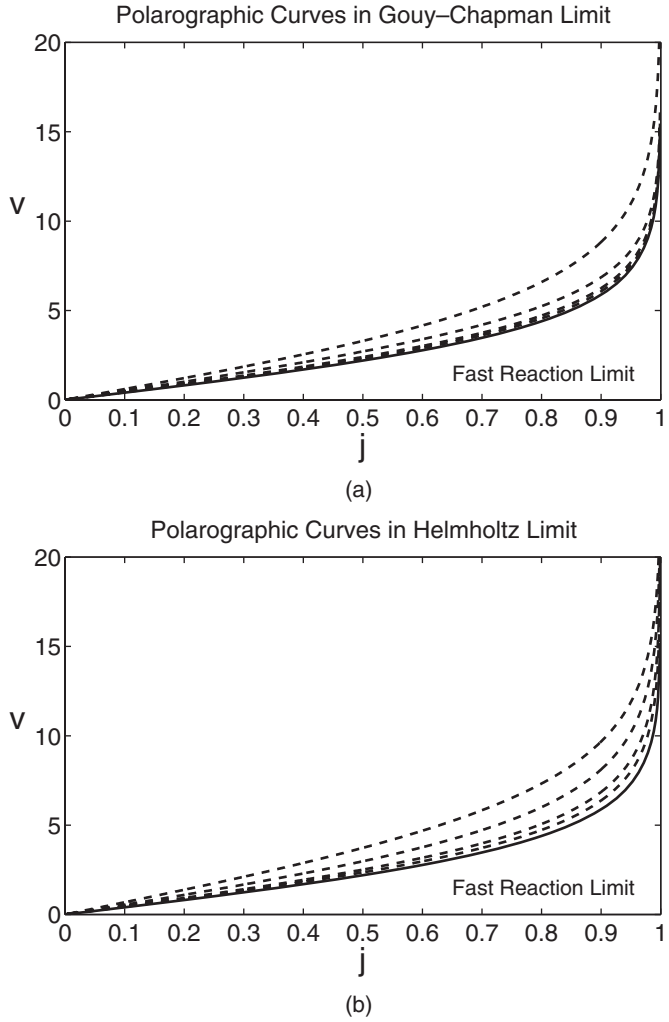


FIG. 3. *Polarographic curves in* (a) *the Gouy–Chapman limit and* (b) *the Helmholtz limit as the reaction-rate constants are increased (dashed lines). For these plots, the reaction-rate constants* ($j_r = 1, 2, 5,$ *and* 10) *increase as the curves shift towards the lower right and are related by* $k_c = j_r/2$. *It should be noted that the fast-reaction limit is reached very quickly; in both plots, the curve closest to the fast-reaction curve has a* j_r *value of only* 10.

Since $\alpha_c + \alpha_a = 1$ for theoretical models of single electron-transfer reactions [1, 11, 30], we can solve explicitly for ϕ_o and $v - \phi_1$ to obtain

$$\phi_o = \ln\left(\frac{j_r}{k_c(1-j)}\right), \qquad v - \phi_1 = \ln\left(\frac{k_c(1+j)}{j_r}\right). \tag{78}$$

Thus, for fast-reaction kinetics, the leading-order cell voltage is given by

$$v(j) = 4\tanh^{-1}(j). \tag{79}$$

Notice that this is exactly the fast-reaction limit of $v(j)$ that we find in both the Gouy–Chapman and Helmholtz limits. It is straightforward to check the validity of the

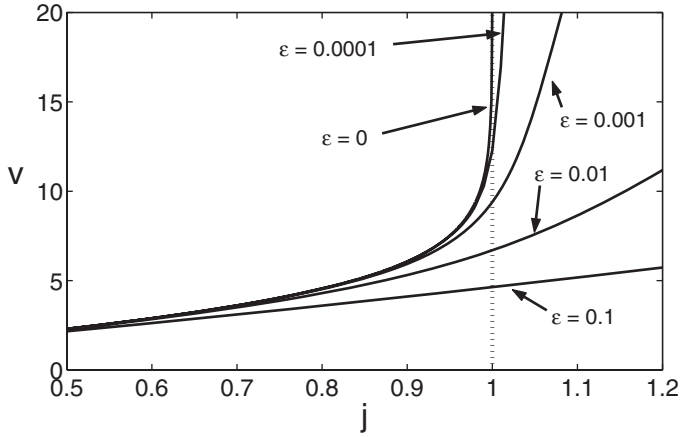


FIG. 4. *Polarographic curves for ϵ values of* 0, 0.0001, 0.001, 0.01, *and* 0.1 *(listed in order from uppermost to lowest curves) with the other physical parameters taken to be $\delta = 0$, $k_c = 10$, and $j_r = 10$. Notice that for any $\epsilon > 0$, the cell has no problem achieving current densities higher than the diffusion-limited current (dashed vertical line). All of these curves, with the exception of the exact $\epsilon = 0$ curve, were generated by numerically solving* (24)–(26) *subject to the boundary conditions* (27)–(31) *using the method of our companion paper* [22].

assumptions made in (76) and (77) by substituting these results into the expressions for the reaction rates and observing that the zeta potentials satisfy the bounds $\zeta_o \leq 0$ and $\zeta_1 \leq \ln(\frac{k_c(1+j)}{j_r - j})$, which follow from the monotonicity of ζ_o and ζ_1 as functions of δ.

4. Thick double layers, $\epsilon = O(1)$. Up to this point, we have examined only the current-voltage characteristics in the singular limit $\epsilon \to 0$, where the current density cannot exceed its diffusion-limited value, $j = 1$. The situation changes for any finite $\epsilon > 0$.

4.1. What limiting current? As is clearly evident in Figure 4, the cell has no problem breaking through the classical limiting current for $\epsilon > 0$. Figure 4 also shows that the ϵ dependence of the polarographic curves becomes significant only at currents approaching the diffusion-limited current; below $j \approx 0.5$, the curves are nearly indistinguishable. Moreover, as ϵ increases, the upper ends of the polarographic curves flatten out and shift downwards. This decrease in the cell voltage for large ϵ values arises because the diffuse charge layers overlap and are able to interact with each other. More precisely, the cell has become so small (relative to the Debye screening length) that the electric fields from the two diffuse layers partially cancel each other out throughout the cell, resulting in a lower total cell voltage. It should be emphasized that this effect is observable only because we are studying a two-electrode system. Single-electrode systems (in addition to being not physically achievable) are not capable of showing this behavior because they always implicitly assume an infinite system size, which effectively discards any interactions from "far away" electrodes.

4.2. Breakdown of the classical approximation. For a diffusion-limited cell, the classical nonlinear asymptotic analysis just presented leads to an aesthetically appealing theory that predicts a limiting current at $j = 1$. The existence of this limiting current fits nicely with our physical intuition that the concentration of cations in a solution must always remain nonnegative. In reality, however, the analysis breaks

down as the current approaches (and exceeds) its limiting value.

The breakdown of the classical asymptotics is evident upon examining the expansions for the bulk field variables as the current is increased toward its diffusion-limited value. Calculating a few of the higher-order terms in the bulk asymptotic expansion, we find that

$$-\bar{E}(x) = \frac{2j}{\bar{c}^{(0)}} + \frac{3}{2}\epsilon^2\frac{(2j)^3}{(\bar{c}^{(0)})^4} + \frac{111}{4}\epsilon^4\frac{(2j)^5}{(\bar{c}^{(0)})^7} + \frac{6045}{4}\epsilon^6\frac{(2j)^7}{(\bar{c}^{(0)})^{10}} + O(\epsilon^8), \tag{80}$$

$$\bar{c}(x) = \bar{c}^{(0)} + \frac{1}{2}\epsilon^2\frac{(2j)^2}{(\bar{c}^{(0)})^2} + \frac{3}{2}\epsilon^4\frac{(2j)^4}{(\bar{c}^{(0)})^5} + \frac{231}{8}\epsilon^6\frac{(2j)^6}{(\bar{c}^{(0)})^8} + O(\epsilon^8), \tag{81}$$

$$\bar{\rho}(x) = 0 + \epsilon^2\frac{(2j)^2}{(\bar{c}^{(0)})^2} + 6\epsilon^4\frac{(2j)^4}{(\bar{c}^{(0)})^5} + \frac{777}{4}\epsilon^6\frac{(2j)^6}{(\bar{c}^{(0)})^8} + O(\epsilon^8). \tag{82}$$

Since $\bar{c}^{(0)} \to 2x$ as $j \to 1$, the higher-order terms are clearly more singular than the leading-order term at the limiting current. Rubinstein and Shtilman make a similar observation from a potentiostatic perspective; they note that the asymptotic expansions are not uniform in the cell voltage [24].

The inconsistency in the classical approximation was apparently first noticed by Levich, who observed that the leading-order solution in the bulk predicts an infinite charge density when the current density reaches 1, which directly contradicts the assumption of bulk charge neutrality [46]. As $j \to 1$, the bulk charge density is given by

$$\bar{\rho} = -\epsilon^2\frac{d^2\bar{\phi}}{dx^2} = \frac{\epsilon^2}{[x + (1-j)/2j]^2} \approx \frac{\epsilon^2}{x^2}, \tag{83}$$

which diverges at the cathode.

Smyrl and Newman first showed that these paradoxical results are related to the breakdown of thermal equilibrium charge profiles near the cathode, leading to a significant expansion of the double layer into the bulk solution [23]. They argue that the assumption of electroneutrality breaks down when $\bar{\rho} \approx \bar{c}$. Since $\bar{c}$ is proportional to x at the limiting current, the bulk approximation fails to be valid for x smaller than $O(\epsilon^{2/3})$, which leads to a boundary layer that is thicker than the usual Debye length. From an alternative perspective, the problems begin when $\bar{\rho}(0)/\bar{c}(0) \approx 1$ (see Figure 5). Using this criterion, we find that the classical asymptotic theory is appropriate only when $c_o \gg (2j\epsilon)^{2/3}$ or, equivalently, $j \ll 1 - (2\epsilon)^{2/3}$. Since the cell voltage is approximately $4\tanh^{-1}(j)$ in many situations, this regime also corresponds to $v = O(|\ln\epsilon|)$. This shows that in thin films, where ϵ is not so small, it is easy to exceed the classical limiting current and achieve rather different charge profiles [22].

5. Conclusion. In summary, we have revisited the classical PNP equations, analyzing for the first time the effect of physically realistic boundary conditions for thin-film galvanic cells and other microelectrochemical systems. In particular, we focused on the effect of Stern-layer capacitance and Faradaic reactions with Butler–Volmer kinetics. Such boundary conditions contain new physics, such as the possibility of a reaction-limited current due to the slow injection of ions at the anode. We also find that the Stern layer generally allows the cell to exceed limiting currents by carrying diverging portions of the cell voltage, which would otherwise end up in the diffuse part of the double layer. We have provided analytical formulae for current-voltage relations that should prove useful in characterizing the differential resistance of thin films, such as those used in on-chip microbatteries. Here, we have focused on the classical nonlinear regime in which thin double layers remain in thermal equilibrium; the more

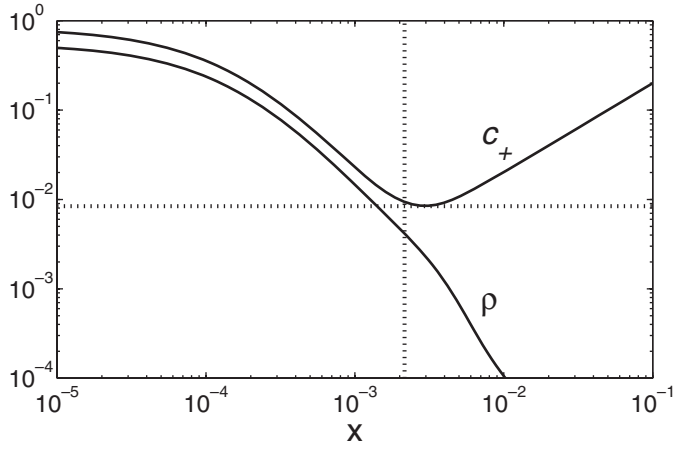


FIG. 5. *Numerical solutions for the dimensionless cation concentration $c_+(x)$ and (full) charge density $2\rho(x)$ at the diffusion-limited current ($j = 1.0$) with physical parameters $k_c = 10$, $j_r = 10$, $\alpha_c = \alpha_a = 0.5$, $\delta = 0.0$, and $\epsilon = 0.0001$. In the cathode region, electroneutrality breaks down as the solution becomes cation rich in order to satisfy the reaction boundary conditions. Note that when $x = O(\epsilon^{2/3})$, c_+ and ρ are both $O(\epsilon^{2/3})$. For reference, the dashed vertical line shows where $x = \epsilon^{2/3}$, and the dashed horizontal line shows where $y = \epsilon^{2/3}[(2+2^{2/3})+4/(2+2^{2/3})^2] \approx c_+(\epsilon^{2/3})$.*

exotic, nonequilibrium regime, which arises at and above the classical limiting current, is analyzed in the companion paper [22].

Appendix. Positivity of ion concentrations. The positivity of the ion concentrations follows directly from the mathematical formulation of the problem. For the anion concentration, (22) can be integrated exactly using the integrating factor e^{ϕ} to yield

$$c_-(x) = Ae^{\phi(x)}, \tag{84}$$

which implies that the sign of $c_-(x)$ is the same across the entire domain. Since the integral constraint (31) requires that $c_-(x)$ is positive somewhere in the domain, $c_-(x)$ must be positive *everywhere* in the domain.

For the cation concentration, we make use of the reaction boundary conditions. Integrating (21) with the integrating factor e^{ϕ}, we obtain

$$c_+(x) = c_+(0)e^{\phi(0)-\phi(x)} + 4je^{-\phi(x)} \int_0^x e^{\phi(y)} dy. \tag{85}$$

Clearly, the integral term is positive because e^{ϕ} is positive everywhere. Moreover, the reaction boundary condition (29) implies that $c_+(0) > 0$ because both j_r and k_c are positive. Thus, we find that the cation concentration is strictly positive.

Acknowledgments. The authors thank F. Argoul, J. J. Chae, H. A. Stone, and W. Y. Tam for many helpful discussions.

REFERENCES

[1] J. NEWMAN, *Electrochemical Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1991.
[2] I. RUBINSTEIN, *Electro-Diffusion of Ions*, SIAM Stud. Appl. Math. 11, SIAM, Philadelphia, PA, 1990.

[3] V. Barcilon, D.-P. Chen, R. S. Eisenberg, *Ion flow through narrow membrane channels: Part* II, SIAM J. Appl. Math., 52 (1992), pp. 1405–1425.
[4] J.-H. Park and J. W. Jerome, *Qualitative properties of steady-state Poisson–Nernst–Planck systems: Mathematical study*, SIAM J. Appl. Math., 57 (1997), pp. 609–630.
[5] V. Barcilon, D.-P. Chen, R. S. Eisenberg, and J. W. Jerome, *Qualitative properties of steady-state Poisson–Nernst–Planck systems: Perturbation and simulation study*, SIAM J. Appl. Math., 57 (1997), pp. 631–648.
[6] N. J. Dudney, J. B. Bates, D. Lubben, and F. X. Hart, *Thin-film rechargeable lithium batteries with amorphous $Li_xMn_2O_4$ cathodes*, in Thin Film Solid Ionic Devices and Materials, J. Bates, ed., The Electrochemical Society, Pennington, NJ, 1995, pp. 201–214.
[7] B. Wang, J. B. Bates, F. X. Hart, B. C. Sales, R. A. Zuhr, and J. D. Robertson, *Characterization of thin-film rechargeable lithium batteries with lithium cobalt oxide cathodes*, J. Electrochem. Soc., 143 (1996), pp. 3204–3213.
[8] B. J. Neudecker, N. J. Dudney, and J. B. Bates, *"Lithium-free" thin-film battery with* in situ *plated Li anode*, J. Electrochem. Soc., 147 (2000), pp. 517–523.
[9] N. Takami, T. Ohsaki, H. Hasabe, and M. Yamamoto, *Laminated thin Li-ion batteries using a liquid electrolyte*, J. Electrochem. Soc., 149 (2002), pp. A9–A12.
[10] Z. Shi, L. Lü, and G. Ceder, *Solid State Thin Film Lithium Microbatteries*, Singapore-MIT Alliance Technical Report: Advanced Materials for Micro- and Nano-Systems Collection, 2003; available online from http://hdl.handle.net/1721.1/3672.
[11] C. M. A. Brett and A. A. O. Brett, *Electrochemistry. Principles, Methods, and Applications*, Oxford Science, Oxford, 1993.
[12] W. Nernst, *Theorie der Reaktionsgeschwindigkeit in heterogenen Systemen*, Z. Phys. Chem., 47 (1904), pp. 52–55.
[13] E. Brunner, *Reaktionsgeschwindigkeit in heterogenen Systemen*, Z. Phys. Chem., 47 (1904), pp. 56–102.
[14] E. Brunner, *Die kathodische und anodische Stromspannungskurve bei der Elektrolyse von Jod-Jodkaliumlösungen*, Z. Phys. Chem., 58 (1907), pp. 1–126.
[15] H.-C. Chang and G. Jaffé, *Polarization in electrolytic solutions. Part* I. *Theory*, J. Chem. Phys., 20 (1952), pp. 1071–1077.
[16] G. Jaffé and C. Z. LeMay, *On polarization in liquid dielectrics*, J. Chem. Phys., 21 (1953), pp. 920–928.
[17] E. M. Itskovich, A. A. Kornyshev, and M. A. Vorotyntsev, *Electric current across the metal-solid electrolyte interface.* I. *Direct current, current-voltage characteristic*, Phys. Stat. Sol. (a), 39 (1977), pp. 229–238.
[18] A. A. Kornyshev and M. A. Vorotyntsev, *Conductivity and space charge phenomena in solid electrolytes with one mobile charge carrier species: A review with original material*, Electrochim. Acta, 26 (1981), pp. 303–323.
[19] A. Bonnefont, F. Argoul, and M. Z. Bazant, *Analysis of diffuse-layer effects on time-dependent interfacial kinetics*, J. Electroanal. Chem., 500 (2001), pp. 52–61.
[20] M. Z. Bazant, K. Thornton, and A. Ajdari, *Diffuse-charge dynamics in electrochemical systems*, Phys. Rev. E (3), 70 (2004), article 021506.
[21] J. Newman, *The polarized diffuse double layer*, Trans. Faraday Soc., 61 (1965), pp. 2229–2237.
[22] K. T. Chu and M. Z. Bazant, *Electrochemical thin films at and above the classical limiting current*, SIAM J. Appl. Math., 65 (2005), pp. 1485–1505.
[23] W. H. Smyrl and J. Newman, *Double layer structure at the limiting current*, Trans. Faraday Soc., 63 (1967), pp. 207–216.
[24] I. Rubinstein and L. Shtilman, *Voltage against current curves of cation exchange membranes*, J. Chem. Soc. Faraday Trans. II, 75 (1979), pp. 231–246.
[25] P. Delahay, *Double Layer and Electrode Kinetics*, Interscience, New York, 1965.
[26] A. J. Bard and L. R. Faulkner, *Electrochemical Methods*, John Wiley & Sons, New York, 2001.
[27] J. R. Macdonald, *Impedance spectroscopy: Old problems and new developments*, Electrochim. Acta, 35 (1990), pp. 1483–1492.
[28] L. A. Geddes, *Historical evolution of circuit models for the electrode-electrolyte interface*, Ann. Biomedical Engrg., 25 (1997), pp. 1–14.
[29] O. Stern, *Zur theorie der elektrolytischen doppelschicht*, Z. Elektrochem., 30 (1924), pp. 508–516.
[30] C. Chidsey, *Kinetics of electrode reactions*, in Physical Chemistry, R. S. Berry, S. A. Rice, and J. Ross, eds., Oxford University Press, New York, 2000, pp. 999–1008.
[31] A. Frumkin, *Wasserstoffüberspannung und struktur der doppelschict*, Z. Phys. Chem., 164A (1933), pp. 121–133.

[32] R. J. Hunter, *Foundations of Colloid Science*, Oxford University Press, Oxford, 2001.
[33] J. Lyklema, *Fundamentals of Interface and Colloid Science. Volume* II: *Solid-Liquid Interfaces*, Academic Press, San Diego, CA, 1995.
[34] D. C. Grahame, *The electrical double layer and the theory of electrocapillarity*, Chem. Rev., 41 (1947), pp. 441–501.
[35] D. C. Grahame, *Differential capacity of mercury in aqueous sodium fluoride solutions.* I. *Effect of concentration at* 25°, J. Amer. Chem. Soc., 76 (1954), pp. 4819–4823.
[36] J. R. Macdonald, *Theory of the differential capacitance of the double layer in unadsorbed electrolytes*, J. Chem. Phys., 22 (1954), pp. 1857–1866.
[37] J. R. Macdonald, *Static space charge and capacitance for a single blocking electrode*, J. Chem. Phys., 29 (1958), pp. 1346–1358.
[38] A. Ajdari, *AC pumping of liquids*, Phys. Rev. E (3), 61 (2000), pp. R45–R48.
[39] M. Z. Bazant and T. M. Squires, *Induced-charge electro-kinetic phenomena: Theory and microfluidic applications*, Phys. Rev. Lett., 92 (2004), article 066101.
[40] T. M. Squires and M. Z. Bazant, *Induced-charge electro-osmosis*, J. Fluid Mech., 509 (2004), pp. 217–252.
[41] H. Helmholtz, *Studien über electrische Grenzschichten*, Ann. Phys. Chem., 7 (1879), pp. 337–382.
[42] M. Gouy, *Sur la Constitution de la Charge Électrique a la Surface d'un Électrolyte*, J. de Phys., 9 (1910), pp. 457–468.
[43] D. L. Chapman, *A contribution to the theory of electrocapillarity*, Philos. Mag., 25 (1913), pp. 475–481.
[44] A. A. Chernenko, *The theory of the passage of direct current through a solution of a binary electrolyte*, Dokl. Akad. Nauk SSSR, 153 (1962), pp. 1129–1131 (in Russian).
[45] A. D. MacGillivray, *Nernst-Planck equations and the electroneutrality and Donnan equilibrium assumptions*, J. Chem. Phys., 48 (1968), pp. 2903–2907.
[46] V. G. Levich, *Physico-chemical Hydrodynamics*, Prentice-Hall, London, 1962.

SIAM J. APPL. MATH.
Vol. 65, No. 5, pp. 1485–1505

ELECTROCHEMICAL THIN FILMS AT AND ABOVE THE CLASSICAL LIMITING CURRENT*

KEVIN T. CHU† AND MARTIN Z. BAZANT†

Abstract. We study a model electrochemical thin film at DC currents exceeding the classical diffusion-limited value. The mathematical problem involves the steady Poisson–Nernst–Planck equations for a binary electrolyte with nonlinear boundary conditions for reaction kinetics and Stern-layer capacitance, as well as an integral constraint on the number of anions. At the limiting current, we find a nested boundary-layer structure at the cathode, which is required by the reaction boundary condition. Above the limiting current, a depletion of anions generally characterizes the cathode side of the cell. In this regime, we derive leading-order asymptotic approximations for the (i) classical bulk space-charge layer and (ii) another nested highly charged boundary layer at the cathode. The former involves an exact solution to the Nernst–Planck equations for a single, unscreened ionic species, which may apply more generally to Faradaic conduction through very thin insulating films. By matching expansions, we derive current-voltage relations well into the space-charge regime. Throughout our analysis, we emphasize the strong influence of the Stern-layer capacitance on cell behavior.

Key words. Poisson–Nernst–Planck equations, electrochemical systems, limiting current, reaction boundary conditions, double-layer capacitance, polarographic curves

AMS subject classifications. 34B08, 34B16, 34B60, 35E05

DOI. 10.1137/040609926

Introduction. Thin-film technologies offer a promising way to construct rechargeable microbatteries, which can be directly integrated into modern electronic circuits [1, 2, 3, 4, 5, 6]. Due to the power-density requirements of many applications, such as portable electronics, microbatteries are likely to be operated at high-current density, possibly exceeding diffusion limitation. In a thin film, very large electric fields are easily produced by applying only small voltages, due to the small electrode separation, which may be comparable to the Debye screening length. Under such conditions, the traditional postulates of macroscopic electrochemical systems [7, 8]—bulk electroneutrality and equilibrium double layers—break down near the classical diffusion-limited current [9]. The mathematical justification for these postulates is based on matched asymptotic expansions in the limit of thin double layers [10, 11, 12], which require subtle modifications at large currents.

The concept of a "limiting current," due to the maximum steady-state flux of diffusion across an electrochemical cell, was introduced by Nernst a century ago [13]. Consider the simplest case of a binary electrolyte between parallel plate electrodes with cation redox reactions and inert anions. Assuming neutrality, the bulk concentration is a linear function of distance (due to steady diffusion) with a gradient proportional to the current. Since the total number of anions is fixed, the total integral of the bulk concentration must also be fixed, which implies that the concentration at the cathode decreases linearly with current. The "diffusion-limited current" corre-

*Received by the editors June 12, 2004; accepted for publication (in revised form) November 1, 2004; published electronically May 12, 2005. This work was supported in part by the MRSEC program of the National Science Foundation under award DMR 02-13282 and in part by the Department of Energy through the Computational Science Graduate Fellowship (CSGF) program provided under grant DE-FG02-97ER25308.

http://www.siam.org/journals/siap/65-5/60992.html

†Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139 (kchu@mit.edu, bazant@mit.edu).

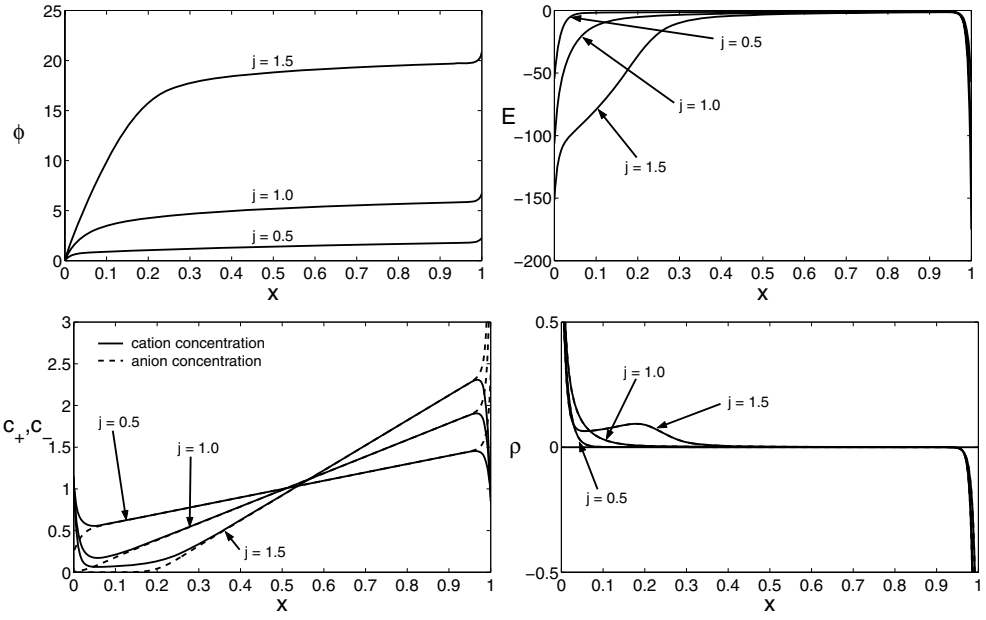


FIG. 1. *Profiles of the dimensionless potential (top left), electric field (top right), total ionic concentration (bottom left), and charge density (bottom right) in three regimes: below the classical diffusion-limited current ($j = 0.5$), at the limiting current ($j = 1$), and above the limiting current ($j = 1.5$). These are numerical solutions to our model problem with the following dimensionless parameters: $\epsilon = 0.01$, $\delta = 0$, $k_c = 10$, $j_r = 10$.*

sponds to a vanishing bulk concentration at the cathode, and, as the name suggests, it can never be reached, except with an infinite voltage.

It was eventually realized that the classical theory is flawed, as illustrated in Figure 1 by numerical solutions to our model problem below. The bulk concentration remains linear, but the system is clearly able to achieve and even exceed the classical limiting current (as shown in the lower left panel of the figure). Levich was perhaps the first to notice that the assumption of bulk electroneutrality yields approximate solutions to the Poisson–Nernst–Planck (PNP) equations, which are not self-consistent near the limiting current, since the predicted charge density eventually exceeds the salt concentration near the cathode [14]. This paradox was first resolved by Smyrl and Newman, who showed that the double layer expands at the limiting current as the Poisson–Boltzmann approximation of thermal equilibrium breaks down [15]. Rubinstein and Shtilman later pointed out that mathematical solutions also exist for larger currents, well above the classical limiting value, characterized by a region of nonequilibrium "space charge" extending significantly into the neutral bulk [16]. As shown in Figure 1, the space-charge layer exhibits anomalously large electric fields and charge densities, compared to the equilibrium double layers at smaller currents.

The possibility of superlimiting currents has been studied extensively in the different context of bulk liquid electrolytes, where a thin space-charge layer drives nonlinear electro-osmotic slip. This phenomenon of "electro-osmosis of the second kind" was introduced by Dukhin for the nonlinear electrophoresis of ion-selective, conducting colloidal particles [17], and Ben and Chang have recently studied it in microfluidics [18]. The mathematical analysis of second-kind electro-osmosis using matched asymptotic expansions, similar to the approach taken here, was first developed by Rubinstein

and Zaltzman for related phenomena at electrodialysis membranes [19, 20]. In earlier studies, the space-charge layer was also invoked by Bruinsma and Alexander [21] to predict hydrodynamic instability during electrodeposition and by Chazalviel [22] in a controversial theory of fractal electrochemical growth.

As in our companion paper on sublimiting currents [9], here we consider (typically solid or gel) thin films, e.g., arising in microbatteries, which approach the classical limiting current without hydrodynamic instability. At micron or smaller length scales, the space-charge layer need not be "thin" compared to the film thickness, so we also analyze currents well above the classical limiting current, apparently for the first time. In both regimes, close to and far above the classical limiting current, we derive matched asymptotic expansions for the concentration profiles and potential, which we compare against numerical solutions. In addition to our focus on superlimiting currents and small systems, a notable difference with the literature on second-kind electro-osmosis is our use of nonlinear boundary conditions for Faradaic electron-transfer reactions, assuming Butler–Volmer kinetics and a compact Stern layer. We also analyze the current-voltage relation, thus extending our analogous results for thin films below the limiting current [9].

1. Statement of problem. Before delving into the analysis (and to make the paper self-contained), we review governing equations and boundary conditions. We shall focus solely on the dimensionless formulation of the problem, derived and discussed in the companion paper [9].

The transport of cations and anions is described by the steady Nernst–Planck equations

$$\frac{d^2c_+}{dx^2} + \frac{d}{dx}\left(c_+\frac{d\phi}{dx}\right) = 0, \tag{1}$$

$$\frac{d^2c_-}{dx^2} - \frac{d}{dx}\left(c_-\frac{d\phi}{dx}\right) = 0, \tag{2}$$

while Poisson's equation relates the electric potential to the charge density,

$$-\epsilon^2\frac{d^2\phi}{dx^2} = \frac{1}{2}\left(c_+ - c_-\right). \tag{3}$$

Here ϵ is a small dimensionless parameter equal to the ratio of the Debye screening length to the electrode separation (or film thickness). Note that this formulation assumes constant material properties, such as diffusivity, mobility, and dielectric coefficient, and neglects any variations which may occur at large electric fields. The factor of $1/2$ multiplying the charge density $c_+ - c_-$ is present merely for convenience. The domain for the system of (1)–(3) is $0 < x < 1$.

The two Nernst–Planck equations are easily integrated under the physical constraint that the boundaries are impermeable to anions (i.e., zero flux of anions at $x = 0$) and taking the nondimensional current density at the electrodes to be $4j$:

$$\frac{dc_+}{dx} + c_+\frac{d\phi}{dx} = 4j, \tag{4}$$

$$\frac{dc_-}{dx} - c_-\frac{d\phi}{dx} = 0. \tag{5}$$

Then by introducing the average ion concentration and (half) the charge density,

$$c = \frac{1}{2}\left(c_+ + c_-\right) \quad \text{and} \quad \rho = \frac{1}{2}\left(c_+ - c_-\right), \tag{6}$$

we can derive a more symmetric form for the coupled PNP equations:

$$\frac{dc}{dx} + \rho\frac{d\phi}{dx} = 2j, \tag{7}$$

$$\frac{d\rho}{dx} + c\frac{d\phi}{dx} = 2j, \tag{8}$$

$$-\epsilon^2\frac{d^2\phi}{dx^2} = \rho. \tag{9}$$

For this system of one second-order and two first-order differential equations, we require four boundary conditions and one integral constraint:

$$\phi(0) - \delta\epsilon\frac{d\phi}{dx}(0) = 0, \tag{10}$$

$$\phi(1) + \delta\epsilon\frac{d\phi}{dx}(1) = v, \tag{11}$$

$$k_c\left[c(0)+\rho(0)\right]e^{\alpha_c\phi(0)} - j_r e^{-\alpha_a\phi(0)} = j, \tag{12}$$

$$-k_c\left[c(1)+\rho(1)\right]e^{\alpha_c(\phi(1)-v)} + j_r e^{-\alpha_a(\phi(1)-v)} = j, \tag{13}$$

$$\int_0^1 \left[c(x)-\rho(x)\right]dx = 1. \tag{14}$$

These conditions, which are often simplified or omitted in electrochemical modeling, are central to our analysis. A detailed discussion can be found in the companion paper [9], so here we simply give an overview.

The first two boundary conditions, (10)–(11), account for the intrinsic capacitance of the compact part of the electrode-electrolyte interface, which is taken to be linear (the "Stern model"). The compact-layer charge could contain solvated ions at the point of closest approach to the electrode, as well as adsorbed ions on the surface. The capacitance also accounts for the dielectric polarization of the solvation layer and/or impurities or coatings on the surface. In these boundary conditions, δ is a dimensionless parameter which measures the strength of the surface capacitance, and v is the total dimensionless voltage drop across the cell.

The next two boundary conditions, (12)–(13), are Butler–Volmer rate equations, which represent the kinetics of Faradaic electron-transfer reactions at each electrode, with an Arrhenius dependence on the compact-layer voltage. In these equations, k_c and j_r are dimensionless reaction-rate constants and α_c and α_a are transfer coefficients for the electrode reaction. It is worth noting that α_c and α_a do not vary too much from system to system; typically they have values between 0 and 1, and often both take on values near 1/2.

Finally, the integral constraint, (14), reflects the fact that the total number of anions is fixed, assuming that anions are not allowed to leave the electrolyte by Faradaic processes or specific adsorption. When solving time-dependent problems with the same mathematical model [23, 24], the constraint is not needed, since the total number of anions is set by the initial condition. Here, we solve for the steady state at different voltages (and currents), assuming the same average concentration of anions to allow a meaningful comparison for the same cell.

It is important to understand that the need for an extra constraint reflects that the current-voltage relationship, $j(v)$, or "polarographic curve," is not given a priori. As usual in one-dimensional problems [9], it is easier to assume galvanostatic forcing at fixed current, j, and then solve for the cell voltage, $v(j)$, by applying the boundary

condition (11), rather than the more common case of potentiostatic forcing at fixed voltage, v. For this reason, we take the former approach in our analysis. For steady-state problems, the two kinds of forcing are equivalent and yield the same (invertible) polarographic curve, $j(v)$ or $v(j)$.

For some of our analysis, it will be convenient to further simplify the problem by introducing the dimensionless electric field, $E \equiv -\frac{d\phi}{dx}$. This transformation is useful because three of the five independent constraints can be expressed in terms of these variables, without explicit dependence on $\phi(x)$, namely, the two Butler–Volmer rate equations,

$$k_c \left(c(0) + \rho(0)\right) e^{-\alpha_c \delta \epsilon E(0)} - j_r e^{\alpha_a \delta \epsilon E(0)} = j, \tag{15}$$

$$-k_c \left(c(1) + \rho(1)\right) e^{\alpha_c \delta \epsilon E(1)} + j_r e^{-\alpha_a \delta \epsilon E(1)} = j, \tag{16}$$

and the integral constraint on the total number of anions, (14). The potential is recovered by integrating the electric field and applying the Stern boundary conditions (10) and (11).

2. Unified analysis at all currents.

2.1. Master equation for the electrostatic potential. We begin our analysis by reducing the governing equations, (7) through (9), to a single master equation for the electrostatic potential. Substituting (9) into (7) and integrating, we obtain an expression for the average concentration:

$$c(x) = \underline{c}_o + 2jx + \frac{\epsilon^2}{2}\left(\frac{d\phi}{dx}\right)^2. \tag{17}$$

Then by applying the integral constraint, (14), we find that the integration constant, $\underline{c}_o$, is given by

$$\underline{c}_o = (1-j) - \epsilon^2 \left[\left(\frac{d\phi}{dx}\right)\Big|_{x=1} - \left(\frac{d\phi}{dx}\right)\Big|_{x=0} + \frac{1}{2}\int_0^1 \left(\frac{d\phi}{dx}\right)^2 dx \right]. \tag{18}$$

Note that when the electric field is $O(1)$, (17) and (18) reduce to the leading-order concentration in the bulk when j is sufficiently below the limiting current [9]. We can now eliminate ρ and c from (8) to arrive at a single master equation for ϕ,

$$\epsilon^2 \left[-\frac{d^3\phi}{dx^3} + \frac{1}{2}\left(\frac{d\phi}{dx}\right)^3 \right] + \left(\underline{c}_o + 2jx\right)\frac{d\phi}{dx} = 2j, \tag{19}$$

or, equivalently, for the electric field E,

$$\epsilon^2 \left[\frac{d^2E}{dx^2} - \frac{1}{2}E^3 \right] - \left(\underline{c}_o + 2jx\right) E = 2j. \tag{20}$$

Once this equation is solved, the concentration, c, and charge density, ρ, are computed using (17) and Poisson's equation, (9).

The master equation has been derived in various equivalent forms since the 1960s. Grafov and Chernenko [25] first combined (4), (5), and (9) to obtain a single nonlinear differential equation for the anion concentration, c_-, whose general solution they expressed in terms of Painlevé's transcendents. The master equation for the

electric field, (20), was first derived by Smyrl and Newman [15] in the special case of the classical limiting current, where $j = 1$ and $\underline{c}_o = 0$, where they discovered a nonequilibrium double layer of width $\epsilon^{2/3}$, which is apparent from the form of the master equation. We shall study the general electric-field and potential equations for an arbitrary current, j, focusing on boundary-layer structure in the limiting and superlimiting regimes.

2.2. Efficient numerical solution. To solve the master equation for the electric field with the boundary conditions and integral constraint, we use the Newton–Kantorovich method [26]. Specifically, we use a Chebyshev pseudospectral discretization to solve the linearized boundary-value problem at each iteration [26, 27]. Our decision to use this method is motivated by its natural ability to resolve boundary layers and its efficient use of grid points. We are able to get accurate results for many parameter regimes very quickly (typically less than a few minutes on a workstation) with only a few hundred grid points, which would not be possible at large currents and/or thin double layers using a naive finite-difference scheme. It is important to stress that the boundary conditions and the integral constraint are explicitly included as part of the Newton–Kantorovich iteration. Therefore, the linear boundary-value problem solved in each iteration is actually an *integrodifferential equation* with boundary conditions that are integroalgebraic equations.

To ensure convergence at high currents, we use continuation in the current density parameter, j, and start with a sufficiently low initial j that the bulk electroneutral solution is a reasonable initial guess; often, initial j values relatively high compared to the diffusion-limited current are acceptable. After a small increase in current, we check that the iteration converges to a correspondingly small perturbation of the previous solution. Analogous continuation in the δ parameter is also sometimes necessary to compute solutions at high δ values.

The results of the numerical method are presented in the figures below and in [9] to test our analytical approximations obtained by asymptotic analysis.

2.3. Recovery of classical results below the limiting current, $j \ll 1 - O(\epsilon^{2/3})$. In the low-current regime, the master equation admits the two distinguished limits around $x = 0$ that arise in the classical analysis: $x = O(1)$ and $x = O(\epsilon)$. When $x = O(1)$, we find the usual bulk electric field from (19) and the bulk concentration from (17). When $x = O(\epsilon)$, the master equation can be rescaled using $x = \epsilon y$ to obtain

$$-\frac{d^3\phi}{dy^3} + \frac{1}{2}\left(\frac{d\phi}{dy}\right)^3 + \underline{c}_o\frac{d\phi}{dy} + 2jy\epsilon\frac{d\phi}{dy} = 2j\epsilon, \tag{21}$$

which is equivalent to the classical theory at leading order [9]. In particular, the Gouy–Chapman structure of the double layer can be derived directly from the Smyrl–Newman equation in this limit [23].

The anode boundary layer comes from a similar $O(\epsilon)$ scaling around $x = 1$. Note that in the $j \ll 1 - \epsilon^{2/3}$ regime, the scaling $x = O(\epsilon^{2/3})$ is *not* a distinguished limit because the $\underline{c}_o(\frac{d\phi}{dx})$ term would dominate all other terms in (19).

3. Nested boundary layers at the limiting current, $j = 1 - O(\epsilon^{2/3})$. In this section, we show that a nontrivial nested boundary-layer structure emerges at the classical limiting current when general boundary conditions are considered.

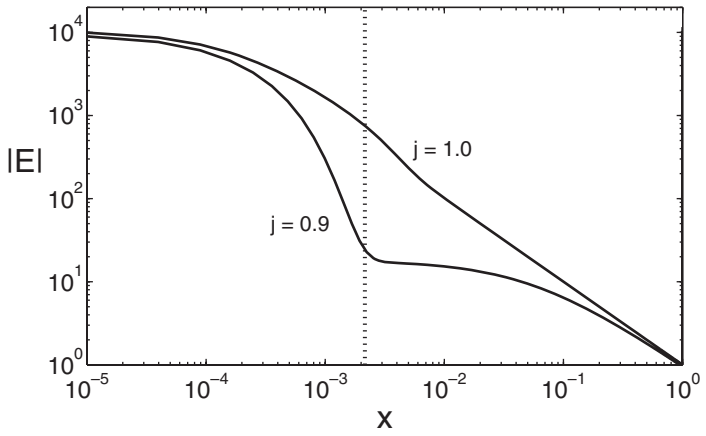


FIG. 2. *Numerical solutions for the dimensionless electric field $E(x)$ at current densities of $j = 0.9$ and $j = 1.0$ demonstrating the expansion of the diffuse layer at the limiting current ($k_c = 1$, $j_r = 2$, $\delta = 0.1$, and $\epsilon = 0.0001$). For reference, the vertical line shows where $x = \epsilon^{2/3}$.*

3.1. Expansion of the double layer out of equilibrium. As discussed in the companion paper [9], the classical analysis breaks down as the current approaches the diffusion-limited current, $j \to 1$. One sign of the problem is that the charge density at $j = 1$ grows near the cathode ($x \to 0$),

$$\rho = \epsilon^2 \frac{d^2\phi}{dx^2} \sim \frac{\epsilon^2}{x^2}. \tag{22}$$

The classical assumption of charged boundary layers of $O(\epsilon)$ width, therefore, fails because the charge density, $\rho = O(1)$, would be much larger than the salt concentration, $c \sim 2x = O(\epsilon)$, at $x = O(\epsilon)$, which violates bulk electroneutrality. This paradox, noted by Levich [14], was resolved by Smyrl and Newman [15], who realized that the structure of the double layer must change near the classical limiting current. In particular, the width of the diffuse part expands to $x = O(\epsilon^{2/3})$, beyond which the bulk charge density remains small, $\rho = O(\epsilon^{2/3})$, as shown in Figure 2. Here, we revisit this problem with more general boundary conditions and also consider currents above the classical limiting current.

Mathematically, the classical asymptotics fails because a new distinguished limit for the master equation appears as $j \to 1$. Rescaling the master equation using $x = \epsilon^{2/3} z$ gives us

$$-\frac{d^3\phi}{dz^3} + \frac{1}{2}\left(\frac{d\phi}{dz}\right)^3 + \frac{\underline{c}_o}{\epsilon^{2/3}}\frac{d\phi}{dz} + 2jz\frac{d\phi}{dz} = 2j, \tag{23}$$

which implies that we have a meaningful distinguished limit if $\underline{c}_o = O(\epsilon^{2/3})$ or, equivalently, $j = 1 - O(\epsilon^{2/3})$. In this regime, the double layer is no longer in Poisson–Boltzmann equilibrium at leading order, and the potential satisfies the more general equation, (23), for $z = O(1)$ or $x = O(\epsilon^{2/3})$.

Unfortunately, at this scale, *all* terms in (23) are $O(1)$, so we are forced to solve the full equation. Although general solutions can be expressed in terms of Painlevé's transcendents [8, 18, 25], these are not convenient for applying our nonlinear boundary conditions or obtaining physical insight. Even when $\underline{c}_o = o(\epsilon^{2/3})$, we are left with a

complicated differential equation which does not admit a simple analytical solution. However, in the case $\underline{c}_o = o(\epsilon^{2/3})$, it is possible to study the asymptotic behavior of the solution in the limits $z \to 0$ and $z \to \infty$ by considering the behavior of the *neighboring* asymptotic layers.

3.2. Nested boundary layers when $|1 - j| = o(\epsilon^{2/3})$. The appearance of the new distinguished limit for $j = 1 - O(\epsilon^{2/3})$ does not destroy the ones that exist in the classical analysis. In particular, the $O(\epsilon)$ boundary layer at $x = 0$ does *not* vanish. This inner layer was overlooked by Smyrl and Newman because they assumed a fixed surface charge density given by the equilibrium zeta potential [15], rather than more realistic boundary conditions allowing for surface charge variations and electrochemical reactions.

In the general case, a set of nested boundary layers must exist when the current is near (or above) the classical limiting current. For convenience, we shall refer to the $x = O(\epsilon^{2/3})$ and the $x = O(\epsilon)$ regions as the "Smyrl–Newman" and "inner diffuse" layers, respectively. It is important to realize that, without the inner layer, it would be impossible to satisfy any reasonable boundary conditions describing the electrochemical reactions which support the current. In the Smyrl–Newman layer, the concentration of the active species (here, cations) nearly vanishes at the limiting current, since $\underline{c}_0 = O(\epsilon^{2/3})$, but this would imply a very small reaction rate density. The paradox of the original Smyrl–Newman solution (which ignores reactions) is that there are very few ions available at the cathode, and yet there is a very large reaction rate and current. The resolution involves an inner layer where the cation concentration increases to $O(1)$.

In the context of our model of electrochemical reactions, we can also understand the nested boundary layers on mathematical grounds. Consider the reaction boundary condition at the cathode, (12). To estimate the c and ρ at the electrode surface, we rescale (17) and Poisson's equation using $x = \epsilon^{2/3} z$ to obtain

$$c = \underline{c}_o + 2j\epsilon^{2/3} z + \frac{\epsilon^{2/3}}{2}\left(\frac{d\phi}{dz}\right)^2, \tag{24}$$

$$\rho = -\epsilon^{2/3}\frac{d^2\phi}{dz^2}, \tag{25}$$

which means that the concentration and charge density are both $O(\epsilon^{2/3})$ since $\underline{c}_o = o(\epsilon^{2/3})$ when $|1 - j| = o(\epsilon^{2/3})$. Then, from the Stern boundary condition, we have $\phi(0) = -\delta\epsilon\bar{E} = -\delta\epsilon^{1/3}\grave{E} = O(\delta\epsilon^{1/3})$. Plugging these estimates into the reaction boundary condition, we find

$$k_c O(\epsilon^{2/3}) e^{\alpha_c \delta \epsilon^{1/3}\grave{E}(0)} = j + j_r e^{-\alpha_a \delta\epsilon^{1/3}\grave{E}(0)} = O(1). \tag{26}$$

This equation cannot be satisfied in the limit $\epsilon \to 0$ with $\delta \geq 0$ fixed, which implies the existence of the inner diffuse layer. In the Gouy–Chapman model without any compact layer ($\delta = 0$), (26) reduces to a contradiction, $O(\epsilon^{2/3}) = j = \text{constant}$, and thus implies the existence of the inner diffuse layer. In the Stern model ($\delta > 0$), it can only be satisfied for very large values, $\delta = O(|\log \epsilon^{2/3}|/\epsilon^{1/3})$, but, since δ is fixed, the nested inner layer must appear as $\epsilon \to 0$. However, this calculation predicts that the magnitude of the concentration at the cathode (within the inner layer) decreases with increasing δ, which is clearly seen in the numerical solutions of Figure 3.

To analyze (23), it is convenient to focus on the electric field rather than the

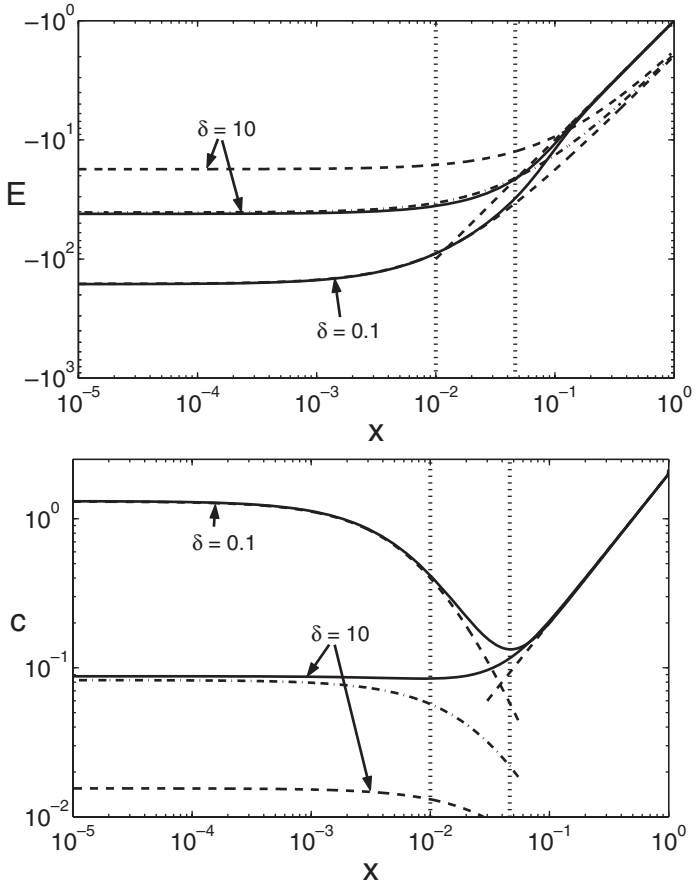


FIG. 3. *Numerical solutions (solid lines) for the dimensionless electric field $E(x)$ and concentration $c(x)$ at the classical diffusion-limited current ($j = 1$) compared with leading-order asymptotic approximations (dashed and dot-dashed lines) for $k_c = 1$, $j_r = 2$, $\epsilon = 0.01$, and $\delta = 0.1, 10$. The leading-order bulk approximations for $E(x)$ and $c(x)$ are given by* (28) *and $c(x) = 2jx$, respectively. In the diffuse layer, the leading-order approximations are given by* (30) *and* (31). *For the $\delta = 10$ curves, the difference between the dashed and dot-dashed curves is that the dashed curve uses an approximate value for B given by* (36), *while the dot-dashed curve uses a B value calculated by numerically solving* (33). *For reference, the vertical lines show where $x = \epsilon$ and $x = \epsilon^{2/3}$. The thin anode diffuse-layer field is not shown.*

potential. In terms of the scaled electric field, $\grave{E}(z) \equiv -\frac{d\phi}{dz} = \epsilon^{2/3} E(x)$, (23) becomes

$$\frac{d^2\grave{E}}{dz^2} - \frac{1}{2}\grave{E}^3 - 2j(z\grave{E} + 1) = \frac{\mathrm{c}_o}{\epsilon^{2/3}}\grave{E}, \tag{27}$$

which we shall refer to as the "Smyrl–Newman equation." From (71) in [9], we know that the first few terms in the expansion for the bulk electric field at the limiting current are

$$\begin{aligned} -\bar{E}(x) &= \frac{1}{x} + \frac{3\epsilon^2}{4x^4} + \frac{111\epsilon^4}{16x^7} + \frac{6045\epsilon^6}{32x^{10}} + \cdots \\ &= \frac{1}{\epsilon^{2/3}}\left(\frac{1}{z} + \frac{3}{4z^4} + \frac{111}{16z^7} + \frac{6045}{32z^{10}} + \cdots\right). \end{aligned} \tag{28}$$

Since the second series is asymptotic for $z \gg 1$, the expansion in the bulk is valid for $x \gg \epsilon^{2/3}$. In order to match the solution in the Smyrl–Newman layer to the bulk, we expect the asymptotic solution to (27) as $z \to \infty$ to be given by the expression in parentheses in (28). We could also have arrived at this result by directly substituting an asymptotic expansion in $1/z$ and matching coefficients. As we can see in Figure 3 the leading-order term in (28) is a good approximation to the exact solution in the bulk and is matched by the solution in the Smyrl–Newman layer as it extends into the bulk.

We now turn our attention towards the "inner diffuse" layer, which gives us the asymptotic behavior of the Smyrl–Newman equation in the limit $z \to 0$. Introducing the scaled variables $y = x/\epsilon = z/\epsilon^{1/3}$ and $\breve{E} = \epsilon\bar{E} = \epsilon^{1/3}\grave{E}$, (27) becomes

$$\frac{d^2\breve{E}}{dy^2} - \frac{1}{2}\breve{E}^3 - 2j\epsilon(y\breve{E} + 1) = \underline{c}_o\breve{E}. \tag{29}$$

Near the limiting current (i.e., $\underline{c}_o = O(\epsilon^{2/3})$), $\breve{E}$ satisfies $\frac{d^2\breve{E}}{dy^2} = \frac{1}{2}\breve{E}^3$ at leading order with the boundary condition $\breve{E} \to 0$ as $y \to \infty$ from the matching condition that $\grave{E}$ remains bounded as $z \to 0$. Integrating this equation twice with the observation that $\frac{d\breve{E}}{dy} > 0$ gives us

$$\breve{E}(y) \sim -\frac{2}{y+b}, \tag{30}$$

where b is a constant determined by applying the Butler–Volmer reaction boundary condition at the cathode. We can estimate $\breve{c}(y)$ and $\breve{\rho}(y)$ by substituting (30) into (17) and Poisson's equation to find

$$\breve{c}(y) = \underline{c}_o + 2jx + \frac{\epsilon^2}{2}\bar{E}(x)^2 = \underline{c}_o + 2j\epsilon y + \frac{1}{2}\breve{E}(y)^2 = \frac{2}{(y+b)^2} + O(\epsilon), \tag{31}$$

$$\breve{\rho}(y) = \epsilon^2\frac{d\bar{E}}{dx} = \frac{d\breve{E}}{dy} = \frac{2}{(y+b)^2} + O(\epsilon). \tag{32}$$

Therefore, b satisfies the following transcendental equation at leading order:

$$k_c\frac{4}{b^2}e^{2\alpha_c\delta/b} = j + j_r e^{-2\alpha_a\delta/b}. \tag{33}$$

While this equation does not admit a simple closed-form solution, we can compute approximate solutions in the limits of small and large δ values. In the small δ limit, we can linearize (33) and expand b in a power series in δ to obtain

$$b \sim 2\sqrt{\frac{k_c}{j+j_r}} + \delta\left(\alpha_c + \frac{\alpha_a j_r}{j+j_r}\right) + O(\delta^2). \tag{34}$$

At the other extreme, for $\delta \gg 1$, (33) can be approximated by

$$k_c\frac{4}{b^2}e^{2\alpha_c\delta/b} \approx j. \tag{35}$$

Then, using fixed-point iteration on the approximate equation, we find that

$$b \sim \frac{2\alpha_c\delta}{\log\kappa - 2\log\log\kappa + O\left(\log\log\log\delta^2\right)}, \tag{36}$$

where $\kappa \equiv j\alpha_c^2\delta^2/k_c$. Figure 3 shows that the leading-order approximation (30) is very good in the inner diffuse layer as long as an accurate estimate for b is used. While the small δ approximation for b is amazingly good (the asymptotic and numerical solutions are nearly indistinguishable), the large δ estimate for b is not as good but is only off by an $O(1)$ multiplicative factor.

Before moving on, it is worth noting that the asymptotic behavior of the concentration and charge density in the Smyrl–Newman layer as $z \to 0$ and $z \to \infty$ suggests that the charge density is low throughout the entire Smyrl–Newman layer. Figure 3 shows how the Smyrl–Newman layer acts as a transition layer, allowing the bulk concentration to become small near the cathode while still ensuring a sufficiently high cation concentration at the cathode surface to satisfy the reaction boundary conditions. The transitional nature of the Smyrl–Newman layer becomes even more pronounced for smaller values of ϵ.

4. Bulk space charge above the limiting current, $1 + O(\epsilon^{2/3}) \ll j \ll O(1/\epsilon)$. As current exceeds the classical limiting value, the overlap region between the inner diffuse and Smyrl–Newman layers grows to become a layer having $O(1)$ width. Following other authors [16, 22], we shall refer to this new layer as the "*space-charge*" layer because, as we shall see, it has a nonnegligible charge density compared to the rest of the bulk. Therefore, in this current regime, the central region of the electrochemical cell is split into two pieces having $O(1)$ width separated by an $o(1)$ transition layer.

In the bulk, the solution remains unchanged except that $\underline{c}_o$ cannot be approximated by $1 - j$; the contribution from the integral term is no longer negligible. The need for this correction arises from the high electric fields required to drive current through the electrically charged space-charge layer. With this minor modification, we find that the bulk solution is

$$\begin{aligned} \bar{c}(x) &= \underline{c}_o + 2jx, \\ \bar{E}(x) &= \frac{1}{x_o - x}, \end{aligned} \tag{37}$$

where $x_o \equiv -\underline{c}_o/2j$ is the point where the bulk concentration vanishes (see Figure 4).

Between the two $O(1)$ layers, there is a small transition layer. Rescaling the master equation using the change of variables $z = (x - x_o)/\epsilon^{2/3}$ and $\acute{E}(z) = \epsilon^{2/3}\bar{E}(x)$, we again obtain the Smyrl–Newman equation, (27), with right-hand side equal to zero. As before, we find that the solution in the transition layer approaches $-1/z$ as $z \to \infty$. In the other direction as $z \to -\infty$, we will find that the appropriate boundary condition is $\acute{E} \to -2\sqrt{j|z|}$ to match the electric field in space-charge layer.

4.1. Structure of the space-charge layer. Physically, we could argue that the concentration of ions in the space-charge layer is very small (i.e., zero at leading order) because the layer is essentially the result of stretching the ionic content of the overlap between the inner diffuse and Smyrl–Newman layers, which is small to begin with, over an $O(1)$ region. This physical intuition is confirmed by the numerical solutions shown in Figures 4, 5, and 6. Therefore, using (17), we obtain the leading-order solution for the electric field

$$\tilde{E} \sim \frac{-2\sqrt{j\,(x_o - x)}}{\epsilon}. \tag{38}$$

Note that the magnitude of the field is exactly what is required to make the integral term in $\underline{c}_o$ an $O(1)$ contribution. From this formula, it is easy to compute the charge

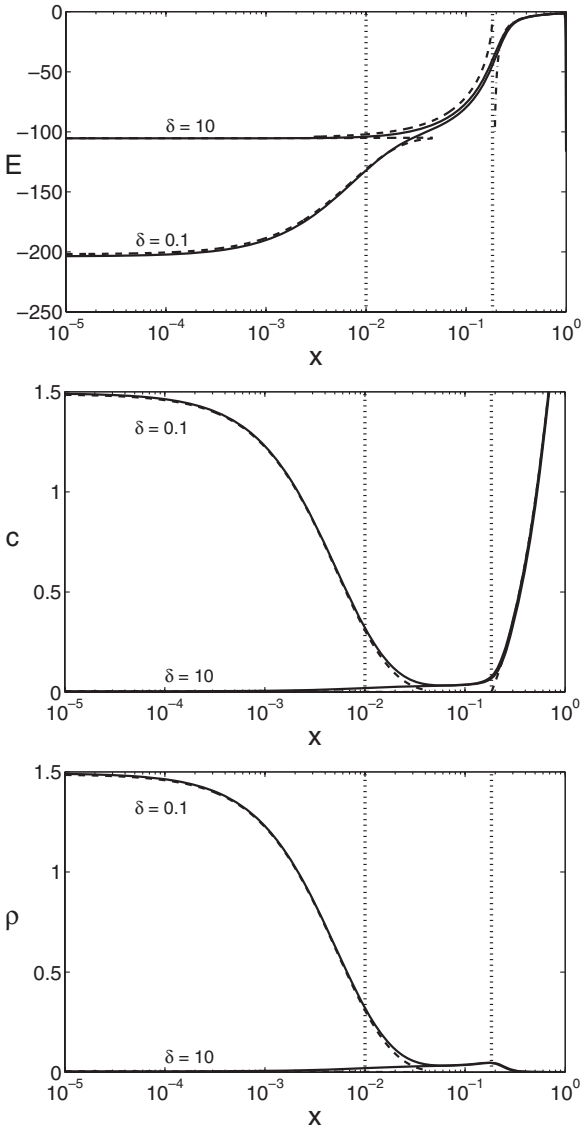


FIG. 4. *Numerical solutions (solid lines) for the dimensionless electric field $E(x)$, average concentration $c(x)$, and charge density $\rho(x)$ above the diffusion-limited current ($j = 1.5$) compared with leading-order asymptotic approximations (dashed lines) for $k_c = 1$, $j_r = 2$, $\epsilon = 0.01$, and $\delta = 0.1, 10$. The leading-order bulk approximations are given by* (37)*. In the space-charge layer, the leading-order electric field is given by* (38)*, and leading-order concentration is* 0*. Finally,* (58) *and* (59) *are the diffuse-layer asymptotic approximations for the electric field and concentration, respectively. For reference, the vertical lines show where $x = \epsilon$ and $x = x_o$.*

density in the space-charge layer:

$$\tilde{\rho} = \epsilon^2 \frac{d\tilde{E}}{dx} \sim \epsilon \sqrt{\frac{j}{x_o - x}}, \tag{39}$$

which is an order of magnitude larger than the $O(\epsilon^2)$ charge density in the bulk. The $O(\epsilon)$ charge density also implies that the concentration must be at least $O(\epsilon)$ because

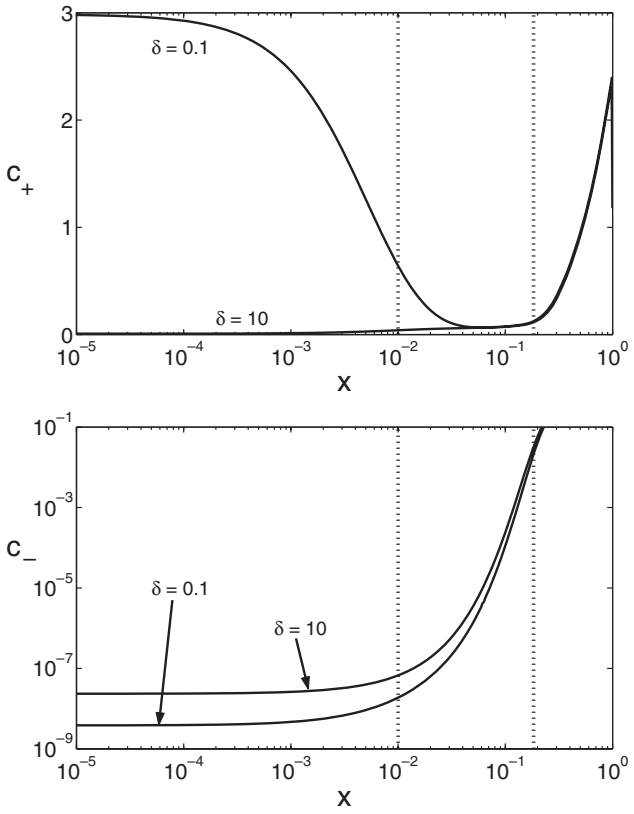


FIG. 5. *Numerical solutions for the dimensionless cation and anion concentrations above the diffusion-limited current ($j = 1.5$) for $k_c = 1$, $j_r = 2$, $\epsilon = 0.01$, and $\delta = 0.1, 10$. For reference, the vertical lines show where $x = \epsilon$ and $x = x_o$.*

the anion concentration, $c - \rho$, is positive.

With the electric field given by (38), we can determine the values of x_o and $\underline{c}_o$ by solving the system of equations given by the definition of x_o and $\underline{c}_o$. Using (18) to calculate $\underline{c}_o$ and noticing that the leading-order contribution to the integral comes from the space-charge layer, we obtain

$$\underline{c}_o \sim 1 - j\left(1 + x_o^2\right). \tag{40}$$

Combining this result with $x_o = -\underline{c}_o/2j$, we find that

$$x_o \sim 1 - j^{-1/2}, \qquad \underline{c}_o \sim 2(j^{1/2} - j), \tag{41}$$

which can be substituted into (37) and (38) to yield the leading-order solutions in the bulk and space-charge layers. It should be noted that the expression for x_o is consistent with the estimate for the width found by Bruinsma and Alexander [21] and Chazaviel [22] in the limits $j - 1 \ll 1$ and small space-charge layer ($x_o \ll 1$), although our analysis also applies to much larger voltages.

The results obtained via physical arguments in the previous few paragraphs motivate an asymptotic series expansion for E whose leading-order term is $O(1/\epsilon)$. Moreover, because we want to be able to balance the current density at second order, we

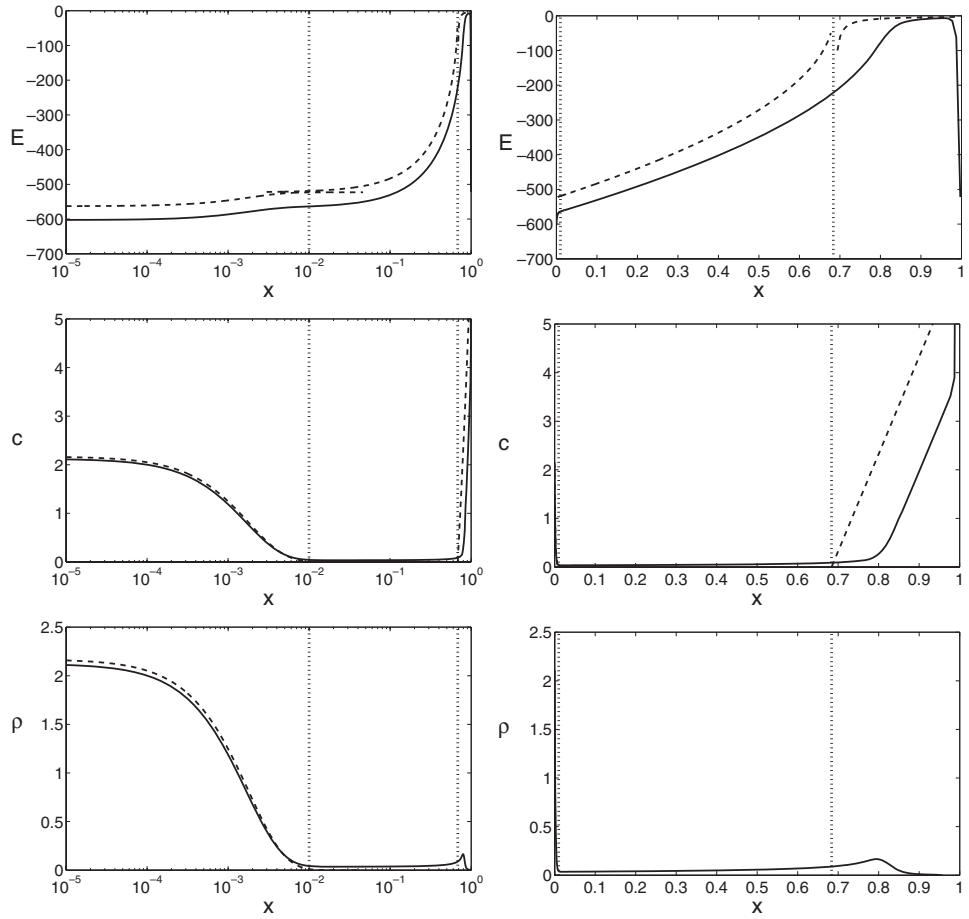


FIG. 6. *Numerical solutions (solid lines) for the dimensionless electric field $E(x)$, average concentration $c(x)$, and charge density $\rho(x)$ far above the diffusion-limited current ($j = 10.0$) compared with leading-order asymptotic approximations (dashed lines) for $k_c = 1$, $j_r = 2$, $\epsilon = 0.01$, and $\delta = 0.1$. Each field is shown twice:* (1) *with x on* log *scale to focus on the cathode region and* (2) *with x on a linear scale to emphasize the interior of the cell. Note that $j\epsilon = 0.1$, so the asymptotic approximations are not as good as at lower current densities. For reference, the vertical lines show where $x = \epsilon$ and $x = x_o$.*

choose the second-order term to be $O(j)$. Thus, we have

$$\tilde{E} = \frac{1}{\epsilon} E_{-1} + E_0 j + \cdots . \tag{42}$$

Note that in this asymptotic series, the first term dominates the second term only as long as $j \ll 1/\epsilon$, so the following analysis holds exclusively for current densities far below $O(1/\epsilon)$. Figure 6 illustrates the breakdown of the leading-order asymptotic solutions at very high current densities. While the qualitative features of the asymptotic approximation are correct (e.g., the shape of $E(x)$ in the diffuse layer and the slope of $c(x)$ in the bulk), the quality of the approximation is clearly less than at lower values of j.

The key advantage of a more systematic asymptotic analysis is that we are able to calculate the leading-order behavior of the space-charge layer concentration $\tilde{c}$, which

is not possible with only knowledge of the leading-order behavior for the electric field. Substituting (42) into the master equation (20), it is straightforward to obtain

$$\tilde{E} \sim -\frac{2}{\epsilon}\sqrt{j\,(x_o - x)} - \frac{1}{2\,(x_o - x)} + \cdots . \tag{43}$$

Using this expression in (17), we find the dominant contribution to $\tilde{c}$ is exactly the same as $\tilde{\rho}$:

$$\tilde{c} \sim \epsilon\sqrt{\frac{j}{x_o - x}}. \tag{44}$$

Since $c_- = c - \rho$, this result leads to an important physical conclusion: *The space-charge layer is essentially depleted of anions*, $c_- = o(\epsilon)$, as is clearly seen in Figures 4 and 5. This contradicts our macroscopic intuition about electrolytes, but, in very thin films, complete anion deplection might occur. For example, in a microbattery developed for on-chip power sources using the $Li/SiO_2/Si$ system, lithium ion conduction has recently been demonstrated in nanoscale films of silicon oxide, where there should not be any counterions or excess electrons [6].

At leading order as $\epsilon \to 0$, the anion concentration, c_-, can be set to zero in the space-charge layer, leaving the following two governing equations:

$$\frac{dc_+}{dx} + c_+\frac{d\phi}{dx} = 4j, \tag{45}$$

$$-\epsilon^2\frac{d^2\phi}{dx^2} = \frac{1}{2}c_+. \tag{46}$$

As with the binary electrolyte case, these equations can be reduced to a single equation for the electric potential:

$$\frac{d^3\phi}{dx^3} + \frac{d^2\phi}{dx^2}\frac{d\phi}{dx} = -\frac{2j}{\epsilon^2}. \tag{47}$$

Integrating this equation once, we obtain a Riccati equation for $\frac{d\phi}{dx}$:

$$\frac{d^2\phi}{dx^2} + \frac{1}{2}\left(\frac{d\phi}{dx}\right)^2 = -\frac{2j}{\epsilon^2}\,(x - x_o) + h, \tag{48}$$

where h is an integration constant. Using the transformations

$$u \equiv e^{\phi/2}, \qquad z \equiv -\frac{j^{1/3}}{\epsilon^{2/3}}\,(x - x_o) + \frac{\epsilon^{4/3}h}{2j^{2/3}}, \tag{49}$$

we find that u satisfies Airy's equation,

$$\frac{d^2u}{dz^2} - zu = 0. \tag{50}$$

Thus, the general solution for $\phi(x)$ is

$$\phi(x) = 2\log\left[a_1 Ai\left(\frac{j^{1/3}}{\epsilon^{2/3}}\,(x_o - x) + \beta h\right) + a_2 Bi\left(\frac{j^{1/3}}{\epsilon^{2/3}}\,(x_o - x) + \beta h\right)\right], \tag{51}$$

where a_1 and a_2 are constants determined by boundary conditions and $\beta = \frac{\epsilon^{4/3}}{2j^{2/3}}$.

To simplify this expression, note that in the limit $\epsilon \to 0$, the potential drop between $x = x_o$ and $x = 0$ is approximately

$$\phi(x_o) - \phi(0) \sim 2\log\left[\frac{a_1 Ai(0) + a_2 Bi(0)}{a_1 Ai\left(\frac{x_o j^{1/3}}{\epsilon^{2/3}}\right) + a_2 Bi\left(\frac{x_o j^{1/3}}{\epsilon^{2/3}}\right)}\right]. \tag{52}$$

Now, using the large argument behavior of the Airy functions, we see that as $\epsilon \to 0$, the argument of the logarithm approaches zero. Thus, we are lead to the conclusion that the electric potential at $x = x_o$ is less than at $x = 0$. However, this is completely inconsistent with our physical intuition and the numerical results, which show that $\phi(x_o) - \phi(0) > 0$. Therefore, it must be the case that $a_2 \approx 0$, so that

$$\phi(x) = 2\log\left[a_1 Ai\left(\frac{j^{1/3}}{\epsilon^{2/3}}(x_o - x) + \beta h\right)\right] \tag{53}$$

and

$$E(x) = \frac{2j^{1/3}}{\epsilon^{2/3}} \frac{Ai'\left(\frac{j^{1/3}}{\epsilon^{2/3}}(x_o - x) + \beta h\right)}{Ai\left(\frac{j^{1/3}}{\epsilon^{2/3}}(x_o - x) + \beta h\right)}. \tag{54}$$

In principle, the integration constants h and a_1 can be determined by matching to the inner diffuse layer, $x = O(\epsilon)$ (described below), and the bulk transition layer, $|x_0 - x| = O(\epsilon^{2/3})$ (described above). Here, the main point is that the leading-order approximation for the electric field when the region is depleted of anions is exactly (38), which follows from the asymptotic form of $Ai(z)$ and $Ai'(z)$ as $z \to \infty$ in (54). The equivalence of the single-ion equations and the full governing equations at leading order mathematically confirms the physically interpretation of the space-charge layer as a region of anion depletion.

4.2. Boundary layers above the limiting current. To complete our analysis of the high-current regime, $1 + O(\epsilon^{2/3}) \ll j \ll O(1/\epsilon)$, we must consider the boundary layers. At the anode, all fields are $O(1)$, so we recover the usual Gouy–Chapman solution with the minor modification that $c_1 = 2\sqrt{j}$, which is the value $\bar{c}$ takes as $x \to 1$. The cathode structure, however, is much more interesting because it is depleted of anions (see Figure 5). To our knowledge, this nonequilibrium inner boundary layer on the space-charge region, related to the reaction boundary condition at the cathode, has not been analyzed before.

As in the space-charge layer, the leading-order governing equations in this layer are those of a single ionic species with no counterions (45) and (46). Rescaling those equations using $x = \epsilon y$, we obtain

$$\frac{d\check{c}_+}{dy} + \check{c}_+ \frac{d\check{\phi}}{dy} = 4j\epsilon \approx 0, \tag{55}$$

$$-\frac{d^2\check{\phi}}{dy^2} = \frac{1}{2}\check{c}_+. \tag{56}$$

From these equations, it is immediately clear that the cations have a Boltzmann equilibrium profile at leading order: $c_+ \propto e^{-\phi(y)}$. As in the analysis for the space-charge layer, it is possible to find a general solution to (55) and (56). By combining

these equations and integrating, we find that the potential in the cathode boundary layer has the form

$$\check{\phi} \sim \log\left[\sinh^2(py+q)\right] + r, \tag{57}$$

where p, q, and r are integration constants. Therefore, the electric field and concentration are

$$\check{E}(y) \sim -2p\coth(py+q), \tag{58}$$

$$\check{c}(y) = \frac{1}{2}\check{c}_+(y) \sim \frac{2p^2}{\sinh^2(py+q)}. \tag{59}$$

Matching the electric fields in the diffuse and space-charge layers, we find that $p \sim \sqrt{jx_o}$. Note that because $p = O(\sqrt{j})$, the electric field in the diffuse charge layer is $O(\sqrt{j}/\epsilon)$, which is the same order of magnitude as in the space-charge layer. To solve for q, we use the expression for p in the cathode Stern and Butler–Volmer boundary conditions, which leads to the following nonlinear equation:

$$\frac{4k_c j x_o}{\sinh^2 q}\exp\left(2\alpha_c\delta\sqrt{jx_o}\coth q\right) - j_r\exp\left(-2\alpha_a\delta\sqrt{jx_o}\coth q\right) = j. \tag{60}$$

In the limit of small δ, we can use fixed-point iteration to obtain an approximate solution,

$$q \sim \sinh^{-1}\left(2\sqrt{\frac{k_c j x_o\exp\left(2\alpha_c\delta\sqrt{jx_o}\coth q_o\right)}{j + j_r\exp\left(-2\alpha_a\delta\sqrt{jx_o}\coth q_o\right)}}\right), \tag{61}$$

where q_o has the same form as q with $(\coth q_o)$ set equal to 1. For $\delta \gg 1$, the leading-order equation is

$$\frac{4k_c j x_o}{\sinh^2 q}\exp\left(2\alpha_c\delta\sqrt{jx_o}\coth q\right) \sim j, \tag{62}$$

which implies that $q \gg 1$, so that the left-hand side can be small enough to balance the current. Thus, by using $\coth q \approx 1$ and $\sinh q \approx \exp(q)/2$, we find that $q \sim \alpha_c\delta\sqrt{jx_o} + \frac{1}{2}\log(16k_c x_o)$. The agreement of these asymptotic approximations with the numerical solutions in the diffuse charge layer is illustrated in Figure 4.

5. Polarographic curves. We are now in a position to compute the leading-order behavior of the polarographic curve at and above the classical limiting current. Recall that the formula for the cell voltage is given by

$$v = -\delta\epsilon E(0) + \int_0^1 -E(x)dx - \delta\epsilon E(1). \tag{63}$$

The integral is the voltage drop through the interior of the cell, and the first and last terms account for the potential drop across the Stern layers.

At the limiting current, $j = 1$, we can estimate the voltage drop across the cell by using the bulk and diffuse-layer electric field to approximate the field in the Smyrl–Newman transition layer to obtain

$$v \sim -\delta\epsilon E(0) + \int_0^{\epsilon^{2/3}} -E(x)dx + \int_{\epsilon^{2/3}}^1 -E(x)dx - \delta\epsilon E(1) \tag{64}$$

$$\sim 2\frac{\delta}{b} + 2\log\left(\frac{\epsilon^{-1/3}+b}{b}\right) - \frac{2}{3}\log\epsilon. \tag{65}$$

Notice that in the small δ limit, this expression reduces to $v \sim -\frac{4}{3}\ln\epsilon$ as $\epsilon \to 0$. The

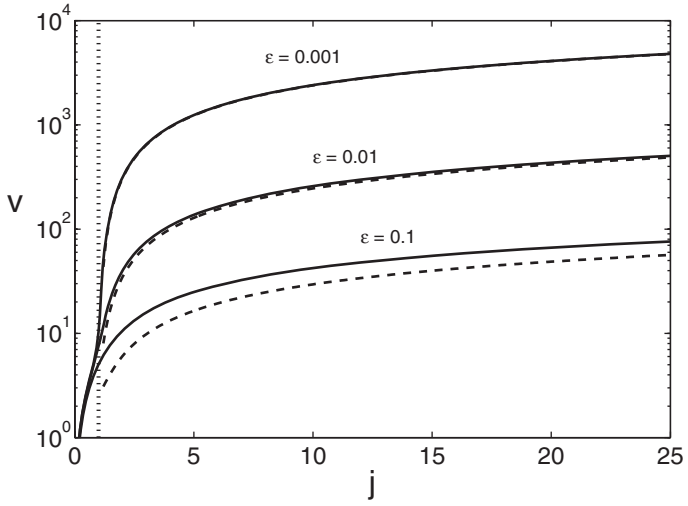


FIG. 7. *Comparison of numerical polarographic curves (dashed lines) with leading-order asymptotic approximations (solid lines) given in* (66) *for several values of* ϵ *with* $\delta = 1.0$, $k_c = 1$, *and* $j_r = 2$. *For* $\epsilon = 0.001$, *the numerical and asymptotic polarographic curves are indistinguishable on this graph. For reference, the vertical dashed line shows the classical diffusion-limited current* $j = 1$.

TABLE 1
Comparison of the asymptotic approximations (65) *and* (66) *with numerically calculated values for the cell voltage at various* ϵ *and* δ *values. These cell voltages were computed with* $k_c = 1$ *and* $j_r = 2$.

		$j = 1.0$		$j = 1.5$	
ϵ	δ	v_{exact}	v_{asym}	v_{exact}	v_{asym}
1e-4	0.01	13.125	12.101	1297.799	1289.621
1e-4	1.00	13.222	12.374	1297.048	1291.101
1e-4	10.0	14.290	13.571	1305.318	1300.129
1e-3	0.01	10.165	9.146	140.207	132.790
1e-3	1.00	10.277	9.475	139.450	134.270
1e-3	10.0	11.552	10.890	147.717	143.299
1e-2	0.01	7.339	6.303	22.434	15.725
1e-2	1.00	7.479	6.729	21.624	17.206
1e-2	10.0	9.228	8.465	29.886	26.234
1e-1	0.01	4.922	3.649	9.479	2.637
1e-1	1.00	5.005	4.219	7.790	4.118
1e-1	10.0	7.995	6.327	16.088	13.146

dependence, $v(j = 0) \propto \ln \epsilon$, is clear in the numerical polarographic curves shown in Figure 7. (See also Figure 4 of the companion paper [9].) Table 1 compares this approximation with the exact cell voltage for a few ϵ and δ values. For small ϵ values ($\epsilon \leq 0.01$), the asymptotic approximations are fairly good (within 5% to 10%).

Above the limiting current, the space-charge layer makes the dominant contribution to the cell voltage. Using (37) and (38) in the formula for the cell voltage, we find that

$$v \sim \frac{4\sqrt{j}}{3\epsilon}\left(1 - j^{-1/2}\right)^{3/2} + 2\delta\left(j - \sqrt{j}\right)^{1/2}\coth q - \frac{1}{2}\log j - 2/3\log\epsilon. \tag{66}$$

The first two terms in this expression estimate the voltage drop across the space-charge and the cathode Stern layers, respectively. The last two terms are the subdominant

contribution from the bulk where we have somewhat arbitrarily taken $x = x_o + \epsilon^{2/3}$ as the boundary between the bulk layer and the Smyrl–Newman transition layer. Notice that we ignore the contribution from the cathode diffuse and Smyrl–Newman layers. It is safe to neglect the diffuse layer because it is an $O(1)$ contribution. However, the Smyrl–Newman layer has a nonnegligible potential drop that we have to accept as error since we do not have an analytic form for the solution in that region.

Figure 7 shows that the asymptotic polarographic curves are quite accurate for sufficiently small ϵ values. In Table 1, we compare the results predicted by the asymptotic formula with numerical results for a few specific values of ϵ and δ. It is interesting that the approximation is also better for large δ values (we explain this observation in the next section). Also, while the $\log \epsilon$ term is subdominant, it makes a significant contribution to the cell voltage for ϵ values as small as 0.01.

As with the width of the space-charge layer, x_o, our expression for the cell voltage, (66), is consistent with the results of Bruinsma and Alexander [21] and Chazaviel [22] near the limiting current, $j \to 1^+$, while remaining valid at much larger currents, $j = O(1/\epsilon)$.

6. Effects of the Stern-layer capacitance. The inclusion of the Stern layer in the boundary conditions allows us to explore the effects of the intrinsic surface capacitance on the structure of the cell. From Figures 3 through 5, we can see that smaller Stern-layer capacitances (i.e., larger δ values) decrease the concentration and electric-field strength in the cathode diffuse layer. This behavior arises primarily from the influence of the electric field on the chemical kinetics at the electrode surfaces. When the capacitance of the Stern layer is low, small electric fields at the cathode surface translate into large potential drops across the Stern layer, (10), which help drive the deposition reaction, (12). As a result, neither the electric field nor the cation concentration need to be very large at the cathode to support high-current densities. These results confirm our physical intuition that it is only important to pay attention to the diffuse layer when the Stern-layer potential drop is negligible (i.e., $\delta \ll 1$).

At high currents, another important effect of the Stern-layer capacitance is that the total cell voltage becomes dominated by the potential drop across the Stern layer at large δ values (i.e., small capacitances). This behavior is clearly illustrated in Figure 8. Notice that for currents below the classical diffusion-limited current, the total cell voltage does not show a strong dependence on δ. However, for $j > 1$, the total cell voltage increases with δ—the increase being driven by the strong δ dependence of the Stern voltage.

7. Conclusion. In summary, we have studied the classical problem of direct current in an electrochemical cell, focusing on the exotic regime of high-current densities. A notable new feature of our study is the use of nonlinear Butler–Volmer and Stern boundary conditions to model a thin film passing a Faradaic current, as in a microbattery. We have derived leading-order approximations for the fields at and above the classical, diffusion-limited current, paying special attention to the structure of the cathodic boundary layer, which must be present to satisfy the reaction boundary conditions. In our analysis of superlimiting current, we have shown that the key feature of the bulk space-charge layer is the depletion of anions. Our exact solution of the leading-order problem in the space-charge region, (51), could thus also have relevance for Faradaic conduction through very thin insulating films.

Using the asymptotic approximations to the fields, we are able to derive a current-voltage relation, (66), which compares well with numerical results, far beyond the limiting current. Combined with the analogous formulae in the companion paper [9],

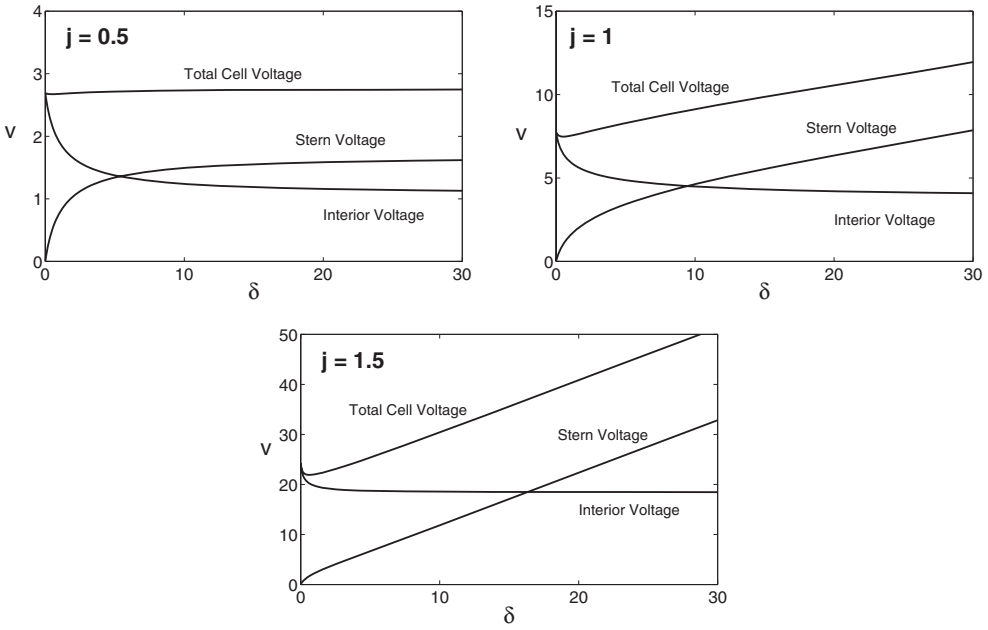


FIG. 8. *These graphs break the total cell voltage into contributions from the cell interior and the Stern layer as a function of* δ *for* $\epsilon = 0.01$, $k_c = 2$, *and* $j_r = 2$. *Note that at and above the classical limiting current, the Stern-layer voltage dominates the total cell voltage for large values of* δ.

which hold below the limiting current, we have essentially analyzed the full range of the current-voltage relation. These results could be useful in interpreting experimental data, e.g., on the internal resistance of thin-film microbatteries.

A general conclusion of this study is that boundary conditions strongly affect the solution. For example, the Stern-layer capacitance, often ignored in theoretical analysis, plays an important role in determining the qualitative structure of the cell near the cathode, as well as the total cell voltage. The nonlinear boundary conditions for Butler–Volmer reaction kinetics also profoundly affect charge distribution and current-voltage relation, compared to the ubiquitous case of Dirichlet boundary conditions. The latter rely on the assumption of surface equilibrium, which is of questionable validity at very large currents.

We leave the reader with a word of caution. The results presented here are valid mathematical solutions of standard model equations, but their physical relevance should be met with some skepticism under extreme conditions, such as superlimiting current. For example, the PNP equations are meant to describe infinitely dilute solutions in relatively small electric fields [7, 28, 29]. Even for quasi-equilibrium double layers, their validity is not so clear when the zeta potential greatly exceeds the thermal voltage, because co-ion concentrations may exceed the physical limit required by discreteness (accounting also for solvation shells) and counterion concentrations may become small enough to violate the continuum assumption. Large electric fields can cause the permittivity to vary, by some estimates up to a factor of ten, as solvent dipoles become aligned. Including such effects, however, introduces further ad hoc parameters into the model, which may be difficult to infer from experimental data. Instead, we suggest using our analytical results (especially current-voltage relations) to test the validity of the basic model equations in thin-film experiments.

Acknowledgments. The authors thank M. Brenner, J. Choi, and B. Kim for helpful discussions.

REFERENCES

[1] N. J. Dudney, J. B. Bates, D. Lubben, and F. X. Hart, *Thin-film rechargeable lithium batteries with amorphous $Li_xMn_2O_4$ cathodes*, in Thin Film Solid Ionic Devices and Materials, J. Bates, ed., The Electrochemical Society, Pennington, NJ, 1995, pp. 201–214.

[2] B. Wang, J. B. Bates, F. X. Hart, B. C. Sales, R. A. Zuhr, and J. D. Robertson, *Characterization of thin-film rechargeable lithium batteries with lithium cobalt oxide cathodes*, J. Electrochem. Soc., 143 (1996), pp. 3204–3213.

[3] B. J. Neudecker, N. J. Dudney, and J. B. Bates, *"Lithium-free" thin-film battery with* in situ *plated Li anode*, J. Electrochem. Soc., 147 (2000), pp. 517–523.

[4] N. Takami, T. Ohsaki, H. Hasabe, and M. Yamamoto, *Laminated thin Li-ion batteries using a liquid electrolyte*, J. Electrochem. Soc., 149 (2002), pp. A9–A12.

[5] Z. Shi, L. Lü, and G. Ceder, *Solid State Thin Film Lithium Microbatteries*, Singapore-MIT Alliance Technical Report: Advanced Materials for Micro- and Nano-Systems Collection, 2003; available online from http://hdl.handle.net/1721.1/3672.

[6] N. Ariel, G. Ceder, D. R. Sadoway, and E. A. Fitzgerald, *Electrochemically-controlled transport of lithium through ultra-thin SiO2 for novel electronic and optoelectronic devices*, submitted for publication.

[7] J. Newman, *Electrochemical Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1991.

[8] I. Rubinstein, *Electro-Diffusion of Ions*, SIAM Stud. Appl. Math. 11, SIAM, Philadelphia, PA, 1990.

[9] M. Z. Bazant, K. T. Chu, and B. J. Bayly, *Current-voltage relations for electrochemical thin films*, SIAM J. Appl. Math., 65 (2005), pp. 1463–1484.

[10] A. A. Chernenko, *The theory of the passage of direct current through a solution of a binary electrolyte*, Dokl. Akad. Nauk SSSR, 153 (1962), pp. 1129–1131 (in Russian).

[11] J. Newman, *The polarized diffuse double layer*, Trans. Faraday Soc., 61 (1965), pp. 2229–2237.

[12] A. D. MacGillivray, *Nernst-Planck equations and the electroneutrality and Donnan equilibrium assumptions*, J. Chem. Phys., 48 (1968), pp. 2903–2907.

[13] W. Nernst, *Theorie der Reaktionsgeschwindigkeit in heterogenen Systemen*, Z. Phys. Chem., 47 (1904), pp. 52–55.

[14] V. G. Levich, *Physico-Chemical Hydrodynamics*, Prentice-Hall, London, 1962.

[15] W. H. Smyrl and J. Newman, *Double layer structure at the limiting current*, Trans. Faraday Soc., 63 (1967), pp. 207–216.

[16] I. Rubinstein and L. Shtilman, *Voltage against current curves of cation exchange membranes*, J. Chem. Soc. Faraday Trans. II, 75 (1979), pp. 231–246.

[17] S. S. Dukhin, *Electrokinetic phenomena of the second kind and their applications*, Adv. Colloid Interface Sci., 35 (1991), pp. 173–196.

[18] Y. Ben and H.-C. Chang, *Nonlinear Smoluchowski slip velocity and micro-vortex generation*, J. Fluid Mech., 461 (2002), pp. 229–238.

[19] I. Rubinstein and B. Zaltzman, *Electro-osmotically induced convection at a permselective membrane*, Phys. Rev. E (3), 62 (2000), pp. 2238–2251.

[20] I. Rubinstein and B. Zaltzman, *Electro-osmotic slip of the second kind and instability in concentration polarization at electrodialysis membranes*, Math. Models Methods Appl. Sci., 11 (2001), pp. 263–299.

[21] R. Bruinsma and S. Alexander, *Theory of electrohydrodynamic instabilities in electrolytic cells*, J. Chem. Phys., 92 (1990), pp. 3074–3085.

[22] J.-N. Chazalviel, *Electrochemical aspects of the generation of ramified metallic electrodeposits*, Phys. Rev. A (3), 42 (1990), pp. 7355–7367.

[23] A. Bonnefont, F. Argoul, and M. Z. Bazant, *Analysis of diffuse-layer effects on time-dependent interfacial kinetics*, J. Electroanal. Chem., 500 (2001), pp. 52–61.

[24] M. Z. Bazant, K. Thornton, and A. Ajdari, *Diffuse charge dynamics in electrochemical systems*, Phys. Rev. E (3), 70 (2004), article 021506.

[25] B. M. Grafov and A. A. Chernenko, *Theory of the passage of a constant current through a solution of a binary electrolyte*, Dokl. Akad. Nauk SSSR, 146 (1962), pp. 135–138 (in Russian).

[26] J. P. Boyd, *Chebyshev and Fourier Spectral Methods*, Dover, Mineola, NY, 2001.

[27] L. N. Trefethen, *Spectral Methods in MATLAB*, SIAM, Philadelphia, PA, 2000.

[28] P. Delahay, *Double Layer and Electrode Kinetics*, Interscience, New York, 1965.

[29] A. J. Bard and L. R. Faulkner, *Electrochemical Methods*, John Wiley & Sons, New York, 2001.

SIAM J. APPL. MATH.
Vol. 65, No. 5, pp. 1506–1525

THREE-DIMENSIONAL PROBABILITY DENSITY FUNCTIONS VIA TOMOGRAPHIC INVERSION*

JULES S. JAFFE†

Abstract. In many experimental observation systems where the goal is to record a three-dimensional observation of an object, or a set of objects, a lower-dimensional projection of the intended subject is obtained. In some situations only the statistical properties of such objects are desired: the three-dimensional probability density function. This article demonstrates that under special symmetries this function can be obtained from either a one- or two-dimensional probability density function which has been obtained from the observed, projected data. Standard tomographic theorems can be used to guarantee the uniqueness of this function, and a natural basis set can be used in computing the three-dimensional function from the one- or two-dimensional projection. The theory of this inversion is explored using theoretical and computational methods with examples of data taken from scientific experiments.

Key words. tomographic inversion, probability

AMS subject classifications. 15A29, 15A09, 60G10

DOI. 10.1137/S003613990342390X

Introduction. In many experimental observation systems whose goal is to record a "true" three-dimensional observation of an object, or a set of objects, a lower-dimensional projection of the intended subject is obtained. In general, the tomographic inversion theorems for many such systems have been known for some time and have been used to expedite the inversion of projected data sets: one- and two-dimensional projections of such structures to obtain the three-dimensional data (Herman, 1980; Kak and Slaney, 1987). For the most part, the data collection and subsequent inversions have considered the reconstruction of individual three-dimensional structures whose unique features are desired.

An alternate problem which utilizes the exact same tomographic theory arises in a situation where the interest is in some statistical parameterization of a set of three-dimensional objects and where the data that have been measured are lower-dimensional projections of these three-dimensional structures. This is true in certain situations such as the x-ray diffraction of molecules in suspension, where the molecules are assumed to be randomly oriented and the observed diffraction patterns are a result of the incoherent superposition of the magnitude of the Fourier transforms of the various orientations. In this case, the correlation distances between atoms are the desirable parameter, and in favorable situations the relationships between the ensemble of projections and the observed data can be used to infer the average distances between such entities (Hukins, 1981).

Another example concerns the case treated in this article: the recording of the displacements of animals using either optical or acoustic systems. In this situation, one type of desired knowledge is the probability density function for velocities $pdf(\vec{v})$, that is to say, the probability that a given animal is moving with velocity between $\vec{v}$ and $\vec{v}+d\vec{v}$, where $\vec{v}$ is a three-dimensional velocity. Assuming that the system is temporally

*Received by the editors March 10, 2003; accepted for publication (in revised form) November 13, 2004; published electronically May 12, 2005.
http://www.siam.org/journals/siap/65-5/42390.html
†Marine Physical Laboratory, Scripps Institution of Oceanography, La Jolla, CA 92093-0238 (jules@mpl.ucsd.edu).

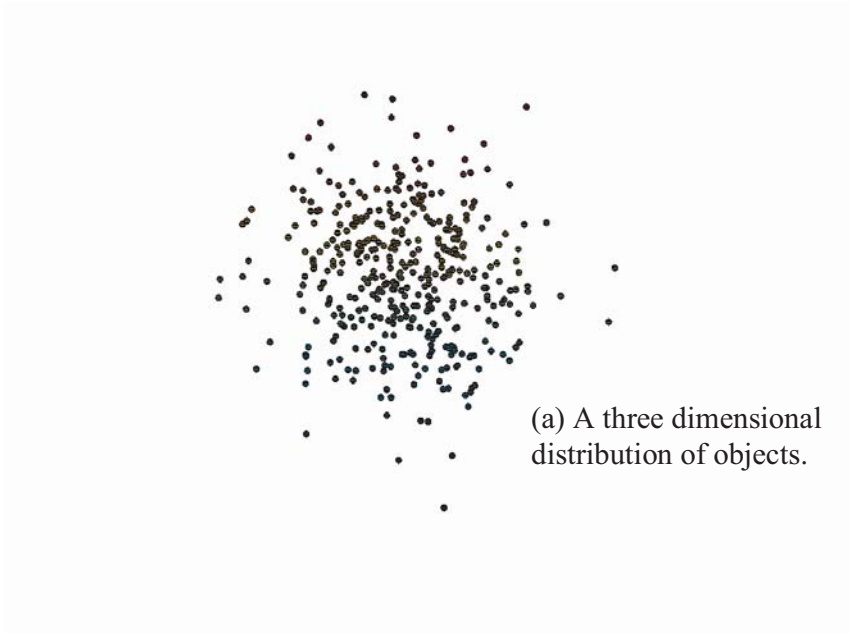


FIG. 1. *A diagram which illustrates the utility of the methodology developed.* (a) *shows a three-dimensional distribution of objects. The distribution can be measured with either* (b) *an essentially one-dimensional sonar system producing a record of animal reflectivity along a line or* (c) *an image of the same phenomenon observed with a camera system.*

stationary, this statistic provides an interesting parameterization of the state of the system. Valuable ecological information can be inferred from such a function which includes the metabolic energy level of the animal (Torres and Childress, 1983) as well as a knowledge of the animal's lifestyle and the factors which govern its success in either foraging or mating.

Two special cases that are treated in the article are illustrated in Figure 1. The native environment for many observational phenomena is three-dimensional. However, in many cases it is much easier to measure either one- or two-dimensional data sets in which the three-dimensional information is embedded. As one example, we remark that the inferences about the behavior of animal aggregations (swarms, flocks)

(Okubo, 1980; Parrish and Edelstein-Keshet, 1999; Parrish and Hamner, 1997) could benefit from the methodology described herein. The figure depicts reflections from an essentially one-dimensional sonar system (b) and images from a camera (c). In the case of the sonar, a very narrow beam can be used to measure the range-dependent locations of animals along a single line. In the case of the camera, the projected two-dimensional locations of the animals can be inferred.

Mathematical formulation. The starting point for our discussion of the algorithm concerns the existence of a data matrix of object positions as a function of time: $\mathcal{R} = \{\vec{r}_{i,j}\}$, where the i,jth element corresponds to the three-dimensional position of object i observed at time t_j. Considering this matrix, a set of displacements can be computed for a set of objects at two time instants, $\{t_j, t_k\}$ as $\{\Delta\vec{r}_i = \vec{r}_{ij} - \vec{r}_{ik} | t_j - t_k, i = 1, M\}$, where there are M objects. Computing this set of displacements over all objects and over all time intervals and assuming temporal stationarity allows the computation of a set of displacements as a function of time interval $\Delta\vec{r}(\Delta T) = \{\Delta\vec{r}_i | \Delta T, i = 1, M\}$. Dividing this data matrix by the time difference $\Delta T = t_j - t_k$ provides an estimate of the instantaneous velocities of the objects. However, we choose to leave the data as positions, as they are a bit easier to visualize.

Forming a histogram of the object displacements provides an estimate of the underlying probability density function for object displacement $pdf(\Delta\vec{r}|\Delta T)$ for a fixed time interval, ΔT. Dividing three-dimensional space into small three-dimensional boxes of dimensions, $\delta\Delta\vec{r}$, the approximate three-dimensional probability density function, derived from the measured data, can be computed by assuming, given $n(\Delta\vec{r})$ as the number of occurrences of the length $\Delta\vec{r}$ in the interval $\Delta\vec{r} - \frac{\delta\Delta\vec{r}}{2} < \Delta\vec{r} \leq \Delta\vec{r} + \frac{\delta\Delta\vec{r}}{2}$, that

$$pdf_{3d}(\Delta\vec{r}|\Delta T) = \lim_{N,t\to\infty} \frac{n(\Delta\vec{r})}{N}.$$

We next consider the data collection process and the consequences of a system that can measure only a subset of the vector components of the range displacement. In one case examined here a set of narrow sonar beams constituted the measurement which transformed the three-dimensional data vector into one that considered only range. A more common occurrence occurs when a conventional optical camera is used to obtain a two-dimensional picture of a set of three-dimensional objects. In this case, a conventional camera system provides a projected view of the animals' three-dimensional positions. In either case, a linear transformation projects the higher-dimensional data onto the measured lower-dimensional coordinates, resulting in a loss of information. For a single object or scene, a multitude of views can be obtained, resulting in a set of projections which can be used to invert for the true three-dimensional object (Kak and Slaney, 1987). In the special case when the object has some inherent degree of symmetry, the structure of the object can be obtained from a reduced set of data. So, for example, in the case of an object which displays complete three-dimensional symmetry, the three-dimensional structure can be obtained from a single projection.

The purpose of this article is to extend the theory of the reconstruction of projected functions to the case of probability density functions. As such, given an ensemble of objects with features $\{\Delta\vec{r}\}$, a probability density function $pdf_{3d}(\Delta\vec{r})$ is used to describe their characteristics. In our case, the resultant measurement of such objects is a new set of objects with reduced dimensionality. This set of objects has lower dimensionality, say ρ (a scalar), and a new and different probability density function

$pdf_{1d}(\rho)$ (in the one-dimensional case) is obtained which is dependent on measurement geometry.

Defining a new matrix $\mathcal{D} = \{\rho_{ij}\}$, where the i,jth element is the projected position of object i at time t_j, the treatment above can be extended in a similar way to obtain an estimate for the probability density function of this one-dimensional function. Given $n(\Delta\rho)$ as the number of occurrences of the length $\Delta\rho$ in the interval $\Delta\rho - \frac{\delta\Delta\rho}{2} < \Delta\rho \leq \Delta\rho + \frac{\delta\Delta\rho}{2}$, then

$$pdf_{1d}(\Delta\rho|\Delta T) = \lim_{N,t\to\infty} \frac{n(\Delta\rho)}{N}.$$

Although we suspect that this theory can be generalized to measurements other than displacements, only the straightforward relationships between the displacements in one, two and three dimensions and their respective probability density functions will be addressed here. In the case of displacement, under the assumption of three-dimensional isotropy, the inversion from $pdf_{3d}(\Delta\rho)$ to $pdf_{3d}(\Delta\vec{r})$ can be formulated using the theory of reconstruction of functions from their projections. The next section demonstrates that the one-dimensional probability density function obtained this way is a projection of the three-dimensional probability density function. Additional theorems are proved relating to our specific formulation as well.

Theorems.

The spherically symmetric case. Our first proof examines the relationship between the three-dimensional probability density function for a fixed time delay ΔT, $pdf_{3d}(\Delta\vec{r}|\Delta T)$, and a one-dimensional projection of it, $pdf_{1d}(\Delta\rho|\Delta T)$. Figure 2 illustrates that the formulation concerns the existence of two sets of data and their respective probability density functions. The two data sets are the "true" three-dimensional set of object displacements which would be obtained from a "true" three-dimensional imaging system $\{\Delta\vec{r}|\Delta T\}$. The measured data $\{\Delta\rho|\Delta T\}$ is obtained via the measurement process $M_{3d\to 1d}$ as shown in the diagram (as, for example, using a pencil sonar beam). The probability density functions, $pdf_{3d}(\Delta\vec{r}|\Delta T)$ and $pdf_{1d}(\Delta\rho|\Delta T)$ are obtained from each of these data sets via the binning transformations as above and represent the underlying statistics of the processes from one point of view. $Pdf_{3d\to 1d}$ is the resultant transformation that occurs when one computes $pdf_{1d}(\Delta\rho|\Delta T)$ from the inherent three-dimensional probability density function by projecting it to a one-dimensional function.

$$\begin{array}{ccc} & M_{3d\to 1d} & \\ \{\Delta\vec{r}|\Delta T\} & \Rightarrow & \{\Delta\rho|\Delta T\} \\ \Downarrow & & \Downarrow \\ & Pdf_{3d\to 1d} & \\ pdf_{3d}(\Delta\vec{r}|\Delta T) & \Rightarrow & pdf_{1d}(\Delta\rho|\Delta T) \end{array}$$

FIG. 2. *A diagram illustrating the relationships (clockwise from upper right) between the data* $\{\Delta\vec{r}|\Delta T\}$*, the projected data* $\{\Delta\rho|\Delta T\}$*, the probability density function for the one-dimensional data* $pdf_{1d}(\Delta\rho|\Delta T)$*, and the probability density function for the three-dimensional data* $pdf_{3d}(\Delta\vec{r}|\Delta T)$*. The transformations* $M_{3d\to 1d}$ *and* $pdf_{1d}(\Delta\rho|\Delta T)$ *are the projection operator on the data set and the projection operator on the three-dimensional data.*

The first theorem concerns the consistency of the above diagram. That is, the arrows indicate that the one-dimensional probability density function can be obtained

in one of two ways: by recording the three-dimensional data and computing the one-dimensional pdf via a projection of the three-dimensional probability density function, or by recording the one-dimensional data and computing the probability density function from this scalar data set. The two methods produce equivalent functions as indicated by the theorem.

THEOREM 1. *The one-dimensional probability density function obtained from the projected data is identical to that obtained via a projection of the three-dimensional probability density function.*

Proof. The question asked here is whether the diagram in Figure 2 "commutes." To demonstrate that this is so, assume an arbitrary three-dimensional probability density function $pdf_{3d}(\Delta\vec{r}|\Delta T)$ and consider a finite realization of this process in that a set of N three-dimensional vectors has been observed, $\{\Delta\vec{r}_1, \Delta\vec{r}_2, \ldots, \Delta\vec{r}_N\}$. From the definition of a probability density function, $pdf_{3d}(\Delta\vec{r}_i) = \lim_{N\to\infty} \frac{n(\Delta\vec{r}_i)}{N}$, where $n(\Delta\vec{r}_i)$ has been defined as the number of occurrences of event $\Delta\vec{r}_i$ in the interval $\Delta\vec{r}_i - \frac{\delta\Delta\vec{r}_i}{2} < \Delta\vec{r}_i \leq \Delta\vec{r}_i + \frac{\delta\Delta\vec{r}_i}{2}$. These cells also have dimensions $\delta\Delta x, \delta\Delta y, \delta\Delta z$, with the corresponding definition of the probability density function as above for $\delta\Delta\vec{r}$. Now the transformation $M_{3d\to 1d}$ maps the data vector $\{\Delta\vec{r}|\Delta T\}$ into $\{\Delta\rho|\Delta T\}$. Without loss of generality, consider the measurement system to be able to resolve objects in the direction Δz so that $\{\Delta\vec{r}|\Delta T\} \to \{\Delta z|\Delta T\}$ and the Δx and Δy components are lost in the measurement process. Then, for a given cell $\delta\Delta z$, $n(\Delta z|\Delta T) = \sum_{\Delta y}\sum_{\Delta x} n(\Delta z, \Delta x, \Delta y|\Delta t)$. Dividing by the total number of observations N and taking the limit $N \to \infty$ permits the integration to be approximated as $\sum_{\Delta y}\sum_{\Delta x} \frac{n(\Delta z,\Delta y,\Delta x|\Delta T)}{N} = \int pdf_{3d}(\Delta\vec{r})dxdy$, which is the marginal probability density function $pdf_{1d}(\Delta z)dxdy$, the projection of $pdf_{3d}(\Delta r)$ onto the z axis. In the limit as $N \to \infty$ the marginal distribution can be computed either directly from the three-dimensional probability density function or by projecting the data to one dimension and then performing the computation of the probability density function.

We next turn our attention to the class of three-dimensional functions which have some special symmetry. The simplest case is when the three-dimensional probability density function has spherical symmetry. In this case, assuming spherical coordinates, the entire function can be represented by a radial slice so that $pdf_{3d}(\Delta\vec{r}) = pdf_{3d}(\rho)$, where ρ is the distance from the origin. Although the ultimate interest is in reconstructing such functions from a single projection, we pause briefly to state and prove a simple theorem.

THEOREM 2. *Given a separable and isotropic probability density function, the projection of the three-dimensional function is identical to the one-dimensional probability density function.*

This theorem is almost trivial to prove; however, we present it for completeness and because it also illustrates the well-known and important fact that if the three-dimensional data are normally distributed, no inversion is necessary, as the one-dimensional probability density function and the three-dimensional probability density function are identical.

So, for example, if (for fixed ΔT)

$$pdf_{3d}(\rho) = pdf_{1d}(x)pdf_{1d}(y)pdf_{1d}(z), \tag{1}$$

then integrating with respect to y and z yields

$$pdf_{3d}(\rho) = pdf_{1d}(x)\int pdf_{1d}(y)dy \int pdf_{1d}(z)dz, \tag{2}$$

so that

$$pdf_{3d}(\rho) = pdf_{1d}(x). \tag{3}$$

As an example, consider a set of object displacements that are normally distributed in three dimensions with mean 0 and variance σ. The three-dimensional probability density function of the displacements can be represented as

$$pdf_{3d}(\Delta\vec{r}) = \frac{1}{2\pi\sigma^2} \exp^{-\frac{\Delta x^2+\Delta y^2+\Delta z^2}{2\sigma^2}}. \tag{4}$$

In this case, the projection of this function onto a one-dimensional axis (here taken to be the z axis with no loss of generality since this function is isotropic) is the marginal probability distribution function

$$pdf_{1d}(\Delta z) = \iint \frac{1}{2\pi\sigma^2} \exp^{-\frac{\Delta x^2+\Delta y^2+\Delta z^2}{2\sigma^2}} dxdy, \tag{5}$$

which is equivalent to

$$pdf_{1d}(\Delta z) = \frac{1}{2\pi\sigma^2} \exp^{-\frac{\Delta z^2}{2\sigma^2}}. \tag{6}$$

Evidently, since the projection of this normally distributed function onto a single axis results in a normally distributed function with equal variance, there is no need to invert for the "true" three-dimensional function.

We next continue to pursue our interest in the reconstruction of the three-dimensional probability density function from the projected data and hence the one-dimensional probability density function. In particular, our interest is in the invertibility of the projection operator $pdf_{3d\to 1d}$. We remark that this inversion is possible for the general class of centrosymmetric functions and can be obtained via some standard techniques such as the projection slice theorem and its ramifications. Therefore, given $\mathcal{F}$, a one-dimensional Fourier transform, and $\mathcal{H}$ as a Hankel transform, with $\mathcal{H}^{-1}$ as an inverse Hankel transform, the transformation $pdf_{1d}(\rho) \to pdf_{3d}(\rho)$ can be obtained neglecting normalization coefficients as

$$pdf_{3d}(\rho) = \mathcal{H}^{-1}\mathcal{F}[pdf_{1d}(\rho)]. \tag{7}$$

In order to pursue this interest we examine more closely the transformation between the three-dimensional centrosymmetric function and its one dimensional projection:

$$pdf_{3d}(\rho) \leftrightarrow pdf_{1d}(\rho). \tag{8}$$

Initially, assume that the three-dimensional probability density function is unimodal so that

$$pdf_{3d}(\rho) = \delta(\rho - \rho_0). \tag{9}$$

In a physical sense, this probability density function corresponds to a set of three-dimensional translations where the displacement is isotropic, though of a fixed value. Under this assumption the set of projected vector lengths, and hence the distribution of the projected lengths, can be computed. The following theorem expresses the relationship between this simple unimodal distribution and its projection.

THEOREM 3. *If* $pdf_{3d}(\rho) = \delta(\rho - \rho_0)$, *then*

$$pdf_{1d}(\rho) = \frac{\Pi(\frac{\rho}{2\rho_0})}{2\rho_0}, \tag{10}$$

where

$$\Pi(x) = 1 \quad \textit{if } |x| \leq 1/2, \tag{11}$$

$$\Pi(x) = 0 \qquad \textit{otherwise.} \tag{12}$$

As an outline of the proof, we first assume a spherical coordinate system with ϕ in the x, y plane and θ as the angle between a vector and the z axis. Next, we compute the probability distribution function for θ and ϕ which will lead to the construction of a set of three-dimensional isotropic vectors of length ρ_0. Following this, the one-dimensional probability density function $pdf_{1d}(\rho)$ is obtained by taking a set of projections and computing the probability density function. Since the distribution is isotropic, we choose the z axis for convenience.

In order to compute the probability density functions for θ and ϕ we first assume that an ensemble of values has been chosen for $0 \leq \theta' \leq \pi$ and $0 \leq \phi' \leq 2\pi$ that are uniformly distributed on these intervals. We seek a set of two functions $g(\theta')$ and $h(\phi')$ which will transform this θ' and ϕ' into a new set of variables, θ and ϕ. Imagine the end points of a set of three-dimensional vectors whose origin is at the coordinate system origin and of length ρ_0 as being uniformly distributed on the surface of a sphere of radius ρ_0. Points drawn from this isotropic distribution should sample the surface area of the sphere uniformly so that, as a function of θ and assuming that $h(\phi')$ is an identity mapping, an increment in total area A is equal to

$$\Delta A = \int_0^{\theta=g(\theta')} r^2 \sin(\theta) d\theta d\phi = 4\pi r^2 (\theta'/\pi). \tag{13}$$

The integral can be solved for θ so that

$$\cos(g(\theta')) = 1 - \frac{2\theta'}{\pi}, \tag{14}$$

yielding

$$\theta = g(\theta') = \cos^{-1}\left(1 - \frac{2\theta'}{\pi}\right). \tag{15}$$

Taking the projection of this distribution is particularly simple with respect to the z axis, since $z = \rho_0 \cos\theta$:

$$z = \rho_0 \left(1 - \frac{2\theta'}{\pi}\right). \tag{16}$$

Since θ' is uniformly distributed on the interval $0 \leq \theta' \leq \pi$, this implies that z is uniformly distributed on the interval $-\rho_0 < z < \rho_0$, and that

$$\begin{aligned} p(z) = 1/2\rho_0 \quad & \text{if } |z| \leq \rho_0, \\ 0 \qquad & \text{otherwise,} \end{aligned} \tag{17}$$

which proves the theorem.

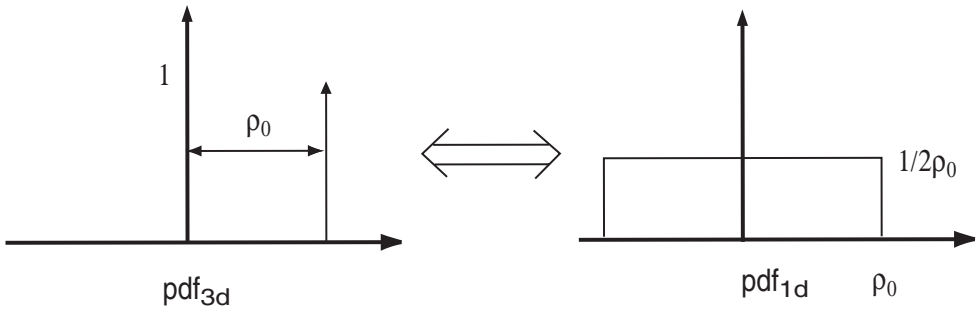


FIG. 3. *The relationship between pdf3d and pdf1d when* $pdf_{3d} = \delta(\rho - \rho_0)$.

We prefer to think of (10) as defining an "impulse response" of the transformation, as illustrated in Figure 3. Since, strictly, the response is a function of position, it is not spatially stationary. Nevertheless, it provides a particularly simple method for inverting the one-dimensional probability density function as shown. Assuming that $pdf_{3d}(\rho)$ can be represented by a set of discrete samples as

$$pdf_{3d}(\rho) = \{pdf_{3d}(\rho_i)|i = 1, N\} = \sum_{i=1}^{N} \int pdf_{3d}(\rho)\delta(\rho - \rho_i)d\rho \tag{18}$$

and noting that the transformation from pdf_{3d} to pdf_{1d} can be made as

$$pdf_{1d}(\rho) = \mathcal{P}[pdf_{3d}(\rho)], \tag{19}$$

where

$$\mathcal{P}\left[\sum_{i=1}^{N} \int pdf_{3d}(\rho)\delta(\rho - \rho_i)d\rho\right] = pdf_{3d}(\rho_i)\frac{\Pi(\frac{\rho}{2\rho_i})}{2\rho_i}, \tag{20}$$

implies that

$$pdf_{1d}(\rho) = \sum_{i=1}^{N} pdf_{3d}(\rho_i)\frac{\Pi(\frac{\rho}{2\rho_i})}{2\rho_i}. \tag{21}$$

This is a system of linear equations which, in principle, can be inverted to obtain an estimate for $pdf_{3d}(\rho)$ from $pdf_{1d}(\rho)$.

An interesting corollary guarantees that these functions are nonincreasing.

COROLLARY. *The one-dimensional probability density function, considered on the positive axis, is always a nonincreasing function.*

The proof follows from Theorem 3 as the final function is a set of nonincreasing functions via (22), and since a sum of nondecreasing functions is nondecreasing, the resultant function is also nondecreasing.

The cylindrically symmetric case. Next, we demonstrate the utility of the general methodology as applied to cylindrically symmetric functions. One example of its use is when a camera system monitors the movements of animals. In this case, a set of two-dimensional projections of the three-dimensional trajectories is obtained. The treatment here is motivated by the fact that many ecosystems can be regarded as

having a distribution of animal motions which are isotropic in the horizontal plane but not the vertical. This is a reasonable assumption when considering some terrestrial and aquatic ecosystems where the force of gravity can permit the animal to differentiate between vertical versus horizontal movement and the environment is azimuthally isotropic in the horizontal plane.

In this case the desired probability density function can be considered to have isotropy in the x, y plane so that the function has cylindrical symmetry, that is, $pdf_{3d}(\rho, \theta, z) = pdf_{3d}(\rho, z)$, where the z, θ, and ρ axes are the ones commonly associated with the cylindrical coordinate system (ρ being the length of a vector in the x, y plane and θ the angle between the vector and the x, y plane). Also, the set of three-dimensional trajectories are projected onto the x, z axes, resulting in the measurement of a set of vectors in the x, z plane. Accordingly, we seek a transform from $pdf_{2d}(x, z)$, the measured data to the "true" probability density function, $pdf_{3d}(\rho, z)$. Note that Δ has been dropped, as it is assumed that the functions are defined on physical variables that obey the geometric requirements of being "projected" via the measurement process.

If ρ and z are independent variables, the three-dimensional probability density function $pdf_{3d}(\rho, z)$ can be expressed as $pdf_{2d}(\rho)pdf_{1d}(z)$, a product of these lower-dimensional functions. In the case where the probability density function cannot be written as a product of the two probability density functions, an alternate strategy for data categorization can be used where the joint probability density function, $pdf_{3d}(\rho, z)$, can be estimated by creating a finite number of bins for the variable z, δz_i, and by observing the probability density function $pdf(x; \delta z_i)$ for each one of these δz_i. In either case, the goal here is to invert for the function $pdf_{2d}(\rho, z)$ given the function $pdf_{1d}(x)$ of measured data. Note that, as before, the one-dimensional function $pdf_{1d}(x)$ of measured (projected) data is not equivalent to $pdf_{3d}(x, 0, 0)$, a slice through the three-dimensional probability density function.

Considering the probability density function only in the plane, we state a theorem similar to Theorem 3, but this time for the two-dimensional to one-dimensional projection.

THEOREM 4. *If* $pdf_{2d}(\rho) = \delta(\rho - \rho_0)$, *then*

$$pdf_{1d}(\rho) = \frac{2}{\pi\rho_0(1 - \frac{\rho^2}{\rho_0^2})^{\frac{1}{2}}} \quad \textit{for } 0 \le \rho \le \rho_0, \tag{22}$$

$$0, \qquad \rho_0 < \rho. \tag{23}$$

The theorem states that if one has a unimodal two-dimensional probability density function which is circularly symmetric in the plane, then the probability density function for the projection of the set of vectors of length ρ_0 will be equal to the above $pdf_{1d}(\rho)$.

For this cylindrically symmetric case, the measurement process will "project" these vectors located in the x, y plane onto the x axis. Here, a linear transformation is derived between the probability density function for the set of projected lengths (uniformly distributed in θ, now assumed to be the angle between the vector ρ and the x axis) and the observed one-dimensional probability density function for their distribution, $pdf_{1d}(x)$.

The proof is as follows: Given a set of vectors in the x, y plane of length ρ_0 and uniformly distributed on $0 \le \theta \le 2\pi$ we seek the $pdf_{1d}(x)$ of their projected lengths. Geometrically, the cumulative distribution function of x in the positive quadrant ($x \ge 0$, $y \ge 0$) for uniform θ is proportional to the length of the arc of a circle

of radius ρ_0 for $0 \leq \theta \leq \pi/2$. Now, considering a new angle, $\theta' = \pi/2 - \theta$ (the angle between the vector and the y axis), the cumulative distribution function for the normalized length of the arc of the circle from $\theta' = 0$ to some value θ' is

$$CDF(\theta') = \frac{2}{\pi\rho_0} \int_0^{\theta'} \rho_0 d\theta = \frac{2\theta'}{\pi}. \tag{24}$$

Transforming back to the original polar angle θ via $\theta' = \frac{\pi}{2} - \theta$ yields

$$CDF(\theta) = 1 - \frac{2\theta}{\pi}. \tag{25}$$

Since $\cos(\theta) = \frac{\rho}{\rho_0}$,

$$CDF(\rho) = 1 - \frac{2}{\pi} \cos^{-1}\left(\frac{\rho}{\rho_0}\right), \tag{26}$$

where we limit

$$0 \leq \cos^{-1}\left(\frac{\rho}{\rho_0}\right) \leq \pi/2. \tag{27}$$

Taking the derivate with respect to x yields the probability density function, so that $pdf_{1d}(\rho) = \delta(\rho - \rho_0)$ is transformed to

$$\frac{2}{\pi\rho_0(1 - \frac{\rho^2}{\rho_0^2})^{\frac{1}{2}}} \quad \text{for} \quad 0 \leq \rho \leq \rho_0, \quad 0 \quad \text{otherwise} \tag{28}$$

by the measurement process.

Interestingly, this function goes to infinity at $\frac{\rho}{\rho_0} = 1$; however, its integral exists and can be used to define the probability of events over any finite interval. The consideration of the other quadrants is accomplished most easily by assuming first that the absolute value of the measurements $|\Delta x|$ are used. This simplifies the bookkeeping needed to keep track of the inverse cosine argument and renders the positive and negative x axes the same. Second, a similar treatment for the negative y axis which defines a new angle which goes from $-\frac{\pi}{2} \leq \theta \leq 0$ yields the same result. Thus, the theorem is proved.

In examining the probability density function, it can be noted that the transformation from two dimensions to one dimension, this time, is somewhat more merciful to the data interpretation in the absence of inversion. So, for example, the most likely value for the projected data is identical to the length of the vector (i.e., when $\frac{\rho}{\rho_0} = 1$). Nevertheless, the transformation does occur and it behooves the experimenter to examine the effects of this projection process in every case. This motivates the development of inversion techniques which will allow the computation of the radially symmetric function from the measured data. As before, a useful view of (22) is as a basis set for the observed data. In this view, the collected data is composed of a superposition of this set of stretched and renormalized functions. The development of inversion techniques that will allow the estimation of $pdf_{2d}(\rho)$ from $pdf_{1d}(\rho)$ will be considered in the next section.

Numerical analysis.

The spherically symmetric case. We next consider the numerical inversion of the set of equations that can be generated from Theorem 3. Assuming discrete

sampling of both the projected data and the proposed inverse solution, a set of matrix equations can be formulated using (21). Assuming the form $\vec{b} = A\vec{x}$, where $\vec{b}$ is the observed data (the one-dimensional probability density function) and $\vec{x}$ is the desired inverse (a radial slice through the three-dimensional probability density function), a numerical inversion can be performed.

A discrete version of the system of equations can be obtained by integrating the observed data $pdf_{1d}(\rho)$ over intervals $\delta\rho$ so that $pdf_{1d}(\delta\rho_j) = \sum_{i=1}^{N} pdf_{3d}(\rho_i)\Pi(\frac{\delta\rho_j}{2\rho_i})$. In addition, considering only positive displacements allows the distribution to be one-sided so that

$$pdf_{1d}(\delta\rho_j) = \sum_{i=1}^{N} pdf_{3d}(\rho_i)\Pi\left(\frac{\delta\rho_j}{\rho_i}\right), \text{ where } \rho_i, \rho_j \geq 0. \tag{29}$$

Assuming bin widths of unit value, the system of equations can be represented as

$$\begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_N \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{2} & . & . & . & \frac{1}{N} \\ & \frac{1}{2} & . & . & . & \frac{1}{N} \\ & & . & . & . & . \\ & & & . & . & . \\ & & & & . & . \\ & & & & & \frac{1}{N} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_N \end{bmatrix}.$$

Since this matrix is upper right triangular, a solution can be written down so that

$$\begin{aligned} x_N &= b_N N, \\ x_{N-1} &= (N-1)(b_{N-1} - b_N), \end{aligned}$$

which implies

$$\begin{aligned} x_i &= i(b_i - b_{i+1}) \quad \text{for } i < N \\ &= ib_i \qquad\qquad\ \ \text{for } i = N. \end{aligned}$$

Thus, the inverse matrix can be written as

$$A^{-1} = \begin{bmatrix} 1 & -1 & & & & \\ & 2 & -2 & & & \\ & & . & & & \\ & & & . & & \\ & & & & (N-1) & -(N-1) \\ & & & & & N \end{bmatrix}.$$

Computing the signal-to-noise of this inversion can be accomplished by adding noise to the observed data vector, $\vec{b}_{obs} = \vec{b} + \vec{\epsilon}$, where $\vec{\epsilon}$ is a noise vector, and then computing the inversion. Substituting back into the expression for x yields a computed $\vec{x}_c$, where

$$x_{c_i} = i((b_i + \epsilon_i) - (b_{i+1} + \epsilon_{i+1})), \tag{30}$$

or

$$x_{c_i} = i(b_{oi} - b_{oi+1}), \tag{31}$$

where b_{oi} and b_{oi+1} are random variables which consist of the observed values of the parameters. Therefore, x_{c_i} is a function of two random variables. If b_{oi} and b_{oi+1} can be described by probability density functions which are independent and, moreover, normally distributed with variance σ_i and σ_{i+1}, then the random variable x_{c_i} has a probability density function which is normal and has variance of

$$\sigma^2_{x_{c_i}} = i^2(\sigma_i^2 + \sigma_{i+1}^2), \tag{32}$$

which demonstrates that the variance increases with the square of value i and is a sum of the variances of the observed data values.

The cylindrically symmetric case. The cylindrically symmetric case can be treated in a manner almost identical to that of the spherically symmetric case. The situation is somewhat more complicated in this case due to the form of the transformation from two-dimensional to one-dimensional probability density function as represented by (22). The two-dimensional probability $pdf_{2d}(\rho)$ can be represented by a discretely sampled version as

$$pdf_{1d}(\rho) = \sum_{i=1}^{N} pdf_{2d}(\rho_i) \frac{2}{\pi\rho_i(1 - \frac{\rho^2}{\rho_i^2})^{\frac{1}{2}}}. \tag{33}$$

As before, we prefer to think of this as a "system response" to a unimodal probability density function ($pdf_{2d}(\rho) = \delta(\rho - \rho_i)$). Integrating this equation over finite bin widths of size $\Delta\rho$ to accommodate the data collection process yields

$$pdf_{1d}(\delta\rho_j) = \int_{\rho_j - \frac{\Delta\rho}{2}}^{\rho_j + \frac{\Delta\rho}{2}} \sum_{i=1}^{N} pdf_{2d}(\rho_i) \frac{2}{\pi\rho_i(1 - \frac{\rho_j^2}{\rho_i^2})^{\frac{1}{2}}} d\rho_j. \tag{34}$$

Taking the integral inside of the sum and noting that

$$\int_{\rho_j - \frac{\Delta\rho}{2}}^{\rho_j + \frac{\Delta\rho}{2}} \frac{2}{\pi\rho_i(1 - \frac{\rho_j^2}{\rho_i^2})^{\frac{1}{2}}} d\rho_j = CDF\left(\rho_j + \frac{\Delta\rho}{2}; \rho_i\right) - CDF\left(\rho_j - \frac{\Delta\rho}{2}; \rho_i\right), \tag{35}$$

where CDF stands for the cumulative distribution function for $pdf_{1d}(\rho)$ as in (26), implies that

$$pdf_{1d}(\delta\rho_j) = \sum_{i=1}^{N} pdf_{2d}(\rho_i) \left(CDF\left(\rho_j + \frac{\Delta\rho}{2}; \rho_i\right) - CDF\left(\rho_j + \frac{\Delta\rho}{2}; \rho_i\right)\right). \tag{36}$$

This equation can then be transformed into a more compact notation as

$$pdf_{1d}(\delta\rho_j) = \sum_{i=1}^{N} pdf_{2d}(\rho_i) \Delta CDF(\rho_j, \rho_i), \tag{37}$$

where

$$\Delta CDF(\rho_j, \rho_i) = CDF\left(\rho_j + \frac{\Delta\rho}{2}; \rho_i\right) - CDF\left(\rho_j + \frac{\Delta\rho}{2}; \rho_i\right) \tag{38}$$

$$= \frac{2}{\pi}\left(\cos\left(\frac{\rho_j - \frac{\Delta\rho}{2}}{\rho_i}\right) - \cos\left(\frac{\rho_j + \frac{\Delta\rho}{2}}{\rho_i}\right)\right). \tag{39}$$

The matrix is, again, upper right triangular because via (23)

$$\Delta CDF(\rho_j, \rho_i) = 0 \qquad \text{if } \rho_i > \rho_j. \tag{40}$$

Therefore, the numerical solution of the set of linear equations can be computed as before. The noise analysis can also be performed; however, we have not derived an analytic expression, as above, as the inverse matrix has a much more complicated structure. We defer these considerations to a future publication.

Experimental results.

The spherically symmetric case. As an example of the use of the method for the analysis of collected data we present results that utilize data generated from the FishTV sonar system (Jaffe et al., 1995). Briefly, the FishTV system is a multibeam sonar system which operates at a frequency of 445 kHz with bandwidth of 25 kHz and at frame rates of up to 4 Hz. The system provides acoustic backscatter from a set of 64 beams whose beam widths are 2 degrees by 2 degrees. The system resolves 512 range bins at a range increment of 0.75 cm, which yields a three-dimensional data set of dimensions 16 degrees by 16 degrees by 3.8 meters. The system has been used to measure the sonar reflectivity (Jaffe, Ohman, and De Robertis, 1998) and behavior (Jaffe, De Robertis, and Ohman, 1999) of small animals in the water column (zooplankton).

Given that the azimuthal resolution of the system is 2 degrees by 2 degrees, at a range of 3 m the cross-track resolution of the system is 10.5 cm by 10.5 cm. In contrast, the range resolution of the system, dictated by the bandwidth of the signal, in excellent signal-to-noise can be as small as 1 cm. Clearly, an algorithm which takes advantage of the superior range resolution in order to estimate the $pdf(\vec{v})$, the probability density function for velocity, is desired. This was especially appealing since the animal motions were suspected to be small, as they are quiescent during some of our observation times (daylight). In addition, during this time, there is good reason to believe that this probability density function is spherically symmetric. One advantage of the algorithm proposed here is that it permits use of this much better range resolution in order to compute the three-dimensional probability density function from the one-dimensional observations.

Data shown here were collected in a fjord in British Columbia over a period of approximately 10 minutes. The sonar was suspended below an anchored ship and oriented into the current via the use of a current vane which was placed on the data collection package. The sonar was operated at a rate of 2 Hz and was aimed slightly downward into the layer of animals that collects at depth during the daytime. Suspended at 85 meters, the system recorded the time varying reflections from targets that were between 7 and 10 meters in range. The recording process yielded a set of time varying reflections from each of the 64 sonar beams. Additional information describing the system can be found in Jaffe et al. (1995).

Extensive software development in conjunction with validation of the system's performance has been performed over a period of several years (De Robertis, 2001). The first step is to identify the presence of individual targets in each of the 64 sonar beams. This is accomplished via the use of a correlation receiver and also by inspecting the neighboring beams for side lobes as our targets are azimuthally spread. The output of this first step is an estimate for the three-dimensional voxel (horizontal and vertical bearing angle and range) in which the target is located, and its acoustic cross section (reflectivity). This set of three-dimensional locations, which are obtained over

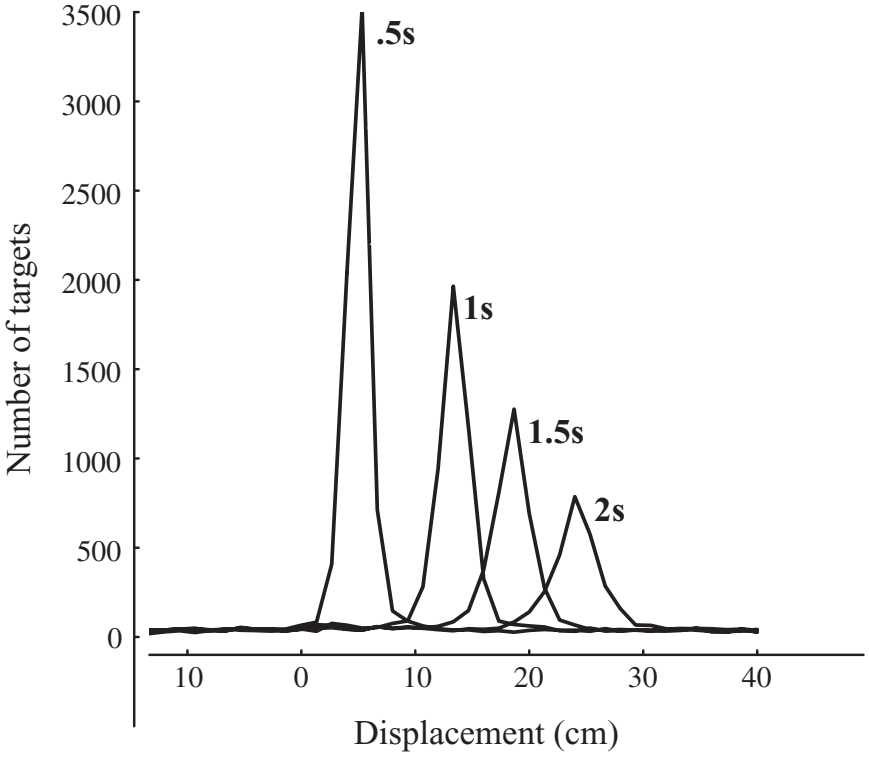


FIG. 4. *A set of displacement distribution functions: the number of targets that underwent translation ΔX in time interval ΔT as a function of time delay.*

a relatively short period of time for one complete system transmission-and-receive cycle, is referred to as a frame.

The next step in our treatment of the data, for the purposes of this article, is to consider the range displacements only. The histogram of these values provides the data from which the probability density function can be estimated. The histogram formed solely from the range estimates forms an estimate for $pdf_{1d}(\rho)$. In order to compute this function from the target lists as a function of frame number, various approaches can be used. Probably the most sophisticated approach would be to track the targets temporally and to estimate their range displacements accordingly. A simpler approach was taken here which should yield approximately the same results. That is, given a time interval, the range displacement vectors were computed for all targets in successive frames, which were separated by the given time interval from the two lists (with reflectivities above a certain value). This is computationally very simple, and under the assumption that the individual target locations are uncorrelated (among themselves), the extra displacements from targets that were not the same should produce a "dc value" or plateau upon which the true target displacements are observable. The displacements were binned into bins of width 1.5 cm. Although the data recording interval results in a range increment of .75 cm, this value was used, as it was closer to the "true" resolution of the system as measured by the width of the autocorrelated pulse.

Figure 4 contains a set of four histograms for these estimated range displacements from this data set. These displacement distributions can easily be converted into an

estimate for $pdf_{1d}(\rho)$ by simply dividing through by the total number of observations. The data indicate that there is a mean translation of the animals' positions (due to the current) which is superimposed upon a broadening of the distribution, which we have assumed is due to animal motion. It is evident that the majority of the animals are hardly moving at all, that is to say, the majority of their displacement is due to the current. This further motivated our interest in using this inversion algorithm to obtain a higher resolution estimate than could be obtained from just the three-dimensional positions, which suffer from poor azimuth resolution (relative to animal displacement).

The data in Figure 4 were obtained by forming the histograms for each of the individual beams for each of the time delays and then shifting them by small amounts in order to line them up. The reason that they had different modes for their displacements was due to the small differences in pointing angles for each of the 64 beams. This resulted in a different value for the projection of the current vector onto each of the beam-pointing angles. The correction assumes that the width of the peaks will change little as a function of this systematic correction.

Since these data were obtained by regarding only changes in animal position in range, the data reflects an estimate for $pdf_{1d}(\Delta\rho|\Delta T)$. Moreover, since it is likely that the "true" animal displacement probability density function is isotropic, the data were treated using the above analysis. Note that this probability density function might be due to both turbulence as well as animal movement. In the case that both of those are isotropic, the analysis considered here can be rightfully applied. In addition, since the current regime where the study was done was extremely laminar, the broadening of the probability density function was interpreted as due to animal movement. The inverse was computed by multiplying the data by the matrix A^{-1} to obtain an estimate for the radial slice through the three-dimensional function (here kept in the form of a displacement histogram). Figure 5 shows the result. An interesting feature of the inverse is that there is a significant decrease in the estimated number of targets that are simply drifting with the current, as contrasted with the estimate that would be obtained by simply taking the one-dimensional displacements histograms.

In order to test whether this inversion differs significantly from the measured data and to explore its stability, a simulation was done. Under the assumption that (1) the measured data was not subject to any systematic error and (2) the underlying probability density function for the measured data is Poisson distributed with expectation and standard deviation equal to the measured value, an ensemble of potential measured waveforms was simulated. Each member of the ensemble was then multiplied by the inverse matrix (as above) to compute a new inverse. This yielded an ensemble of inverses whose standard deviations are shown as dashed lines in Figure 5. The error curves indicate that, at the origin, the computed inverse is significantly different from the measured data; however, at larger displacements the inversion is not. This is an interesting and significant observation. Perhaps some degree of active swimming is needed for survival.

The cylindrically symmetric case. In this section we consider the utilization of the methodology which was developed for the cylindrically symmetric case. In the above spherically symmetric case we described the application of the methodology to data that were measured using a specially designed sonar system. Here, standard camera technology has been used to view the behavior of animals from a single direction. Under the assumption that the underlying probability density function for velocity has cylindrical symmetry, a temporally changing set of data from this sin-

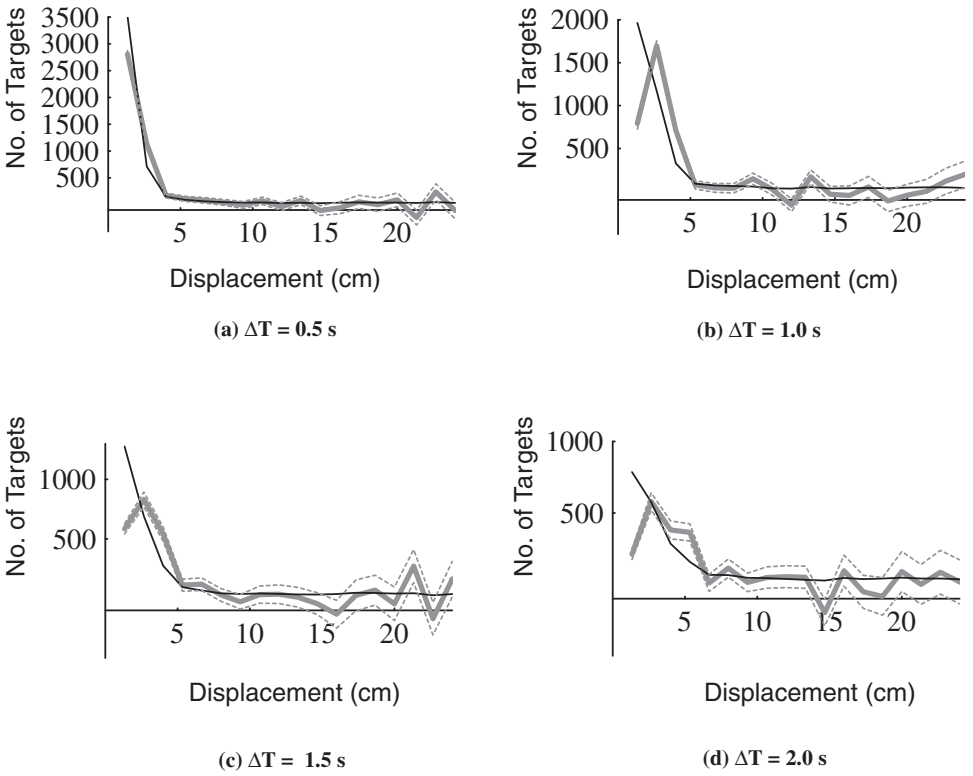


FIG. 5. *The result of performing the matrix inversion A^{-1} on the one-dimensional sonar data for different time delays. The solid black line is the observed histogram of the range displacement data. The gray line is the result obtained after multiplying by the inverse matrix. The dashed gray lines indicate plus or minus one standard deviation. The output of this process is an estimate for a radial slice through the spherically symmetric, three-dimensional histogram.*

gle camera can be used to infer a radial line through the cylindrically symmetric distribution.

Here we consider experiments that were performed to observe the behavior of small underwater animals (zooplankton of size 1 mm) in order to study the differences in their behavior in the presence and absence of food. The animals were placed in an aquarium and photographed with a video camera under dim red light. Postprocessing of the data yielded a set of displacement vectors for the animals $\{\Delta x_i, \Delta z_i; t, i = 1, N\}$ as a function of time.

This section contains the results of processing the set of five animal trajectories which consisted of approximately 100 positions each, yielding about 500 animal displacements (Leising and Franks, 2002). These vectors were then assembled into two histograms of displacements, one for Δx and the other for Δz, under the assumption that the two probability density functions are independent, as described above. The displacements for both positive and negative values were combined under the assumption that the distribution is symmetric.

As described above, the forward problem was modeled using a finite number of bins, and the algorithm was tested on several simulated distributions. Here, the experimental data of the recorded animal positions were used to obtain a set of animal displacements which were binned into 11 size classes. In order to compute the inver-

sion, an 11 by 11 matrix was created for the forward problem. The forward matrix was computed as

1.	.356	.227	.167	.132	.110	.093	.081	.072	.065	.059
0.	.644	.249	.174	.136	.111	.095	.082	.073	.066	.059
0.	0.	.523	.206	.148	.118	.095	.082	.073	.065	.060
0.	0.	0.	.452	.180	.131	.105	.089	.077	.068	.061
0.	0.	0.	0.	.404	.162	.119	.096	.820	.072	.064
0.	0.	0.	0.	0.	.368	.148	.110	.089	.076	.067
0.	0.	0.	0.	0.	0.	.341	.138	.102	.084	.072
0.	0.	0.	0.	0.	0.	0.	.319	.129	.096	.079
0.	0.	0.	0.	0.	0.	0.	0.	.301	.122	.091
0.	0.	0.	0.	0.	0.	0.	0.	0.	.285	.116
0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	.272

The inverse matrix will not be shown here; however, the inverted data set is shown in Figure 6, along with the original data histogram of the displacements. As the figure indicates, the estimate for the *true* data set is somewhat different than the observed data. Curves of plus or minus one standard deviation, computed in an identical way to the spherically symmetric case, are also shown. Interestingly, in the same way as in the spherically symmetric case, a decrease in targets that were practically still is evident. This inversion then also suggests that there are fewer targets moving with very small velocities than would be measured from just computing the probability density function from the projected data. In this case, it is widely known that the animals execute a "hop and sink" strategy for foraging, indicating that they indeed spend little of their time not moving at all. Since the animals are somewhat negatively buoyant, they cannot simply suspend movement and maintain their depth.

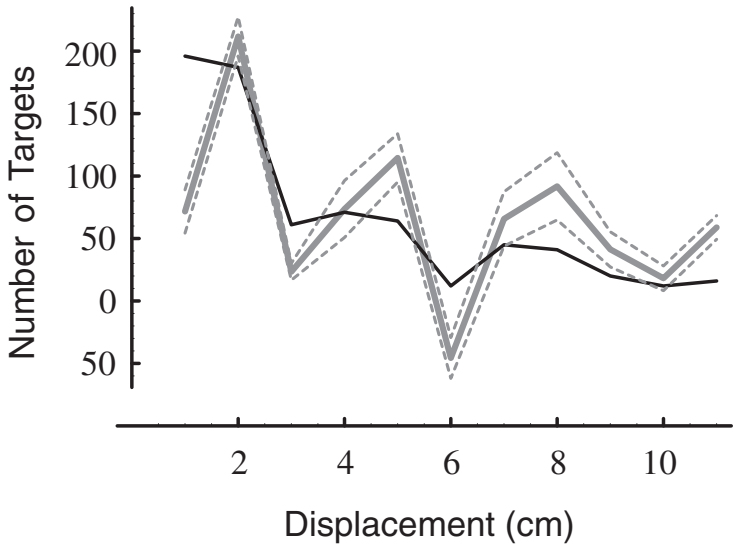


FIG. 6. *A histogram of observed animal displacements, proportional to* $pdf_{1d}(\Delta x)$ *(solid black), and the inversion for the "true" three-dimensional histogram, proportional to* $pdf_{3d}(\Delta r)$ *(gray). The dashed gray lines indicate plus or minus one standard deviation. The data was collected with a camera system which viewed animal displacements in dim red light.*

Discussion and conclusions. In this article a theory of tomographic inversions for probability density functions from observed data has been proposed. The theory is motivated by the many data collection systems that observe a lower-dimensional projection of the data where both one-dimensional and two-dimensional projections of three-dimensional vectors are observed. In the two cases considered here, a three-dimensional sonar system and an optical camera, animal displacement distributions were processed to obtain an estimate for radial slices through the two- and three-dimensional displacement probability density functions. The identical nature of this problem to the standard theories relating to tomographic inversion is emphasized. Several theorems are proved which demonstrate the utility of inverting symmetric marginal probability density functions using tomographic theorems. The utility of these ideas is demonstrated via the application of the numerical implications to experimentally collected data.

From the scientific point of view there are a host of interesting problems that could be explored using the methodology described here. In this regard, this author's primary interest has been in tracking individual animals in various aquatic environments. Other applications of the methodology may relate to tracking animals both on land and in the air. As demonstrated, even one-dimensional range information from a simple device can be transformed into a radial slice of a multidimensional probability density function using the methodology presented here. In the case of animals that fly, the similarities are clear between observing birds or insects and the aquatic examples in the applications given here. Other interesting probability moments exist as well. So, for example, one might be interested in the joint distribution of some predator and prey via their relative distances in some reduced-dimensional coordinate system. Inversion for the true joint probability density function from this joint distribution, under assumptions here of various symmetries, therefore seems like an interesting future application of the technique.

In a similar context, we remark that other work by this author involves the projection of fluorescent organisms onto a plane (Jaffe, Franks, and Leising, 1998; Franks and Jaffe, 2001). Under various assumptions about symmetry, the correlation distances can be computed.

From the point of view of the reconstruction of multidimensional functions from projections, there are a number of theoretical questions which arise as a result of this work. A significant issue concerns the underlying assumptions of isotropy, either in three dimensions or in two, and how strictly isotropic the distributions need to be in order for the results to apply. So, for example, in the analysis of the sonar system's measurements of one-dimensional displacements, it was assumed that the animals' movements could be described by an isotropic probability density function. However, since the animals remain at a depth strata during the day, it cannot be strictly true that their distribution is isotropic, as over time they would diffuse into the entire volume. If the underlying three- or two-dimensional probability density function is not strictly isotropic, the above theorems do not apply; however, if the distribution is "almost" isotropic, then one might hope that some variation of the theorems can be used. Moreover, when the system is not truly isotropic, perhaps one can use some fewer number of projections than would be required for a true multidimensional inversion without arbitrary assumptions about the underlying structure of the desired object. The application of the theory and methodology described here to situations of pragmatic interest is an interesting area for future research.

Additional issues which relate to the numerical solution of the system of equations remain. These issues are generic to the theory of the reconstruction of functions from

projections and thus relate to the field in general. Under the special symmetries considered here, the inversions would benefit from additional work. So, for example, only a straightforward multiplication by the matrix inverse was performed here. However, in the presence of noise, this inversion is probably neither a maximum likelihood nor maximum a priori estimate for the solution. So, for example, in both cases considered here, it is likely that there is some noise amplification as a result of the simplicity of the inversions.

Another issue relates to the numerical implementation of the inverse. Tomographic reconstructions are typically performed through the use of filtered backprojection. It would be interesting to characterize the performance of that algorithm in comparison to the more analytically based algorithms that capitalize on the symmetry which have been proposed here.

Finally, perhaps the most surprising and significant conclusion is that it is productive to perform tomographic inversions on marginal probability density functions. As shown, these inversions can provide an estimate for a true underlying two- or three-dimensional probability density function when performed on one- and two-dimensional projected data. Collection of these lower-dimensional data sets can vastly simplify an experimental setup, as the collection of true three-dimensional data requires both careful calibration and additional hardware when compared with the lower-dimensional case.

Acknowledgments. The author would like to thank both the National Science Foundation and the Office of Naval Research for supporting his work over the years. Thanks are also extended to A. Leising (National Marine Fisheries Service, Pacific Grove, CA) for providing the displacement data for cylindrically symmetric analysis and to Simon Levin (Princeton University) for encouraging remarks after reading an earlier version of this manuscript.

REFERENCES

A. De Robertis, *Validation of acoustic echo counting for studies of zooplankton behavior*, ICES J. Marine Sci., 58 (2001), pp. 543–561.

P. J. S. Franks and J. S. Jaffe, *Microscale distributions of phytoplankton: Initial results from a two-dimensional imaging flourometer, OSST*, Marine Ecology Progress Ser., 220 (2001), pp. 59–72.

G. T. Herman, *Image Reconstruction from Projections: The Fundamentals of Computerized Tomography*, Academic Press, San Francisco, 1980.

D. W. L. Hukins, *X-Ray Diffraction by Disordered and Ordered Systems: Covering X-Ray Diffraction by Gases, Liquids, and Solids and Indicating How the Theory of Diffraction by These Different States of Matter Is Related and How It Can Be Used to Solve Structural Problems*, Pergamon Press, Oxford, New York, 1981.

J. S. Jaffe, P. J. S. Franks, and A. W. Leising, *Simultaneous imaging of phytoplankton and zooplankton distributions*, Oceanography, 11 (1998), pp. 24–29.

J. S. Jaffe, E. Reuss, D. McGehee, and G. Chandran, *FTV, a sonar for tracking macrozooplankton in three-dimensions*, Deep Sea Res., 42 (1995), pp. 1495–1512.

J. S. Jaffe, M. D. Ohman, and A. De Robertis, *OASIS in the sea: Measurement of the acoustic reflectivity of zooplankton with concurrent optical imaging*, Deep Sea Res., 45 (1998), pp. 1239–1253.

J. S. Jaffe, A. De Robertis, and M. D. Ohman, *Sonar estimates of daytime activity levels of Euphausia pacifica in Saanich Inlet*, Can. J. Fisheries & Aquatic Sci., 56 (1999), pp. 2000–2010.

A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*, IEEE Press, New York, 1987.

A. W. Leising and P. J. S. Franks, *Does Acartia clausi (Copepoda: Calanoida) use an area-restricted search foraging strategy to find food?*, Hydrobiologia, 480 (2002), pp. 193–207.

D. McGehee and J. S. Jaffe, *Three-dimensional swimming behavior of individual zooplankters: Observations using the acoustical imaging system FishTV*, ICES J. Marine Sci., 53 (1996), pp. 363–369.

A. Okubo, *Diffusion and Ecological Problems: Mathematical Models*, Springer-Verlag, Berlin, 1980.

J. K. Parrish and L. Edelstein-Keshet, *Complexity, pattern, and evolutionary trade-offs in animal aggregation*, Science, 284 (1999), pp. 99–101.

J. K. Parrish and W. M. Hamner, *Animal Groups in Three Dimensions*, Cambridge University Press, Cambridge, UK, 1997.

J. J. Torres and J. J. Childress, *Relationship of oxygen consumption to swimming speed in Euphausia pacifica* 1. *Effects of temperature and pressure*, Marine Biol., 74 (1983), pp. 79–86.

SIAM J. APPL. MATH.
Vol. 65, No. 5, pp. 1526–1546

MULTIPLE EQUILIBRIA IN COMPLEX CHEMICAL REACTION NETWORKS: I. THE INJECTIVITY PROPERTY*

GHEORGHE CRACIUN† AND MARTIN FEINBERG‡

Abstract. The capacity for multiple equilibria in an isothermal homogeneous continuous flow stirred tank reactor is determined by the reaction network. Examples show that there is a very delicate relationship between reaction network structure and the possibility of multiple equilibria. We suggest a new method for discriminating between networks that have the capacity for multiple equilibria and those that do not. Our method can be implemented using standard computer algebra software and gives answers for many reaction networks for which previous methods give no information.

Key words. equilibrium points, chemical reaction networks, chemical reactors, mass-action kinetics

AMS subject classifications. 80A30, 37C25, 65H10

DOI. 10.1137/S0036139904440278

1. Introduction.

We are interested in studying the uniqueness of positive equilibrium points of a special but large class of systems of nonlinear ordinary differential equations (ODEs): those that derive from chemical reaction networks. In order to understand how these equations arise, we will first look informally at an example of a reaction network and see how it induces a system of ODEs.

Consider some chemical species A, B, C, D, W, X, Y, and Z, and suppose that the chemical reactions occurring among these species are

$$A + B \rightleftharpoons C, \quad X \rightleftharpoons 2A + D \rightleftharpoons Y, \quad D \rightleftharpoons C + W, \quad B + D \rightleftharpoons Z. \tag{1.1}$$

We will study a particular kind of reactor, called a continuous flow stirred tank reactor (CFSTR; see [3]) by chemical engineers. Think of a CFSTR as just some enclosed volume endowed with a feed stream and an outflow stream. Suppose that its contents are kept at constant temperature and are spatially uniform. Now imagine that a liquid mixture of species A, B, C, D, W, X, Y, and Z is continuously supplied to some CFSTR at a constant volumetric flow rate g (volume/time). Also, the contents of the CFSTR are continuously removed at the same volumetric flow rate g. Chemical reactions occur in the CFSTR, according to (1.1). We would like to investigate the temporal evolution of the composition of the mixture within the CFSTR. Let us denote by $c_A^f, c_B^f, \ldots, c_Z^f$ the molar concentrations (moles/volume) in the feed stream and by $c_A(t), c_B(t), \ldots, c_Z(t)$ the molar concentrations within the CFSTR (and effluent stream) at time t. We will denote the vector of all molar concentrations within the CFSTR by $c(t)$. We get the picture shown in Figure 1.1.

One source of change in composition is the occurrence of chemical reactions. It is generally assumed that the occurrence rate of each reaction at time t depends just

*Received by the editors January 27, 2004; accepted for publication June 26, 2004; published electronically May 12, 2005.
http://www.siam.org/journals/siap/65-5/44027.html
†Mathematical Biosciences Institute, Ohio State University, 125 W. 18th Avenue, Columbus, OH 43210 (gcraciun@mbi.ohio-state.edu). This author was supported by the National Science Foundation under agreement 0112050.
‡Department of Chemical Engineering and Department of Mathematics, Ohio State University, 125 Koffolt Laboratories, 140 W. 19th Avenue, Columbus, OH 43210 (feinberg.14@osu.edu).

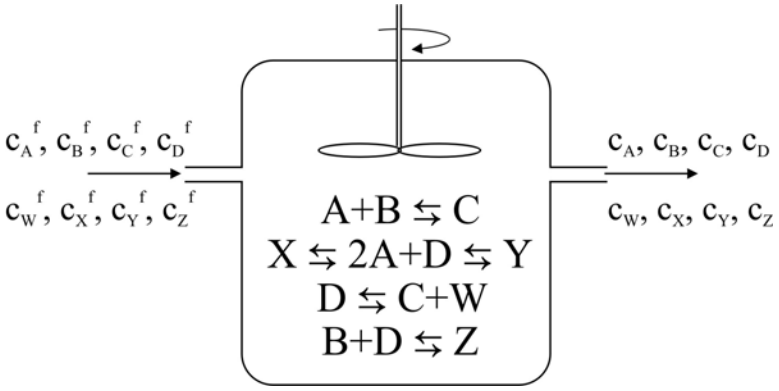


FIG. 1.1. *The CFSTR of the reaction network* (1.1).

on the mixture composition $c(t)$. For example, for the reaction $A + B \to C$ there exists a nonnegative real-valued *rate function* $K_{A+B\to C}$ such that $K_{A+B\to C}(c)$ is the occurrence rate of reaction $A + B \to C$ per unit volume of mixture when the mixture composition is given by the vector c. Let us now think about the instantaneous rate of change of c_A. Whenever the reaction $A + B \to C$ occurs we lose one molecule of A. Also, whenever the reaction $C \to A + B$ occurs we gain one molecule of A. Similarly, whenever the reaction $X \to 2A + C$ occurs we gain two molecules of A, and so on.

The other source of changes in composition is the difference between the composition c^f in the feed stream and the composition c in the effluent stream. (Note that the composition of the effluent stream is presumed to be identical to that of the homogeneous mixture within the vessel.) If V is the total volume of the mixture within the CFSTR,[1] we get

$$\begin{aligned} V\dot{c}_A = g(c_A^f - c_A) &- VK_{A+B\to C}(c) + VK_{C\to A+B}(c) \\ &- 2VK_{2A+D\to X}(c) + 2VK_{X\to 2A+D}(c) \\ &- 2VK_{2A+D\to Y}(c) + 2VK_{Y\to 2A+D}(c). \end{aligned} \tag{1.2}$$

We will now look more closely at the structure of the rate functions. In most cases chemists suppose the rate functions to be of *mass-action* type (see [26]). This means that, for example, for the reaction $A + B \to C$, the more A there is in the CFSTR, the more occurrences of the reaction there will be, and similarly for B. More precisely, we presume that the occurrence rate of the reaction $A + B \to C$ is proportional to the probability of A and B meeting in the CFSTR, which, in turn, is proportional to the value of $c_A c_B$. Thus, we write

$$K_{A+B\to C}(c) = k_{A+B\to C} c_A c_B,$$

where $k_{A+B\to C}$ is a positive *rate constant* for the reaction $A + B \to C$. For the reaction $2A + D \to X$ an occurrence requires two molecules of A and one molecule of D to meet in the CFSTR, and we consider the probability of this encounter to be proportional to $c_A^2 c_D$. Therefore we get

$$K_{2A+D\to X}(c) = k_{2A+D\to X} c_A^2 c_D,$$

[1] We assume hereafter that the densities of the feed and the effluent streams are identical and time-invariant. This implies that V is constant in time. We also assume throughout that the temperature of the reacting mixture is held constant.

where $k_{2A+D\to X}$ is the rate constant for the reaction $2A + D \to X$. In the case of a reaction such as $D \to C + W$ it is presumed that the occurrence rate is simply proportional to the molar concentration of D, i.e.,

$$K_{D\to C+W}(c) = k_{D\to C+W}c_D.$$

The rate constants are usually either approximated on the basis of chemical principles or are deduced from experiments. If we assume mass-action kinetics for the network (1.1), then we get the following *associated system of differential equations*:

$$\begin{aligned}
(1.3)\quad \dot{c}_A &= (g/V)(c_A^f - c_A) - k_{A+B\to C}c_Ac_B + k_{C\to A+B}c_C - 2k_{2A+D\to X}c_A^2c_D \\
&\quad + 2k_{X\to 2A+D}c_X - 2k_{2A+D\to Y}c_A^2c_D + 2k_{Y\to 2A+D}c_Y, \\
\dot{c}_B &= (g/V)(c_B^f - c_B) - k_{A+B\to C}c_Ac_B + k_{C\to A+B}c_C \\
&\quad + k_{Z\to B+D}c_Z - k_{B+D\to Z}c_Bc_D, \\
\dot{c}_C &= (g/V)(c_C^f - c_C) + k_{A+B\to C}c_Ac_B - k_{C\to A+B}c_C \\
&\quad + k_{D\to C+W}c_D - k_{C+W\to D}c_Cc_W, \\
\dot{c}_D &= (g/V)(c_D^f - c_D) + k_{X\to 2A+D}c_X - k_{2A+D\to X}c_A^2c_D \\
&\quad + k_{Y\to 2A+D}c_Y - k_{2A+D\to Y}c_A^2c_D - k_{D\to C+W}c_D + k_{C+W\to D}c_Cc_W \\
&\quad - k_{B+D\to Z}c_Bc_D + k_{Z\to B+D}c_Z, \\
\dot{c}_W &= (g/V)(c_W^f - c_W) + k_{D\to C+W}c_D - k_{C+W\to D}c_Cc_W, \\
\dot{c}_X &= (g/V)(c_X^f - c_X) - k_{X\to 2A+D}c_X + k_{2A+D\to X}c_A^2c_D, \\
\dot{c}_Y &= (g/V)(c_Y^f - c_Y) + k_{2A+D\to Y}c_A^2c_D - k_{Y\to 2A+D}c_Y, \\
\dot{c}_Z &= (g/V)(c_Z^f - c_Z) + k_{B+D\to Z}c_Bc_D - k_{Z\to B+D}c_Z.
\end{aligned}$$

Therefore we obtain a system of ODEs where all equations are determined by the reaction network up to some constants: c_A^f, $c_B^f, \ldots, c_Z^f$, g/V, and $k_{A+B\to C}$, $k_{C\to A+B}, \ldots, k_{Z\to B+D}$. We are now going to ask the question: does this system of ODEs have no more than one positive equilibrium for all positive values of g/V, all positive values of the rate constants, and all nonnegative values of the feed concentrations c_A^f, $c_B^f, \ldots, c_Z^f$?

This question is motivated by experiments. For homogeneous liquid phase CFSTRs, there are very few reports of reaction networks with more than one positive equilibrium, despite hundreds of reaction networks being studied (see [9] for one such report). We are asking this question for *all* positive rate constants since in practice there is poor knowledge of the rate constants of reactions.

This question is not easy to answer, in general. Even if, for the simple example above, we could decide one way or the other by some ad-hoc method, there will be thousands of other reaction networks for which we will still not know the answer. There are important reaction networks with hundreds of reactions. Ideally, there will be a simple way to decide on the uniqueness of equilibria.

We say that a mass-action network *has the capacity for multiple positive equilibria* (in an isothermal homogeneous CFSTR context) if there are positive values of the flow rate, the volume, the rate constants, and nonnegative values of the feed concentrations such that the resulting differential equations admit two or more distinct positive equilibria.

According to [30], there are examples of very similar reaction networks with very different capacities for multiple positive equilibria (see Table 1.1). Networks (i) and

TABLE 1.1
Some examples of reaction networks and their capacity for multiple positive equilibria [30].

Reaction network		Has the capacity for multiple equilibria?
(i)	$A+B \rightleftharpoons P$ $B+C \rightleftharpoons Q$ $C \rightleftharpoons 2A$	Yes
(ii)	$A+B \rightleftharpoons P$ $B+C \rightleftharpoons Q$ $C+D \rightleftharpoons R$ $D \rightleftharpoons 2A$	No
(iii)	$A+B \rightleftharpoons P$ $B+C \rightleftharpoons Q$ $C+D \rightleftharpoons R$ $D+E \rightleftharpoons S$ $E \rightleftharpoons 2A$	Yes
(iv)	$A+B \rightleftharpoons P$ $B+C \rightleftharpoons Q$ $C \rightleftharpoons A$	No
(v)	$A+B \rightleftharpoons F$ $A+C \rightleftharpoons G$ $C+D \rightleftharpoons B$ $C+E \rightleftharpoons D$	Yes
(vi)	$A+B \rightleftharpoons 2A$	No
(vii)	$2A+B \rightleftharpoons 3A$	Yes
(viii)	$A+2B \rightleftharpoons 3A$	No

(iii) in Table 1.1 have the capacity for multiple positive equilibria, but the "middle case" network (ii) does not. Similarly, network (iv) is almost identical to (i), but does not have the capacity for multiple positive equilibria. Moreover, network (v) is an example that shows that we don't need two or more copies of the same species to appear in the same reaction for the network to admit multiple positive equilibria. Also, changing (vi) to (vii) does bring in multiple positive equilibria, but changing (vi) to (viii) does not. Therefore, a good theory of multiple positive equilibria in CFSTRs should be able to differentiate between these subtle differences.

Let us look again at the system of ODEs in (1.3). If we are just interested in equilibria, we set all the left-hand side terms equal to zero, and we get a system of polynomial (algebraic) equations. Let us also move the feed terms $c_A^f, \dots, c_Z^f$ to the other side of the equations. We choose units such that $g/V = 1$. If we now change signs in both sides and rearrange terms, then we get the following system of eight polynomial equations:

$$
\begin{aligned}
c_A^f &= c_A + k_{A+B\to C}c_Ac_B - k_{C\to A+B}c_C + 2k_{2A+D\to X}c_A^2c_D \\
&\quad - 2k_{X\to 2A+D}c_X + 2k_{2A+D\to Y}c_A^2c_D - 2k_{Y\to 2A+D}c_Y, \\
c_B^f &= c_B + k_{A+B\to C}c_Ac_B - k_{C\to A+B}c_C - k_{Z\to B+D}c_Z \\
&\quad + k_{B+D\to Z}c_Bc_D, \\
c_C^f &= c_C - k_{A+B\to C}c_Ac_B + k_{C\to A+B}c_C - k_{D\to C+W}c_D \\
&\quad + k_{C+W\to D}c_Cc_W,
\end{aligned}
\tag{1.4}
$$

$$
\begin{aligned}
c_D^f &= c_D - k_{X\to 2A+D}c_X + k_{2A+D\to X}c_A^2c_D - k_{Y\to 2A+D}c_Y \\
&\quad + k_{2A+D\to Y}c_A^2c_D + k_{D\to C+W}c_D - k_{C+W\to D}c_Cc_W \\
&\quad + k_{B+D\to Z}c_Bc_D - k_{Z\to B+D}c_Z, \\
c_W^f &= c_W - k_{D\to C+W}c_D + k_{C+W\to D}c_Cc_W, \\
c_X^f &= c_X + k_{X\to 2A+D}c_X - k_{2A+D\to X}c_A^2c_D, \\
c_Y^f &= c_Y - k_{2A+D\to Y}c_A^2c_D + k_{Y\to 2A+D}c_Y, \\
c_Z^f &= c_Z - k_{B+D\to Z}c_Bc_D + k_{Z\to B+D}c_Z.
\end{aligned}
$$

Let us denote by k the vector formed by the parameters $k_{A+B\to C}$, $k_{C\to A+B}, \ldots,$ $k_{Z\to B+D}$. Now we denote by $p(c,k)$ the vector of right-hand sides of the system of polynomial equations (1.4), and we call it *the polynomial function associated to the reaction network* (1.1). We regard $p(c,k)$ as a vector-valued function of a (positive) composition vector c and depending on a (positive) vector of rate constants k.

We say that the reaction network (1.1) is an *injective reaction network* if the function $c \to p(c,k)$ is injective for all positive k.

The following simple fact is a key observation: *If a reaction network has the capacity for multiple positive equilibria, then there exists some choice of positive vector k_0 such that the function $c \to p(c,k_0)$ is not injective.* In particular, $p(c^*,k_0) = p(c^\#,k_0) = c^f$ for some feed composition c^f and some distinct compositions c^*, $c^\#$. In other words, an injective reaction network does not have the capacity for multiple positive equilibria; i.e., injectivity is a sufficient condition for the absence of multiple positive equilibria.

Remark 1.1. Injectivity is *not* a necessary condition for the absence of multiple positive equilibria. The reason is that, for a network to admit multiple positive equilibria, there must be a k_0 such that $p(\cdot,k_0)$ maps two distinct compositions not only into the same vector, but, in fact, also into a *nonnegative* feed composition c^f (see (1.4), (3.10)). Were it not for this nonnegativity condition, injectivity would be equivalent to uniqueness of equilibria.

Nevertheless, the class of injective reaction networks subsumes the largest class of reaction networks for which the answer was previously found in [20, 30, 31]. *The main purpose of this paper is to describe a method that allows us to decide whether a given reaction network is injective or not.*

Remark 1.2. In general, it is of course very difficult to check whether a given multidimensional polynomial function is injective or not. Moreover, the function $c \to p(c,k)$ involves several unknown parameters. Our method derives, first, from a theoretical observation about the function $p(\cdot,\cdot)$ and, second, from a rather remarkable empirical observation.

The theoretical observation, discussed in section 3, is that a reaction network is injective whenever its associated polynomial function has the property that $\frac{\partial p(c,k)}{\partial c}$ is nonsingular for all positive c and all positive k. (There is no claim here that any such assertion is true for polynomial functions in general; rather, the assertion is made specifically for polynomial functions that derive, in the manner indicated, from chemical reaction networks.)

To describe the empirical observation, we first note that the nonsingularity property is, of course, equivalent to the requirement that $\det(\frac{\partial p(c,k)}{\partial c})$ be nonzero for all positive c and all positive k. For moderately large networks, the calculation of $\det(\frac{\partial p(c,k)}{\partial c})$ will result in hundreds or thousands of terms, even after combining all similar monomials. Each resulting nonzero term will be a monomial in the (positive)

species concentrations and the (positive) rate constants, with each term containing an integer coefficient. Thus the sign of each term is carried by the sign of its integer coefficient. The empirical observation is this: *For very large and robust classes of networks it is the case that, despite the huge number of terms, the integer coefficient in every term is positive!* In this case, $\det(\frac{\partial p(c,k)}{\partial c})$ cannot vanish, and injectivity of the network is ensured (as is the impossibility of multiple positive equilibria). In fact, we will show that positivity of all nonzero coefficients is *both necessary and sufficient* for injectivity of the network.

In a subsequent article we intend to characterize, in graph-theoretical terms, large classes of networks for which all coefficients are positive. In the meantime, we observe that, for a given network of interest, checking for positivity of the coefficients is a matter that can be resolved by presently available computer algebra systems.

By way of example, we show in (1.5) the first few terms of the expansion of $\det(\frac{\partial p(c,k)}{\partial c})$ for network (1.1):

$$
\begin{aligned}
(1.5)\quad & \det\left(\frac{\partial p(c,k)}{\partial c}\right) \\
&= 10k_{C\to A+B}k_{D\to C+W}k_{2A+D\to X}c_Ac_D^2k_{W\to 0}k_{B+D\to Z}k_{X\to 0}k_{Y\to 0}k_{Z\to 0} \\
&+ 4k_{C\to A+B}k_{D\to 0}k_{2A+D\to X}c_Ac_D^2k_{W\to 0}k_{B+D\to Z}k_{X\to 0}k_{Y\to 2A+D}k_{Z\to 0} \\
&+ 4k_{C\to A+B}k_{D\to 0}k_{2A+D\to Y}c_Ac_Dk_{W\to 0}k_{B\to 0}k_{X\to 0}k_{Y\to 0}k_{Z\to 0} \\
&+ k_{C\to A+B}k_{2A+D\to X}c_A^2k_{A\to 0}k_{W\to 0}k_{B+D\to Z}c_Dk_{X\to 0}k_{Y\to 2A+D}k_{Z\to 0} \\
&+ 4k_{C\to A+B}k_{D\to 0}k_{2A+D\to Y}c_Ac_Dk_{W\to 0}k_{B\to 0}k_{X\to 2A+D}k_{Y\to 0}k_{Z\to 0} \\
&+ 6k_{C\to A+B}k_{D\to C+W}k_{2A+D\to Y}c_Ac_Dk_{W\to 0}k_{B\to 0}k_{X\to 2A+D}k_{Y\to 0}k_{Z\to B+D} \\
&+ 9k_{C\to 0}k_{2A+D\to Y}c_A^2c_Dk_{A+B\to C}k_{C+W\to D}c_Ck_{B+D\to Z}c_Bk_{Z\to 0}k_{X\to 0}k_{Y\to 0} \\
&+ 9k_{C\to 0}k_{2A+D\to Y}c_A^2c_Dk_{A+B\to C}k_{C+W\to D}c_Ck_{B+D\to Z}c_Bk_{Z\to 0}k_{X\to 2A+D}k_{Y\to 0} \\
&+ \cdots.
\end{aligned}
$$

In Table 1.2 we exhibit the (computer-generated) set of all coefficients that would have resulted had the expansion been completed. Note that all the entries are positive. Thus, we conclude that network (1.1) does not have the capacity for multiple positive equilibria in an isothermal CFSTR context.

Our claim that, across wide varieties of reaction networks, it is common for all coefficients to be positive is consistent with the paucity of experimental observations of multiple equilibria in isothermal homogeneous CFSTRs.

In section 3 we provide elaboration on the remarks made here.

Before we describe our results we would like to specify their place in the general landscape of chemical reaction network theory.

Stability results are discussed in [6, 7, 8, 11, 12, 13, 14, 15, 16, 18, 20, 24].

In [12, 13, 14, 15, 16, 17, 18, 19, 20, 25] reaction networks are classified by means of a nonnegative integer index called the *deficiency*. It is then shown how, for reaction networks of *small* deficiency, one can decide whether they have the capacity for multiple positive equilibria (see also the software package [21]).

On the other hand, it is also shown (see [27]) that the deficiency-oriented theory is not likely to give information for a large class of isothermal homogeneous CFSTRs. Work that is *complementary* to the deficiency-oriented theory, and aimed specifically at CFSTRs, was originated in [29] and then substantially broadened in [30, 31].

In [30, 31] Schlosser and Feinberg associate to any reaction network a graph called the *Species-Complex-Linkage (SCL) graph* of the reaction network. Then they describe

TABLE 1.2

The list of all nonzero coefficients in the expansion of the determinant of the Jacobian of the function $c \to p(c,k)$ for the reaction network (1.1). *Note that they are all positive.*

10	4	4	1	4	6	9	9	4	4	4	1	4	1	4	4	4	4	9	4	4
1	4	4	1	1	4	4	1	1	1	4	4	4	4	4	4	6	4	4	4	4
1	1	4	1	1	4	1	1	1	4	4	1	1	15	4	1	4	4	4	1	1
9	1	4	9	4	4	4	1	1	4	15	4	1	9	1	1	1	1	1	1	1
3	3	3	4	1	4	4	4	1	1	4	4	9	1	1	4	4	4	4	15	1
1	4	4	1	1	4	1	6	4	4	4	4	1	1	4	4	4	4	10	1	4
4	4	4	4	6	1	1	4	4	4	6	4	2	1	2	1	1	1	4	10	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	3	1	3	1	3	1	1	4
4	4	1	1	1	1	1	4	4	1	1	6	4	4	1	4	1	1	9	1	1
1	4	1	1	1	1	4	1	4	4	4	2	1	10	4	4	4	4	1	4	1
1	1	4	1	1	1	1	4	1	4	2	1	1	6	4	4	4	15	1	1	6
2	4	1	1	4	4	1	4	1	4	4	1	4	4	4	4	1	4	1	1	2
4	4	4	4	4	4	4	1	4	4	1	1	1	1	4	4	1	4	1	1	1
1	4	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	3	1	1	1	1	3	1	1	1	1	4	4	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	4	1	1	1	1	1	1	1	1	1	1	1	4	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1																			

a criterion in terms of the SCL graph that implies that the CFSTR associated to the reaction network does not have the capacity for multiple positive equilibria. This SCL graph criterion of [30, 31] describes large classes of reaction networks that do not have the capacity for multiple positive equilibria. On the other hand, it is not conclusive for some reaction networks (including (1.1)), and it is not easy to implement as a computer algorithm.

In Theorems 3.1–3.3 we describe equivalent formulations of the injectivity criterion that allow us to decide whether a given reaction network is injective or not using a simple computer algorithm (recall that an injective reaction network cannot have the capacity for multiple positive equilibria). Moreover, the injectivity criterion is less restrictive than the SCL graph criterion in [30, 31]: if the SCL graph criterion can be applied, then our criterion can be applied as well, but sometimes the SCL graph criterion is not conclusive, while our criterion is conclusive.

Applications of chemical reaction network theory are very diverse. There has been a recent surge of interest in applications of dynamics arising from complex reaction networks in biology. A very interesting discussion of biological applications appears in [4]. Also, recent articles address the role of reaction networks in cellular biochemistry [1, 2, 5, 10], in genetics [22, 23, 33], in bioengineering [32], and immunology [34].

In section 2 we give a precise definition of a reaction network, and we discuss some associated ideas. In section 3 we prove equivalent formulations of injectivity for reaction networks (recall that injectivity implies the absence of multiple positive equilibria). We will see that some of these equivalent formulations of injectivity allow us to test whether a given reaction network is injective or not, using a very simple algorithm. In section 4 we describe a condition which implies that a reaction network *does* have the capacity for multiple positive equilibria. Section 5 contains concluding remarks.

2. Definitions and notation. We denote by $\mathbb{R}_+$ the set of strictly positive real numbers, and by $\bar{\mathbb{R}}_+$ the set of nonnegative real numbers. For an arbitrary finite set I we denote by $\mathbb{R}^I$ the real vector space of all formal sums $\sum_{i\in I}\alpha_i i$ for all $\alpha_i \in \mathbb{R}$. Note that I becomes a basis of $\mathbb{R}^I$. By $\mathbb{R}^I_+$ we mean the set of sums $\sum_{i\in I}\alpha_i i$ in which all α_i are strictly positive. By $\bar{\mathbb{R}}^I_+$ we mean the set of sums $\sum_{i\in I}\alpha_i i$ in which all α_i are nonnegative.

In the following definition the *complexes* of a reaction network are to be understood as the objects (such as $A + B$) at the heads and tails of the reaction arrows.

DEFINITION 2.1 (see [14, 18]). *A* chemical reaction network *consists of three finite sets:*

(i) *a set* $\mathscr{S}$ *of* species *of the network;*
(ii) *a set* $\mathscr{C} \subset \bar{\mathbb{R}}^{\mathscr{S}}_+$ *of* complexes *of the network;*
(iii) *a set* $\mathscr{R} \subset \mathscr{C} \times \mathscr{C}$ *of* reactions*, with the following properties:*
  (a) $(y, y) \notin \mathscr{R}$ *for any* $y \in \mathscr{C}$*;*
  (b) *for each* $y \in \mathscr{C}$ *there exists* $y' \in \mathscr{C}$ *such that* $(y, y') \in \mathscr{R}$ *or such that* $(y', y) \in \mathscr{R}$.

When $(y, y') \in \mathscr{R}$ we say that the complex y *reacts to* complex y'. When this is the case we will write $y \to y'$, since it is the usual notation in chemistry.

If we look at the differential equations in (1.3), it is clear that, for CFSTRs in general, there will be not only terms that derive from the occurrence of chemical reactions but also linear terms (such as $-(g/V)c_A$) that derive from the presence of the outflow stream, and constant terms (such as $(g/V)c^f_A$) that derive from the presence of the feed stream. So that all such terms can be brought into a common reaction network theory framework, it will be useful to regard such "flow" terms as having derived from formal chemical "reactions" such as $A \to 0$ (corresponding to the outflow of A) and $0 \to A$ (corresponding to the feed of A); see [12, 25]. Here we view "0" as the zero vector of $\mathbb{R}^{\mathscr{S}}$. If we imagine $A \to 0$ to be governed by mass-action kinetics with rate constant $k_{A\to 0} = g/V$, then the contribution to $\dot{c}_A$ in (1.3) will be precisely $-(g/V)c_A$. We adopt the convention that the mass-action rate of a reaction of the form $0 \to A$ is constant (and equal to the associated rate constant $k_{0\to A}$). Thus, if we choose $k_{0\to A} = (g/V)c^f_A$, then the contribution of the reaction $0 \to A$ to $\dot{c}_A$ is just $(g/V)c^f_A$. In this way, the formal "flow reactions" $A \to 0$ and $0 \to A$ account for the flow terms that appear in the equation for $\dot{c}_A$. More generally, there are advantages to viewing CFSTR mass-action differential equations as having derived from a mass-action system in which the set of "true" reactions is augmented with the set of "flow reactions," with appropriately chosen rate constants. (Recall that we have chosen units such that $g/V = 1$ so that, for us, $k_{s\to 0} = 1$ for all $s \in \mathscr{S}$.)

Hereafter, we shall regard the operative reaction network under discussion to be the augmented one. If, for example, all species are present in the feed stream, then we augment the set of reactions in (1.1) by adding the following flow reactions:

$$\begin{aligned} &0 \to A,\ 0 \to B,\ 0 \to C,\ 0 \to D,\ 0 \to W,\ 0 \to X,\ 0 \to Y,\ 0 \to Z,\\ &A \to 0,\ B \to 0,\ C \to 0,\ D \to 0,\ W \to 0,\ X \to 0,\ Y \to 0,\ Z \to 0. \end{aligned} \tag{2.1}$$

If a certain species, say W, is deemed absent from the feed stream (i.e., if $c^f_W = 0$), then the reaction $0 \to W$ would be omitted. (With respect to injectivity considerations, the presence or absence of certain species in the feed is of no consequence.) Flow reactions of type $0 \to A$ are called *feed reactions*, and flow reactions of type $A \to 0$ are called *outflow reactions*. As Figure 1.1 indicates, all species are deemed present

in the effluent stream, so there is an outflow reaction for each species. (In a future paper we will discuss the implications of relaxing this assumption.)

So, the augmented network corresponding to the reaction network (1.1) has the set of species $\mathscr{S} = \{A, B, C, D, W, X, Y, Z\}$ and the set of complexes $\mathscr{C} = \{A, B, C, D, W, X, Y, Z, 0, A + B, 2A + D, C + W, B + D\}$. It contains the ten true reactions in (1.1) and, when all species are deemed to be in the feed, the sixteen flow reactions in (2.1).

In general, we denote by $\mathscr{R}_t$ the set of true reactions, by $\mathscr{R}_f$ the set of feed reactions, and by $\mathscr{R}_o$ the set of outflow reactions.

DEFINITION 2.2. *A* mass-action system *is a reaction network* $(\mathscr{S}, \mathscr{C}, \mathscr{R})$ *taken together with an element* $k \in \mathbb{R}^{\mathscr{R}}_+$. *The number* $k_{y\to y'}$ *is the* rate constant *of the reaction* $y \to y' \in \mathscr{R}$.

For two vectors in $\bar{\mathbb{R}}^{\mathscr{S}}_+$, say $u = \sum_{s\in\mathscr{S}} u_s s$ and $v = \sum_{s\in\mathscr{S}} v_s s$, we denote $u^v = \prod_{s\in\mathscr{S}} (u_s)^{v_s}$. Here we use the convention that $0^0 = 1$.

We will now show how, by using the notation above, we can express the system of ODEs associated to a reaction network as a very compact formula. For example, note that the term $k_{A+B\to C} c_A c_B$ on the first line in (1.3) can be written as $k_{y\to y'} c^y$; here $y = A + B$ and $y' = C$ are regarded as vectors in $\mathbb{R}^{\mathscr{S}}$, where $\mathscr{S}$ is the set of species. Also, the term $-2k_{2A+D\to Y} c_A^2 c_D$ on the second line in (1.3) can be written as $-2k_{y\to y'} c^y$, where $y = 2A + D, y' = Y$.

If we look for *all* appearances of $k_{2A+D\to Y} c_A^2 c_D$ in (1.3), we notice that they take place in equations corresponding to species A, D, Y, i.e., exactly the species that appear in the complexes y, y'. Moreover, the coefficient of each species in the reaction $2A + D \to Y$ is equal (up to sign) to the coefficient of the monomial $k_{2A+D\to Y} c_A^2 c_D$ in the equation corresponding to that species. The sign is minus for species in y and plus for species in y'. Therefore the factor $k_{2A+D\to Y} c_A^2 c_D$ contributes to the right side of the (vector) ODE precisely as the term $k_{y\to y'} c^y (y' - y)$, where $y = 2A + D, y' = Y$.

Then we get the following two definitions (see [14, 19]).

DEFINITION 2.3. *The* species-formation-rate function *(or simply the* rate function*) for a mass-action system* $(\mathscr{S}, \mathscr{C}, \mathscr{R}, k)$ *is the function* $r(\cdot, k) : \bar{\mathbb{R}}^{\mathscr{S}}_+ \to \mathbb{R}^{\mathscr{S}}$, *defined by*

$$r(c, k) = \sum_{y\to y'\in\mathscr{R}} k_{y\to y'} c^y (y' - y).$$

DEFINITION 2.4. *The* system of differential equations associated to a mass-action system $(\mathscr{S}, \mathscr{C}, \mathscr{R}, k)$ *is given by*

$$\dot{c} = r(c, k).$$

We see here again that the reaction network $(\mathscr{S}, \mathscr{C}, \mathscr{R})$ and the vector k uniquely determine the system of differential equations associated to a mass-action system.

DEFINITION 2.5. *A* positive equilibrium *of a mass-action system* $(\mathscr{S}, \mathscr{C}, \mathscr{R}, k)$ *is an element* $c \in \mathbb{R}^{\mathscr{S}}_+$ *such that* $r(c, k) = 0$.

DEFINITION 2.6. *We say that a reaction network* $(\mathscr{S}, \mathscr{C}, \mathscr{R})$ has the capacity for multiple positive equilibria *if there exist* $k \in \mathbb{R}^{\mathscr{R}}_+$, $a \in \mathbb{R}^{\mathscr{S}}_+$, $b \in \mathbb{R}^{\mathscr{S}}_+$, $a \neq b$, *such that* $r(a, k) = r(b, k) = 0$.

To formulate the following definition recall that we have $\mathscr{R} = \mathscr{R}_f \cup \mathscr{R}_o \cup \mathscr{R}_t$, where $\mathscr{R}_f \cup \mathscr{R}_o$ is the set of flow reactions ($\mathscr{R}_f$ is the set of feed reactions, $\mathscr{R}_o$ is the set of outflow reactions), and $\mathscr{R}_t$ is the set of true reactions.

DEFINITION 2.7. *Given a chemical reaction network* $\mathcal{N} = (\mathscr{S}, \mathscr{C}, \mathscr{R})$, *its* associated polynomial function $p_{\mathcal{N}}(\cdot,\cdot) : \mathbb{R}_+^{\mathscr{S}} \times \mathbb{R}_+^{\mathscr{R}_t \cup \mathscr{R}_o} \to \mathbb{R}^{\mathscr{S}}$ *is*

$$p_{\mathcal{N}}(c,k) = \sum_{y \to y' \in \mathscr{R}_t \cup \mathscr{R}_o} k_{y\to y'} c^y (y - y').$$

Note that

$$r(c,k) = -p_{\mathcal{N}}(c,k) + \sum_{y \to y' \in \mathscr{R}_f} k_{y\to y'} c^y (y' - y).$$

With $\mathscr{S}_f$ denoting the set of species in the feed stream, note also that

$$\sum_{y \to y' \in \mathscr{R}_f} k_{y\to y'} c^y (y' - y) = \sum_{s \in \mathscr{S}_f} k_{0\to s} s.$$

The last equation results from the fact that $\mathscr{R}_f = \{0 \to s : s \in \mathscr{S}_f\}$ and, for $y = 0$, $c^y = 1$. Finally, note that the equilibrium equation $r(c,k) = 0$ is equivalent to

$$p_{\mathcal{N}}(c,k) = \sum_{s \in \mathscr{S}_f} k_{0\to s} s, \tag{2.2}$$

and the sum on the right side of (2.2) is constant. Therefore, if the polynomial function $c \to p_{\mathcal{N}}(c,k)$ is injective for every value of the parameter $k \in \mathbb{R}_+^{\mathscr{R}_t \cup \mathscr{R}_o}$, then there cannot exist multiple positive equilibria.

DEFINITION 2.8. *We say that a chemical reaction network* $\mathcal{N} = (\mathscr{S}, \mathscr{C}, \mathscr{R})$ *is* injective *if the polynomial function* $c \to p_{\mathcal{N}}(c,k)$ *is injective for all* $k \in \mathbb{R}_+^{\mathscr{R}_t \cup \mathscr{R}_o}$.

Remark 2.9. Our consideration of CFSTRs suggests that, for the outflow reactions (i.e., those of the form $s \to 0$), we should require the rate constants to be *identical* for all $s \in \mathscr{S}$. Recall that these rate constants were identified with g/V, which we set to 1. It would appear then that our requirement of injectivity of $p_{\mathcal{N}}(\cdot,k)$ for *all* $k \in \mathbb{R}_+^{\mathscr{R}_t \cup \mathscr{R}_o}$ is stronger than it need be for the application we have in mind. However, it is not hard to show that if $p_{\mathcal{N}}(\cdot,k)$ is injective for all $k \in \mathbb{R}_+^{\mathscr{R}_t \cup \mathscr{R}_o}$ satisfying the restriction $k_{s\to 0} = 1$ for every $s \in \mathscr{S}$, then $p_{\mathcal{N}}(\cdot,k)$ is injective for all $k \in \mathbb{R}_+^{\mathscr{R}_t \cup \mathscr{R}_o}$.

In fact, suppose that, for some $k^* \in \mathbb{R}_+^{\mathscr{R}_t \cup \mathscr{R}_o}$, there are distinct $a^* \in \mathbb{R}_+^{\mathscr{S}}$, $b^* \in \mathbb{R}_+^{\mathscr{S}}$ such that $p_{\mathcal{N}}(a^*,k^*) = p_{\mathcal{N}}(b^*,k^*)$. Now choose $k^\#$, $a^\#$, $b^\#$ as follows:

$$\begin{aligned} a_s^\# &= a_s^* k_{s\to 0}^* \quad \forall s \in \mathscr{S}, \\ b_s^\# &= b_s^* k_{s\to 0}^* \quad \forall s \in \mathscr{S}, \\ k_{y\to y'}^\# &= k_{y\to y'}^* / \prod_{s \in \mathscr{S}} {k_{s\to 0}^*}^{y_s} \quad \forall y \to y' \in \mathscr{R}_t \cup \mathscr{R}_o. \end{aligned}$$

Then $k_{s\to 0}^\# = 1$ for all $s \in \mathscr{S}$, and $p_{\mathcal{N}}(a^\#,k^\#) = p_{\mathcal{N}}(b^\#,k^\#)$. This is to say that if $p_{\mathcal{N}}(\cdot,k^*)$ is not injective for some unrestricted k^*, then there is a restricted $k^\#$ such that $p_{\mathcal{N}}(\cdot,k^\#)$ also fails to be injective.

3. Characterizations of the injectivity property. In this section we prove some equivalent characterizations of the injectivity property that make it possible to check whether a given reaction network is injective by using standard computer algebra software.

Recall that for each reaction network $\mathcal{N} = (\mathscr{S}, \mathscr{C}, \mathscr{R})$ we defined its associated polynomial function $p_{\mathcal{N}}(\cdot,\cdot) : \mathbb{R}^{\mathscr{S}}_+ \times \mathbb{R}^{\mathscr{R}_t \cup \mathscr{R}_o}_+ \to \mathbb{R}^{\mathscr{S}}$.

THEOREM 3.1. *A reaction network $\mathcal{N}$ is injective if and only if we have*

$$\det\left(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k)\right) \neq 0 \quad \forall c \in \mathbb{R}^{\mathscr{S}}_+ \text{ and } \forall k \in \mathbb{R}^{\mathscr{R}_t \cup \mathscr{R}_o}_+. \tag{3.1}$$

Remark 3.1. Note that there is some similarity between this theorem and the Jacobian conjecture[2] over the field of real numbers, since we are concluding injectivity from the nonsingularity of the Jacobian of a polynomial function. Of course, there are also important differences, e.g., the fact that the domain of the function $p(\cdot,k)$ is restricted to $\mathbb{R}^{\mathscr{S}}_+$, and (3.1) holds for all positive values of the parameter k.

Proof. We will show a chain of equivalences from the negation of (3.1) to the noninjectivity of $p_{\mathcal{N}}(\cdot,k)$. The derivative of $p_{\mathcal{N}}(\cdot,k)$ at some point $c \in \mathbb{R}^{\mathscr{S}}_+$ is a linear transformation from $\mathbb{R}^{\mathscr{S}}$ to $\mathbb{R}^{\mathscr{S}}$. According to [18], the result of applying the derivative of $p_{\mathcal{N}}(\cdot,k)$ to an arbitrary vector $\gamma \in \mathbb{R}^{\mathscr{S}}$ can be written as

$$\left(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k)\right)(\gamma) = \sum_{y \to y' \in \mathscr{R}_t \cup \mathscr{R}_o} k_{y\to y'} c^y (y * \gamma)(y - y'),$$

where "$*$" is a special scalar product in $\mathbb{R}^{\mathscr{S}}$, defined by

$$v * w = \sum_{s \in \mathscr{S}} (v_s w_s / c_s).$$

(Here we use the fact that all the components of c are strictly positive.) Note that to say that (3.1) is *not true* is equivalent to

$$\left(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k)\right)(\gamma) = 0 \text{ for some } c \in \mathbb{R}^{\mathscr{S}}_+, k \in \mathbb{R}^{\mathscr{R}_t \cup \mathscr{R}_o}_+, \gamma \in \mathbb{R}^{\mathscr{S}}, \gamma \neq 0, \tag{3.2}$$

which is also equivalent to

$$\sum_{y \to y' \in \mathscr{R}_t \cup \mathscr{R}_o} k_{y\to y'} c^y (y * \gamma)(y - y') = 0 \text{ for some } c \in \mathbb{R}^{\mathscr{S}}_+,\ k \in \mathbb{R}^{\mathscr{R}_t \cup \mathscr{R}_o}_+ \text{ and some } \gamma \in \mathbb{R}^{\mathscr{S}},\ \gamma \neq 0. \tag{3.3}$$

Using the change of variables $\eta_{y\to y'} = k_{y\to y'} c^y$ and $\delta_s = \gamma_s / c_s$ we notice that condition (3.3) is equivalent to

$$\sum_{y \to y' \in \mathscr{R}_t \cup \mathscr{R}_o} \eta_{y\to y'} (y \cdot \delta)(y - y') = 0 \text{ for some } \eta \in \mathbb{R}^{\mathscr{R}_t \cup \mathscr{R}_o}_+ \text{ and some } \delta \in \mathbb{R}^{\mathscr{S}},\ \delta \neq 0, \tag{3.4}$$

where "$\cdot$" is the usual scalar product in $\mathbb{R}^{\mathscr{S}}$. The condition (3.4) in turn is equivalent to

$$\sum_{y \to y' \in \mathscr{R}_t \cup \mathscr{R}_o} K_{y\to y'} (e^{y\cdot\delta} - 1)(y - y') = 0 \text{ for some } K \in \mathbb{R}^{\mathscr{R}_t \cup \mathscr{R}_o}_+ \text{ and some } \delta \in \mathbb{R}^{\mathscr{S}},\ \delta \neq 0, \tag{3.5}$$

[2]The Jacobian conjecture over the field of real numbers says that if a polynomial function $f : \mathbb{R}^m \to \mathbb{R}^m$ has nonsingular Jacobian everywhere, then f is injective. This conjecture was proved to be false in [28].

since the signs of $y \cdot \delta$ and $e^{y\cdot\delta} - 1$ are the same, regardless of the value of $y \cdot \delta$. Then the condition (3.5) is equivalent to

$$\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o} K_{y\to y'}\left(\frac{b^y}{a^y}-1\right)(y-y') = 0 \text{ for some } K \in \mathbb{R}_+^{\mathscr{R}_t\cup\mathscr{R}_o} \text{ and some } a \neq b \in \mathbb{R}_+^{\mathscr{S}} \tag{3.6}$$

via another change of variables, such that $\frac{b_s}{a_s} = e^{\delta_s}$ for all $s \in \mathscr{S}$. Note that $a \neq b$ if and only if $\delta \neq 0$. Condition (3.6) is equivalent to

$$\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o} \kappa_{y\to y'}(b^y - a^y)(y-y') = 0 \text{ for some } \kappa \in \mathbb{R}_+^{\mathscr{R}_t\cup\mathscr{R}_o} \text{ and some } a \neq b \in \mathbb{R}_+^{\mathscr{S}}, \tag{3.7}$$

where $\kappa_{y\to y'} = \frac{K_{y\to y'}}{a^y}$. Now, note that this is equivalent to saying that for some value of $\kappa \in \mathbb{R}_+^{\mathscr{R}_t\cup\mathscr{R}_o}$ the function $p_{\mathcal{N}}(\cdot,\kappa)$ is not injective on $\mathbb{R}_+^{\mathscr{S}}$. Therefore, we showed that the reaction network $\mathcal{N}$ is injective if and only if (3.1) is true. ∎

It is perhaps worthwhile to consider a small example, which is easily worked by hand. Consider network (3.8):

$$A + B \rightleftharpoons C, \quad A \to 2B. \tag{3.8}$$

The system of CFSTR differential equations associated to (3.8) is

$$\begin{aligned}
\dot{c}_A &= c_A^f - c_A - k_{A+B\to C}c_Ac_B + k_{C\to A+B}c_C - k_{A\to 2B}c_A,\\
\dot{c}_B &= c_B^f - c_B - k_{A+B\to C}c_Ac_B + k_{C\to A+B}c_C + 2k_{A\to 2B}c_A,\\
\dot{c}_C &= c_C^f - c_C + k_{A+B\to C}c_Ac_B - k_{C\to A+B}c_C,
\end{aligned} \tag{3.9}$$

where we supposed that $g/V = 1$. If we now again look for equilibria and rearrange terms, we get

$$\begin{aligned}
c_A^f &= c_A + k_{A+B\to C}c_Ac_B - k_{C\to A+B}c_C + k_{A\to 2B}c_A,\\
c_B^f &= c_B + k_{A+B\to C}c_Ac_B - k_{C\to A+B}c_C - 2k_{A\to 2B}c_A,\\
c_C^f &= c_C - k_{A+B\to C}c_Ac_B + k_{C\to A+B}c_C.
\end{aligned} \tag{3.10}$$

Therefore the associated polynomial function for the reaction network (3.8) is

$$\begin{aligned}
p(c,k) = (&c_A + k_{A+B\to C}c_Ac_B - k_{C\to A+B}c_C + k_{A\to 2B}c_A,\\
&c_B + k_{A+B\to C}c_Ac_B - k_{C\to A+B}c_C - 2k_{A\to 2B}c_A,\\
&c_C - k_{A+B\to C}c_Ac_B + k_{C\to A+B}c_C).
\end{aligned} \tag{3.11}$$

Then, for the reaction network (3.8), we have

$$\begin{aligned}
&\det\left(\frac{\partial p}{\partial c}(c,k)\right)\\
&= \det\begin{bmatrix} 1 + k_{A+B\to C}c_B + k_{A\to 2B} & k_{A+B\to C}c_A & -k_{C\to A+B}\\ k_{A+B\to C}c_B - 2k_{A\to 2B} & 1 + k_{A+B\to C}c_A & -k_{C\to A+B}\\ -k_{A+B\to C}c_B & -k_{A+B\to C}c_A & 1 + k_{C\to A+B}\end{bmatrix}\\
&= 1 + k_{C\to A+B} + k_{A+B\to C}c_A + k_{A+B\to C}c_B\\
&\quad + 3k_{A\to 2B}k_{A+B\to C}c_A + k_{A\to 2B}k_{C\to A+B}.
\end{aligned} \tag{3.12}$$

Notice that in (3.12) all coefficients[3] of the monomials in the expansion of the determinant are 1, except the coefficient of $k_{A\to 2B}k_{A+B\to C}c_A$, which is 3. In particular, they are all positive numbers. Therefore, in this case, $\det(\frac{\partial f}{\partial c}(c,k)) > 0$ for all $c \in \mathbb{R}^n_+$ and for all $k \in \mathbb{R}^m_+$, so the reaction network (3.8) is injective as well.

Compare this to $\det(\frac{\partial p}{\partial c}(c,k))$ for the polynomial function associated to the reaction network (vii) in Table 1.1, which is

$$\det\left(\frac{\partial p}{\partial c}(c,k)\right) = 1 + k_{2A+B\to 3A}c_A^2 - 2k_{2A+B\to 3A}c_Ac_B + 3k_{3A\to 2A+B}c_A^2. \tag{3.13}$$

The reaction network (vii) in Table 1.1 does admit multiple positive equilibria, and, as we have seen above, the determinant of the Jacobian of its associated polynomial function has a monomial with a negative coefficient.

Now we are in a position to review and elaborate further on what was said in Remark 1.2. It is worth repeating here that $\det(\frac{\partial p}{\partial c}(c,k))$ can be calculated using currently available computer algebra software and that the result of such a computation will sometimes have hundreds or even thousands of terms, each a monomial in the (positive) species concentrations and the (positive) rate constants. It is remarkable that, more often than not, *all* such monomials will have positive coefficients, so that $\det(\frac{\partial p}{\partial c}(c,k))$ is positive for all positive c and all positive k (recall Table 1.2). Indeed, for large networks the positivity of the monomial coefficients can also be checked with computer algebra software. In this way, Theorem 3.1 provides a (surprisingly robust) way to ensure that a given network is injective and, therefore, incapable of multiple positive equilibria.

In fact, Theorem 3.1 provides the information that networks (ii) and (iv) in Table 1.1 cannot give rise to multiple positive equilibria. On the other hand, Theorem 3.1 by itself stands silent on the capacity for multiple positive equilibria of the very similar networks (i) and (iii). In section 4 we will discuss extensions of Theorem 3.1 that do give information about networks (i) and (iii).

For polynomials in general, it is not necessary that each coefficient be positive in order for the polynomial to take strictly positive values for all positive values of the variables. (The polynomial $x^2 - xy + y^2$ is, of course, an elementary counterexample.) On the other hand, we will show that, for the class of polynomials considered here, positivity of the numerical coefficients is also necessary if positive values of the polynomial are to result for all positive values of the variables (i.e., the species concentrations and rate constants). In turn, this will imply that positivity of all (nonzero) coefficients is not only sufficient but also necessary for a network's injectivity (see Theorem 3.3).

In the following theorem we draw a relationship between the underlying network of chemical reactions and the numerical coefficients in the expansion of $\det(\frac{\partial p}{\partial c}(c,k))$. This relationship will have some importance not only here but also in a subsequent paper, in which we describe large classes of networks for which all (nonzero) coefficients are positive.

THEOREM 3.2. *Consider some reaction network $\mathcal{N}$ with n species. Then for each coefficient in the expansion of* $\det(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k))$ *there is a set of n reactions $\{y_1 \to y_1', \ldots, y_n \to y_n'\}$ (taken from the true and outflow reactions) such that the coefficient is equal to*

$$\det([y_1, \ldots, y_n])\det([y_1 - y_1', \ldots, y_n - y_n']). \tag{3.14}$$

[3]We are looking at coefficients of monomials with respect to the coordinates of c and k.

Moreover, for each choice of n reactions such that (3.14) *is not zero, there is a corresponding coefficient in the expansion of* $\det(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k))$.

Proof. Recall that, with the notation from the proof of Theorem 3.1, we have

$$\left(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k)\right)(\gamma) = \sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o} \eta_{y\to y'}(y*\gamma)(y-y'),$$

where $\eta_{y\to y'} = k_{y\to y'}c^y$. With $\{e_1,\dots,e_n\}$ denoting the canonical basis of $\mathbb{R}^{\mathscr{S}}$, we have

$$\det\left(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k)\right)$$
$$= \det\left[\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o} \eta_{y\to y'}(y*e_1)(y-y'),\dots,\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o} \eta_{y\to y'}(y*e_n)(y-y')\right],$$

and, according to the definition of "$*$", it follows that

$$\left(\prod_{i=1}^n c_i\right)\det\left(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k)\right)$$
$$= \det\left[\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o} \eta_{y\to y'}(y\cdot e_1)(y-y'),\dots,\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o} \eta_{y\to y'}(y\cdot e_n)(y-y')\right].$$

Therefore the coefficients in the expansion of $\det(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k))$ are exactly the coefficients in the expansion of

$$\det\left[\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o} \eta_{y\to y'}(y\cdot e_1)(y-y'),\dots,\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o} \eta_{y\to y'}(y\cdot e_n)(y-y')\right].$$

Note now that each term in the expansion of the determinant above is a scalar multiple of a product of the form $\prod_{i=1}^n \eta_{y_i\to y_i'}$, where $y_1\to y_1',\dots,y_n\to y_n'$ are some reactions in $\mathscr{R}_t\cup\mathscr{R}_o$.

Let us look at some fixed set $\{y_1\to y_1',\dots,y_n\to y_n'\}\subset\mathscr{R}_t\cup\mathscr{R}_o$. With S_n denoting the set of all permutations of $\{1,\dots,n\}$, the coefficient of $\prod_{i=1}^n \eta_{y_i\to y_i'}$ in the expansion of the determinant above is

$$\begin{aligned}
&\sum_{\sigma\in S_n} \det[(y_{\sigma(1)}\cdot e_1)(y_{\sigma(1)}-y'_{\sigma(1)}),\dots,(y_{\sigma(n)}\cdot e_n)(y_{\sigma(n)}-y'_{\sigma(n)})]\\
&= \sum_{\sigma\in S_n} \det[y^1_{\sigma(1)}(y_{\sigma(1)}-y'_{\sigma(1)}),\dots,y^n_{\sigma(n)}(y_{\sigma(n)}-y'_{\sigma(n)})]\\
&= \sum_{\sigma\in S_n} y^1_{\sigma(1)}y^2_{\sigma(2)}\dots y^n_{\sigma(n)}\det[(y_{\sigma(1)}-y'_{\sigma(1)}),\dots,(y_{\sigma(n)}-y'_{\sigma(n)})]\\
&= \sum_{\sigma\in S_n} y^1_{\sigma(1)}y^2_{\sigma(2)}\dots y^n_{\sigma(n)}\mathrm{sgn}(\sigma)\det[(y_1-y_1'),\dots,(y_n-y_n')]\\
&= \left(\sum_{\sigma\in S_n} y^1_{\sigma(1)}y^2_{\sigma(2)}\dots y^n_{\sigma(n)}\mathrm{sgn}(\sigma)\right)\det[(y_1-y_1'),\dots,(y_n-y_n')]\\
&= \det[y_1,\dots,y_n]\det[(y_1-y_1'),\dots,(y_n-y_n')].
\end{aligned}$$

Therefore all coefficients in the expansion of $\det(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k))$ are of the form

$$\det[y_1,\dots,y_n]\det[y_1-y_1',\dots,y_n-y_n']$$

for some set $\{y_1\to y_1',\dots,y_n\to y_n'\}\subset\mathscr{R}_t\cup\mathscr{R}_o$. ∎

Remark 3.2. In a future paper we will use the result of Theorem 3.2 to explain why, for large classes of reaction networks, all coefficients of the monomials in the expansion of $\det(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k))$ are nonnegative (i.e., our empirical observation).

Note that Theorem 3.2 gives us a way of computing the coefficients of $\det(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k))$ one by one. In particular, it suggests a simple parallel computation algorithm for checking injectivity.

We prove now that the injectivity of a reaction network $\mathcal{N}$ is completely characterized by the signs of the coefficients of $\det(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k))$.

THEOREM 3.3. *A reaction network $\mathcal{N}$ is injective if and only if all the coefficients in the expansion of* $\det(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k))$ *are nonnegative.*

Proof. Suppose that all the coefficients in the expansion of $\det(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k))$ are nonnegative. We want to show that $\mathcal{N}$ is injective.

Consider the function $f:\mathbb{R}_+^{\mathscr{S}}\times\mathbb{R}_+^{\mathscr{R}_t\cup\mathscr{R}_o}\to\mathbb{R}$ defined by

$$f(c,k)=\left(\prod_{s\in\mathscr{S}}c_s\right)\det\left(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k)\right).$$

Note that f vanishes if and only if $\det(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k))$ vanishes. As in the proof of Theorem 3.2, the terms in the expansion of $f(c,k)$ are of the form

$$\det[y_1,\dots,y_n]\det[y_1-y_1',\dots,y_n-y_n']\left(\prod_{i=1}^{n}\eta_{y_i\to y_i'}\right),\tag{3.15}$$

where $\eta_{y\to y'}=k_{y\to y'}c^y$, and with each term corresponding to some choice of n reactions from the set $\mathscr{R}_t\cup\mathscr{R}_o$. Note that $\prod_{s\in\mathscr{S}}c_s$ and $\prod_{i=1}^{n}\eta_{y_i\to y_i'}$ are strictly positive, since c and k are regarded to have strictly positive coordinates. Then, to show injectivity, it is enough to show that there exists some set $\{y_1\to y_1',\dots,y_n\to y_n'\}\subset\mathscr{R}_t\cup\mathscr{R}_o$ such that $\det[y_1,\dots,y_n]\det[y_1-y_1',\dots,y_n-y_n']\neq 0$. But if we just choose the set $\{y_1\to y_1',\dots,y_n\to y_n'\}$ to be $\mathscr{R}_o$, we have

$$\det[y_1,\dots,y_n]\det[y_1-y_1',\dots,y_n-y_n']=1,$$

because, up to a permutation, $y_i=e_i$ and $y_i'=0$ for $i=1,\dots,n$. Therefore (3.1) is true, and, according to Theorem 3.1, $\mathcal{N}$ is injective.

Suppose now that $\mathcal{N}$ is injective. We want to show that all the coefficients in the expansion of $\det(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k))$ are nonnegative. Of course, the coefficients in the expansion of $\det(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k))$ are the same as the coefficients in the expansion of $f(c,k)$. We will show that all the coefficients in the expansion of $f(c,k)$ are nonnegative. Note that $f(c,k)$ equals a homogeneous polynomial of degree n of the coordinates of η. Note also that, since we can write the terms in the expansion of $f(c,k)$ as in (3.15), it follows that each monomial in this expansion contains a product $\prod_{i=1}^{n}\eta_{y_i\to y_i'}$ for some set of n *distinct* reactions $\{y_1\to y_1',\dots,y_n\to y_n'\}$, and there is no other monomial with the same set of n reactions. Suppose now that there is some monomial with a negative coefficient in the expansion of $f(c,k)$. Then, by choosing some $\eta\in\mathbb{R}_+^{\mathscr{R}_t\cup\mathscr{R}_o}$ such that the coordinates of η that appear in the negative monomial are very large, and

all other coordinates of η are very small (i.e., very close to zero), we conclude that f takes a negative value somewhere in its domain. Similarly, by using a monomial with a positive coefficient (for example, the monomial with the coefficient "1" that we mentioned above) we conclude that f takes a positive value somewhere in its domain. Since the domain of f is connected, it follows that f is zero somewhere in its domain. According to Theorem 3.1, this contradicts the hypothesis that $\mathcal{N}$ is injective. Therefore there cannot exist any monomial with a negative coefficient in the expansion of $f(c,k)$, so there cannot exist any monomial with a negative coefficient in the expansion of $\det(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k))$. □

Remark 3.3. Theorem 3.3 allows us to show that, although injectivity is *sufficient* to conclude that a reaction network does not admit multiple positive equilibria, *it is not a necessary condition*. One such example is the reaction network (vi) in Table 1.1. Indeed, that reaction network does not admit multiple positive equilibria but has

$$\det\left(\frac{\partial p}{\partial c}(c,k)\right) = 1 + k_{A+B\to 2A}c_A - k_{A+B\to 2A}c_B + 2k_{2A\to A+B}c_A,$$

which does have one negative coefficient, so the network is not injective.

Remark 3.4. Theorems 3.2 and 3.3 imply that, given a reaction network with n species and m reactions, it is only the structure of its subnetworks of exactly n reactions (some of which could be outflow reactions) that dictates whether the reaction network is injective or not. Also, given some reaction network that *does* admit multiple positive equilibria, Theorems 3.2 and 3.3 allow us to pinpoint the subnetwork or subnetworks that create the capacity for multiple positive equilibria as exactly the ones for which the product of determinants $\det[y_1,\dots,y_n]\det[y_1-y_1',\dots,y_n-y_n']$ is negative. Or, consider some finite family of reaction networks, each containing exactly n species. According to Theorem 3.2, we can enumerate all possible "bad" subnetworks in that family (i.e., subnetworks that have exactly n reactions, and for which the product of determinants above is negative). Then, in that family, only the reaction networks that contain a copy of some "bad" subnetwork can have the capacity for multiple positive equilibria.

Remark 3.5. Up to now we have considered only reaction networks where all species are in the outflow. If $\mathcal{N}$ is a reaction network such that not all species are in the outflow, but there are n reactions $\{y_1\to y_1',\dots,y_n\to y_n'\}$ in $\mathcal{N}$ (some of which could be outflow reactions) such that $\det[y_1,\dots,y_n]\det[y_1-y_1',\dots,y_n-y_n']>0$, then Theorem 3.3 remains valid.

4. Sufficient conditions for existence of multiple positive equilibria. Recall that, as we mentioned in section 1, the injectivity property is not a necessary condition for the absence of multiple positive equilibria (see also Remark 3.3). In other words, if a network $\mathcal{N}$ is not injective, this does not imply that $\mathcal{N}$ has the capacity for multiple positive equilibria. Theorems 4.1 and 4.2 below say that if $\mathcal{N}$ is not injective and satisfies an additional condition, then $\mathcal{N}$ does have the capacity for multiple positive equilibria. We begin with a lemma.

LEMMA 4.1. *Let $\mathcal{N}=(\mathscr{S},\mathscr{C},\mathscr{R})$ be some reaction network (augmented to include the flow reactions). Suppose that there is some $c\in\mathbb{R}_+^{\mathscr{S}}$ and some $k\in\mathbb{R}_+^{\mathscr{R}_t\cup\mathscr{R}_o}$ such that*

$$\det\left(\frac{\partial p_{\mathcal{N}}}{\partial c}(c,k)\right)=0$$

and

$$\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o} k_{y\to y'}c^y(y-y')\in\mathbb{R}^{\mathscr{S}}_+.$$

Then $\mathcal{N}$ does have the capacity for multiple positive equilibria.

Proof. The reaction network $\mathcal{N}$ admits multiple positive equilibria if and only if there is some $\kappa\in\mathbb{R}^{\mathscr{R}_t\cup\mathscr{R}_o}_+$ and some $a\neq b\in\mathbb{R}^{\mathscr{S}}_+$ such that

$$\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o}\kappa_{y\to y'}a^y(y-y')=\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o}\kappa_{y\to y'}b^y(y-y')=c^f$$

for some $c^f\in\bar{\mathbb{R}}^{\mathscr{S}}_+$ (recall Remark 1.1). Consider $\eta\in\mathbb{R}^{\mathscr{R}_t\cup\mathscr{R}_o}_+$ such that $\eta_{y\to y'}=k_{y\to y'}c^y$ for each reaction $y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o$, where c and k are as in the theorem statement. Then, as in the proof of Theorem 3.1, there exists some $\delta\in\mathbb{R}^{\mathscr{S}},\ \delta\neq 0$, such that

$$\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o}\eta_{y\to y'}(y\cdot\delta)(y-y')=0.$$

Consider $a\in\mathbb{R}^{\mathscr{S}}_+$ given by $a_s=1$ for every $s\in\mathscr{S}$, and consider $b\in\mathbb{R}^{\mathscr{S}}_+$ given by $b_s=e^{\delta_s}$ for every $s\in\mathscr{S}$. Note that $\delta\neq 0$ implies $a\neq b$. Denote by $\kappa\in\mathbb{R}^{\mathscr{R}_t\cup\mathscr{R}_o}_+$ the vector given by $\kappa_{y\to y'}=\frac{y\cdot\delta}{e^{y\cdot\delta}-1}\eta_{y\to y'}$ for all $y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o$ with $y\cdot\delta\neq 0$, and $\kappa_{y\to y'}=\eta_{y\to y'}$ for all $y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o$ with $y\cdot\delta=0$.

Then we have

$$\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o}\kappa_{y\to y'}(b^y-a^y)(y-y')=\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o}\eta_{y\to y'}(y\cdot\delta)(y-y')=0.$$

Note that, without loss of generality, we can suppose that the norm of δ is very small. On the other hand we have

$$\begin{aligned}&\lim_{\delta\to 0}\left(\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o}\kappa_{y\to y'}a^y(y-y')\right)\\&=\lim_{\delta\to 0}\left(\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o}\kappa_{y\to y'}(y-y')\right)\\&=\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o}\eta_{y\to y'}(y-y')\\&=\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o}k_{y\to y'}c^y(y-y')\ \in\ \mathbb{R}^{\mathscr{S}}_+.\end{aligned}$$

Then, for small enough δ, it follows that $\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o}\kappa_{y\to y'}a^y(y-y')\ \in\ \mathbb{R}^{\mathscr{S}}_+$. ∎

THEOREM 4.1. *Consider some reaction network $\mathcal{N}=(\mathscr{S},\mathscr{C},\mathscr{R})$ (augmented to include the flow reactions). For $\eta\in\mathbb{R}^{\mathscr{R}_t\cup\mathscr{R}_o}_+$ let $T_\eta:\mathbb{R}^{\mathscr{S}}\to\mathbb{R}^{\mathscr{S}}$ be defined by*

$$T_\eta(\delta)=\sum_{y\to y'\in\mathscr{R}_t\cup\mathscr{R}_o}\eta_{y\to y'}(y\cdot\delta)(y-y'),\tag{4.1}$$

and let

$$f(\eta)=\det(T_\eta).\tag{4.2}$$

Suppose that for some $\eta^* \in \mathbb{R}_+^{\mathscr{R}_t \cup \mathscr{R}_o}$ *we have*

$$f(\eta^*) < 0, \tag{4.3}$$

$$\sum_{y \to y' \in \mathscr{R}_t \cup \mathscr{R}_o} \eta^*_{y \to y'}(y - y') \in \mathbb{R}_+^{\mathscr{S}}. \tag{4.4}$$

Then $\mathcal{N}$ *has the capacity for multiple positive equilibria.*

Proof. Consider some $\eta^\# \in \mathbb{R}_+^{\mathscr{R}_t \cup \mathscr{R}_o}$ such that for all $y \to y' \in \mathscr{R}_o$ the number $\eta^\#_{y \to y'}$ is very large, and for all $y \to y' \in \mathscr{R}_t$ the number $\eta^\#_{y \to y'}$ is very small. Then condition (4.4) holds for $\eta^\#$, and, for reasons similar to those in the proof of Theorem 3.3, $f(\eta^\#) > 0$.

Suppose now that there is some $\eta^* \in \mathbb{R}_+^{\mathscr{R}_t \cup \mathscr{R}_o}$ such that both (4.3) and (4.4) are true. Because the set of vectors η that satisfy (4.4) is convex, and because the function f is continuous, it follows that on the line segment that connects $\eta^\#$ and η^* there will be some $\tilde{\eta}$ such that condition (4.4) holds for $\tilde{\eta}$, and $f(\tilde{\eta}) = 0$.

Now, for some fixed $\tilde{c} \in \mathbb{R}_+^{\mathscr{S}}$, choose $\tilde{k} \in \mathbb{R}_+^{\mathscr{R}_t \cup \mathscr{R}_o}$ such that $\tilde{\eta}_{y \to y'} = \tilde{k}_{y \to y'} \tilde{c}^y$ for all $y \to y' \in \mathscr{R}_t \cup \mathscr{R}_o$. According to the chain of equivalences in the proof of Theorem 3.1 (from (3.1) to (3.4)) we have

$$\det\left(\frac{\partial p_{\mathcal{N}}}{\partial c}(\tilde{c}, \tilde{k})\right) = 0.$$

Also, note that

$$\sum_{y \to y' \in \mathscr{R}_t \cup \mathscr{R}_o} \tilde{k}_{y \to y'} \tilde{c}^y (y - y') \in \mathbb{R}_+^{\mathscr{S}}.$$

Then the hypothesis of Lemma 4.1 is satisfied, so its conclusion is also true. ∎

Remark 4.1. Note that if some vector $\eta^* \in \mathbb{R}_+^{\mathscr{R}_t \cup \mathscr{R}_o}$ satisfies (4.3) and (4.4), then $\lambda \eta^*$ also satisfies (4.3) and (4.4) for any positive number λ. Therefore, if there is some η^* that satisfies (4.3) and (4.4) and has all coordinates positive, then there is some η^{**} that satisfies (4.3) and (4.4) and has all coordinates positive *and of total sum* 1. Then Theorem 4.1 can be implemented by considering the polynomial optimization problem (4.5)–(4.8), with linear constraints on a compact domain:

$$\text{minimize} f(\eta) \tag{4.5}$$

subject to the constraints

$$\eta_{y \to y'} \geq \varepsilon \quad \forall y \to y' \in \mathscr{R}_t \cup \mathscr{R}_o, \tag{4.6}$$

$$\sum_{y \to y' \in \mathscr{R}_t \cup \mathscr{R}_o} \eta_{y \to y'} = 1, \tag{4.7}$$

$$\sum_{y \to y' \in \mathscr{R}_t \cup \mathscr{R}_o} \eta_{y \to y'}(y_s - y'_s) \geq \varepsilon \quad \forall s \in \mathscr{S}, \tag{4.8}$$

where ε is some very small positive number. Note that, from the point of view of applying Theorem 4.1, it is enough to find *some* vector η^* satisfying (4.6)–(4.8) and such that $f(\eta^*) < 0$ (i.e., we don't need to find the global minimum, as we are just interested in knowing if the minimum is negative).

THEOREM 4.2. *Consider some reaction network $\mathcal{N} = (\mathscr{S}, \mathscr{C}, \mathscr{R})$ (augmented to include the flow reactions). Suppose that there is a set of n reactions $\{y_1 \to y_1', \ldots, y_n \to y_n'\}$ (where n is the number of species) such that*

$$\det(y_1, \ldots, y_n) \det(y_1 - y_1', \ldots, y_n - y_n') < 0 \tag{4.9}$$

and

$$\sum_{i=1}^{n} \eta_i (y_i - y_i') \in \mathbb{R}^{\mathscr{S}}_+ \text{ for some positive numbers } \eta_1, \ldots, \eta_n. \tag{4.10}$$

Then $\mathcal{N}$ does have the capacity for multiple positive equilibria.

Proof. Consider some $\eta^* \in \mathbb{R}^{\mathscr{R}_t \cup \mathscr{R}_o}_+$ such that for all $y \to y' \in \{y_1 \to y_1', \ldots, y_n \to y_n'\}$ the number $\eta^*_{y \to y'}$ is very large, and for all other $y \to y'$ the number $\eta^*_{y \to y'}$ is very small. Then, as in the proof of Theorem 3.3, it follows that $f(\eta^*) < 0$, because in the expansion of $\det(T_{\eta^*})$ the negative term corresponding to the subnetwork $\{y_1 \to y_1', \ldots, y_n \to y_n'\}$ dominates all other terms.

Suppose in particular that $\eta^*_{y_i \to y_i'} = \lambda \eta_i$ for some (very large) number λ. Then (4.4) holds for this η^*, since $\sum_{i=1}^n \eta_i (y_i - y_i') \in \mathbb{R}^{\mathscr{S}}_+$, and $\sum_{y \to y' \in \mathscr{R}_t \cup \mathscr{R}_o} \eta^*_{y \to y'} (y - y')$ is very close to $\lambda \sum_{i=1}^n \eta_i (y_i - y_i')$. Therefore we can apply Theorem 4.1. ∎

Remark 4.2. Note that if some numbers $\eta_1, \ldots, \eta_n$ satisfy $\sum_{i=1}^n \eta_i (y_i - y_i') \in \mathbb{R}^{\mathscr{S}}_+$, then the numbers $\lambda\eta_1, \ldots, \lambda\eta_n$ satisfy $\sum_{i=1}^n \lambda\eta_i (y_i - y_i') \in \mathbb{R}^{\mathscr{S}}_+$ for any positive number λ. Then, when implementing Theorem 4.2, we can replace condition (4.10) with the systems of inequalities

$$\eta_i \geq 1 \quad \text{for } i = 1, \ldots, n, \tag{4.11}$$

$$\sum_{i=1}^{n} \eta_i (y_{is} - y_{is}') \geq 1 \quad \forall s \in \mathscr{S}. \tag{4.12}$$

Remark 4.3. For networks (i), (iii), (v) in Table 1.1 the less powerful but easily applied Theorem 4.2 already affirms the capacity for multiple positive equilibria. For network (vii) Theorem 4.1 affirms the capacity for multiple positive equilibria, while Theorem 4.2 stands silent.

Remark 4.4. Suppose that we are given a reaction network $\mathcal{N}$ having n species, and we would like to know if $\mathcal{N}$ has the capacity for multiple positive equilibria (in the isothermal homogeneous CFSTR context). An algorithm that investigates this problem proceeds as follows: First, check[4] if there is any subnetwork of n reactions such that (4.9) holds. If (4.9) is false for all such subnetworks, then, according to Theorems 3.2 and 3.3, $\mathcal{N}$ does not have the capacity for multiple positive equilibria. If one or more subnetworks of $\mathcal{N}$ satisfy (4.9), then check[5] if Theorem 4.2 applies for any such subnetwork. If Theorem 4.2 remains indecisive, then try to apply the more computationally intensive method given by Theorem 4.1 and described in Remark 4.1.

[4]For example, this can be done by computing $\det(\frac{\partial p_{\mathcal{N}}(c,k)}{\partial c})$ in order to recover the coefficients for the various subnetworks (recall Theorem 3.2).

[5]This will be very easy, since we only have to check the feasibility of the system of linear inequalities (4.11) and (4.12).

5. Concluding remarks. We believe that the theorems presented here have broad utility in deciding the capacity of a complex mass-action system to engender multiple positive steady states in a homogeneous isothermal CFSTR context. That these techniques should be robust relies heavily on our assertion that, despite the presence of hundreds or even thousands of terms in the expansion of $\det(\frac{\partial p_{\mathcal{N}}(c,k)}{\partial c})$ for a complex reaction network, it will typically be the case that all (nonzero) coefficients are positive. (When there are negative coefficients for a given network, they will typically be very few in number.) Although we have given examples to support this assertion, we have not, in this paper, tried to explain why positivity of the coefficients is to be expected broadly. Nor have we tried to identify those aspects of reaction network structure that give rise to negative coefficients. We intend to take up these questions in a future paper. There we will show how certain representations of reaction networks in graph-theoretical terms give surprisingly rapid and incisive information.

Acknowledgment. The authors are grateful for support from the United States National Science Foundation.

REFERENCES

[1] B. Aguda, *A quantitative analysis of the kinetics of the G_2 DNA damage checkpoint system*, Proc. Natl. Acad. Sci., 96 (1999), pp. 11352–11357.
[2] B. Aguda and Y. Tang, *The kinetic origins of the restriction point in the mammalian cell cycle*, Cell Proliferation, 32 (1999), pp. 321–335.
[3] R. Aris, *Elementary Chemical Reactor Analysis*, Dover Publications, Mineola, NY, 2000.
[4] J. Bailey, *Complex biology with no parameters*, Nature Biotech., 19 (2001), pp. 503–504.
[5] N. Barkai and S. Leibler, *Robustness in simple biochemical networks*, Nature, 387 (1997), pp. 913–917.
[6] E. Beretta, F. Vetrano, F. Solimano, and C. Lazzari, *Some results about nonlinear chemical systems represented by trees and cycles*, Bull. Math. Bio., 41 (1979), pp. 641–664.
[7] M. Chaves and E. Sontag, *State-estimators for chemical reaction networks of Feinberg-Horn-Jackson zero deficiency type*, European J. Control, 8 (2002), pp. 343–359.
[8] B. Clarke, *Stability of complex reaction networks*, Adv. Chem. Phys., 42 (1980), pp. 1–213.
[9] P. De Kepper and J. Boissonade, *Theoretical and experimental analysis of phase diagrams and related dynamical properties in the Belousov-Zhabotinskii system*, J. Chem. Phys., 75 (1981), pp. 189–195.
[10] A. Europa, A. Gambhir, P.-C. Fu, and W.-S. Hu, *Multiple steady states with distinct cellular metabolism in continuous culture of mammalian cells*, Biotech. Bioengrg., 67 (2000), pp. 25–34.
[11] M. Feinberg, *Complex balancing in general kinetic systems*, Arch. Rational Mech. Anal., 49 (1972), pp. 187–194.
[12] M. Feinberg and F. Horn, *Dynamics of open chemical systems and the algebraic structure of the underlying reaction network*, Chem. Engrg. Sci., 29 (1974), pp. 775–787.
[13] M. Feinberg, *Mathematical aspects of mass action kinetics*, in Chemical Reactor Theory: A Review, N. Amundson and L. Lapidus, eds., Prentice-Hall, Englewood Cliffs, NJ, 1977, pp. 1–78.
[14] M. Feinberg, *Lectures on Chemical Reaction Networks*, written version of lectures given at the Mathematical Research Center, University of Wisconsin, Madison, WI, 1979. Available online from www.chbmeng.ohio-state.edu/~feinberg/LecturesOnReactionNetworks.
[15] M. Feinberg, *Chemical oscillations, multiple equilibria, and reaction network structure*, in Dynamics of Reactive Systems, W. Stewart, W. Rey, and C. Conley, eds., Academic Press, New York, 1980, pp. 59–130.
[16] M. Feinberg, *Chemical reaction network structure and the stability of complex isothermal reactors.* I. *The deficiency zero and deficiency one theorems*, Chem. Engrg. Sci., 42 (1987), pp. 2229–2268.
[17] M. Feinberg, *Chemical reaction network structure and the stability of complex isothermal reactors.* II. *Multiple steady states for networks of deficiency one*, Chem. Engrg. Sci., 43 (1988), pp. 1–225.

[18] M. Feinberg, *Existence and uniqueness of steady states for a class of chemical reaction networks*, Arch. Rational Mech. Anal., 132 (1995), pp. 311–370.
[19] M. Feinberg, *Multiple steady states for chemical reaction networks of deficiency one*, Arch. Rational Mech. Anal., 132 (1995), pp. 371–406.
[20] M. Feinberg, *Some recent results in chemical reaction network theory*, in Patterns and Dynamics in Reactive Media, R. Aris, D. G. Aronson, and H. L. Swinney, eds., IMA Vol. Math. Appl., Springer, Berlin, 37 (1991), pp. 43–70.
[21] M. Feinberg, *The Chemical Reaction Network Toolbox*, Version 1.02 (1995), and Version 1.1 (1999, with P. Ellison). Available for download from http://www.chbmeng.ohio-state.edu/~feinberg/crnt.
[22] C. Forst, *Molecular evolution of catalysis*, J. Theoret. Bio., 205 (2000), pp. 409–431.
[23] T. Gardner, C. Cantor, and J. Collins, *Construction of a genetic toggle switch in Escherichia coli*, Nature, 403 (2000), pp. 339–342.
[24] F. Horn, *Necessary and sufficient conditions for complex balancing in chemical kinetics*, Arch. Rational Mech. Anal., 49 (1972), pp. 172–186.
[25] F. Horn and R. Jackson, *General mass action kinetics*, Arch. Rational Mech. Anal., 47 (1972), pp. 81–116.
[26] *Law of Mass Action*, in Britannica Concise Encyclopedia, Encyclopedia Britannica, 2004. Available online from http://concise.britannica.com/ebc/article?eu=396792
[27] T. Leib, D. Rumschitzki, and M. Feinberg, *Multiple steady states in complex isothermal CFSTRs:* I. *General Considerations*, Chem. Engrg. Sci., 43 (1988), pp. 321–328.
[28] S. Pinchuk, *A counterexample to the real Jacobian conjecture*, Math. Z., 217 (1994), pp. 1–4.
[29] D. Rumschitzki and M. Feinberg, *Multiple steady states in complex isothermal CFSTRs:* II. *Homogeneous reactors*, Chem. Engrg. Sci., 43 (1988), pp. 329–337.
[30] P. Schlosser and M. Feinberg, *A theory of multiple steady states in isothermal homogeneous CFSTRs with many reactions*, Chem. Engrg. Sci., 49 (1994), pp. 1749–1767.
[31] P. Schlosser, *A Graphical Determination of the Possibility of Multiple Steady States in Complex Isothermal CFSTRs*, Ph.D. thesis, Department of Chemical Engineering, University of Rochester, Rochester, NY, 1988.
[32] J. Shu, P. Wu, and M. Shuler, *Biostability in ammonium-limited cultures of Escherichia Coli B/r*, Chem. Engrg. Comm., 58 (1987), pp. 185–194.
[33] P. Schuster, *Landscapes and molecular evolution*, Phys. D, 107 (1997), pp. 351–365.
[34] E. Sontag, *Structure and stability of certain chemical networks and applications to the kinetic proofreading model of T-cell receptor signal transduction*, IEEE Trans. Automat. Control, 46 (2001), pp. 1028–1047.

SIAM J. APPL. MATH.
Vol. 65, No. 5, pp. 1547–1575

AN ANALYSIS OF HIGHER ORDER BOUNDARY CONDITIONS FOR THE WAVE EQUATION*

JULIEN DIAZ† AND PATRICK JOLY†

Abstract. Thanks to the use of the Cagniard–De Hoop method, we derive an analytic solution in the time domain for the half-space problem associated with the wave equation with Engquist–Majda higher order boundary conditions. This permits us to derive new convergence results when the order of the boundary condition tends to $+\infty$, as well as error estimates. The theory is illustrated by numerical results.

Key words. Cagniard–De Hoop method, absorbing boundary conditions, wave equation, error estimates, convergence, Green's function

AMS subject classifications. 35L05, 35L20, 35C05, 65M99, 44E99

DOI. 10.1137/S0036139903436145

1. Introduction.

The design of accurate absorbing boundary conditions (ABCs) for the numerical calculation of waves in the time domain is already an old subject since the major work of Engquist and Majda [11], [12] in the late 1970s. Their main contribution was the construction and analysis of a hierarchy of local boundary conditions for the wave equation. Let us concentrate on the two-dimensional (2D) acoustic wave equation:

$$\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} - \Delta u = 0, \qquad x = (x_1, x_2) \in \mathbb{R}^2, \quad t > 0. \tag{1.1}$$

Assuming that the data (initial data) of the problem are supported in the upper half-space $\mathbb{R}^2_+ = \{x_2 > 0\}$, it is natural to try to reduce the effective numerical computations to this half-space by imposing adequate absorbing boundary conditions on the artificial boundary $\Gamma = \partial\mathbb{R}^2_+$. In [11], Engquist and Majda proposed the following condition (the integer N is a parameter meant to be large):

$$\mathcal{B}^N u = 0 \quad \text{on } \Gamma, \tag{1.2}$$

where the operators

$$\mathcal{B}^N = \mathcal{B}^N\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}\right), \quad N \geq 0,$$

are a family of homogeneous differential operators defined inductively by

$$\begin{cases} \mathcal{B}^0 = 2I, \quad \mathcal{B}^1 = \dfrac{\partial}{\partial t} - c\dfrac{\partial}{\partial x_2}, \\ \mathcal{B}^{N+1} = \dfrac{\partial}{\partial t}\mathcal{B}^N - \dfrac{c^2}{4}\dfrac{\partial^2}{\partial x_1^2}\mathcal{B}^{N-1}. \end{cases} \tag{1.3}$$

*Received by the editors October 14, 2003; accepted for publication (in revised form) October 18, 2004; published electronically May 12, 2005.
http://www.siam.org/journals/siap/65-5/43614.html
†INRIA Rocquencourt, Domaine de Voluceau, B. P. 105, 78153 Le Chesnay, France (julien.diaz@inria.fr, patrick.joly@inria.fr).

Note that $\mathcal{B}^N$ can be rewritten in the form

$$\mathcal{B}^N = S_{N-1}\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial x_1}\right)\frac{\partial}{\partial x_2} - Q_N\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial x_1}\right),$$

where Q_N and S_{N-1} are homogeneous polynomials of two variables of respective degrees N and $N-1$. In particular, $\mathcal{B}^N$ remains of first order with respect to x_2; the condition (1.2) can be seen as a Dirichlet-to-Neumann (or impedance) type boundary condition since it can be formally rewritten as

$$\frac{\partial u}{\partial x_2} - \frac{Q_N(\frac{\partial}{\partial t}, \frac{\partial}{\partial x_1})}{S_{N-1}(\frac{\partial}{\partial t}, \frac{\partial}{\partial x_1})}u = 0. \tag{1.4}$$

Remark 1.1. For smooth solutions (up to the boundary Γ) of the wave equation, the boundary condition (1.2) can be rewritten in terms of t and x_2 derivatives only. Indeed, $\mathcal{B}^N$ is obviously even with respect to the x_1 variable, and, thanks to the wave equation, the second order derivative with respect to x_1 can be replaced by t and x_2 derivatives:

$$\frac{\partial^2}{\partial x_1^2} \longrightarrow \frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x_2^2}.$$

As a consequence, one can show that [21]

$$\mathcal{B}^N u = 0 \Longleftrightarrow \widetilde{\mathcal{B}}^N u = 0, \quad \widetilde{\mathcal{B}}^N = \left(\frac{\partial}{\partial t} - c\frac{\partial}{\partial x_2}\right)^N. \tag{1.5}$$

This remark will be useful in section 3.

Let us recall that the initial boundary value problems (IBVPs) for linear hyperbolic systems (1.2), or (1.4) or (1.5), are constructed as an approximation of an exact or transparent boundary condition

$$\mathcal{B}u = 0, \quad \mathcal{B} = \frac{\partial}{\partial x_2} - \mathcal{L}, \tag{1.6}$$

where $\mathcal{L}$ is a pseudodifferential operator in (x_1, t) whose symbol is known explicitly. More precisely, if one uses the Laplace–Fourier transform in the (t, x_1) plane (see (3.3) and (3.2)),

$$\varphi(x_1, t) \to \widetilde{\varphi}(k, s),$$

one has the formula

$$\widetilde{\mathcal{L}\varphi}(k, s) = \left(k^2 + \frac{s^2}{c^2}\right)^{\frac{1}{2}}\varphi(k, s), \qquad \mathcal{R}e\left(k^2 + \frac{s^2}{c^2}\right)^{\frac{1}{2}} \geq 0. \tag{1.7}$$

This comes from the fact that if u is a solution of the wave equation in the lower half-space $\mathbb{R}^2_- = \{x_2 < 0\}$ with zero initial data, its partial Laplace–Fourier transform in t and x_1, $\widetilde{u}(k, x_2, s)$, satisfies

$$\widetilde{u}(k, x_2, s) = \widetilde{u}(k, 0, s)e^{(k^2 + \frac{s^2}{c^2})^{\frac{1}{2}}x_2}, \quad x_2 \leq 0,$$

which yields in particular

$$\frac{d\widetilde{u}}{dx_2}(k,0,s) - \left(k^2 + \frac{s^2}{c^2}\right)^{\frac{1}{2}} \widetilde{u}(k,0,s) = 0.$$

The presence of the square root in the symbol of $\mathcal{L}$ makes the operator $\mathcal{L}$, and consequently the boundary condition (1.6), nonlocal in space and time, which is a priori very unpleasant from the numerical point of view. The approximate condition simply comes from a rational approximation of the symbol of $\mathcal{L}$ in such a way that the resulting boundary condition can be expressed in terms of differential operators, which is much more tractable from the numerical point of view. If one writes

$$\left(k^2 + \frac{s^2}{c^2}\right)^{\frac{1}{2}} = \frac{s}{c}\left(1 + \frac{c^2k^2}{s^2}\right)^{\frac{1}{2}},$$

the problem is reduced to the rational approximation of the function of one variable:

$$f(z) = (1+z^2)^{\frac{1}{2}}.$$

Noticing that $f(z)$ is a solution of the fixed point equation,

$$f(z) = 1 + \frac{z^2}{1+f(z)}, \tag{1.8}$$

one obtains a rational approximation (or continuous fraction expansion) of $f(z)$ with the following fixed point algorithm:

$$f_{n+1}(z) = 1 + \frac{z^2}{1+f_n(z)}, \quad f_1(z) = 1. \tag{1.9}$$

The condition (1.2) is obtained by replacing in (1.6) $\mathcal{L}$ by $\mathcal{L}_N$, whose symbol is $\frac{s}{c} f_N(\frac{ck}{s})$. It is then relatively easy to deduce the induction formula (1.3) from (1.9).

Remark 1.2. It is easy to show that the sequence $f_n(z)$ converges, for large n, to $f(z)$ only if $|z| < 1$. Moreover, the convergence is uniform and exponential in any compact of the unit circle. For $|z| > 1$, $f_n(z)$ converges to $-f(z)$, which is the other solution of (1.8). However, it is not a problem for the application to ABCs, as will be shown in this paper.

It is also well known that (1.9) provides the sequence of $\{n, n-1\}$ (for even n) and $\{n-1, n-1\}$ (for odd n) Padé approximants [3] of $f(z)$ at the neighborhood of the origin:

$$f_2(z) = 1 + \frac{z^2}{2}, \qquad f_3(z) = 1 + \frac{2z^2}{4+z^2}, \dots ;$$

and in particular one has

$$f_n(z) - f(z) = O((z^2)^N), \quad z \to 0. \tag{1.10}$$

That is why the boundary condition (1.2) is known as the Engquist–Majda condition of order $2m$. Equation (1.10) shows that the rational approximation of the symbol of $\mathcal{L}$ given by (1.9) is better for the small values of ck/s, which has a physical interpretation (see below).

During recent years, abundant research has been devoted to various improvements (including in particular "better" rational approximations) and extensions (including in

particular the application to other wave equations) of the Engquist–Majda conditions. It is not possible to give here an exhaustive bibliography, and we will refer the reader to recent review papers on the subject by Hagström [18], [19] and Givoli [13]. In the last decade, alternative solutions have been progressively developed and, especially, researchers have tried to promote again the use of exact nonlocal boundary either by using specific geometries for the absorbing boundaries, as in the works by Grote and Keller [14], [15], or by exploiting the recent progress in rapid algorithms (multipoles) and rational approximation, as in the work of Alpert, Greengard, and Hagström [2], [1]. Approximately during the same period, the introduction by Bérenger of the perfectly matched layers (PMLs) technique [6], [5] partly revolutionized the subject. The philosophy here is to replace the absorbing boundary with an absorbing layer (or sponge layer) which is such that any wave propagating in the computational domain is transmitted to the absorbing layer without being reflected. This method quickly attracted many researchers in different fields of application, in particular because of its good practical performances and its easy implementation.

All these methods (local higher order ABCs, nonlocal ABCs, and PMLs) have been successfully introduced in a number of different computational codes. Of course, for anybody who wants to use such codes, the natural question is, Which is the best method for the absorption of waves? Our feeling is that there is no universal answer to such a question and that a response should include some criteria: nature of the problem to be addressed, accuracy, speed of calculation, ease of implementation, long time behavior, etc. However, even with given criteria, the answer would be delicate, in particular because no complete and fair comparison has been done between the three classes of methods. The first reason, which is easy to understand, is that there is probably nobody in the world who has implemented the three methods with the same amount of care. The second reason is a lack of analysis, which is hard in particular if one is interested in getting convincing error estimates. The objective of the present paper is to fill partially this lack in the theory in the case of local ABCs.

Of course, there are a lot of available theoretical results about higher order ABCs. The first question that was raised by Engquist and Majda in their original papers was that of the well-posedness of the IBVP "wave equation—ABC." This is not a trivial question since it is known that polynomial approximations of degree greater than 2 of the function $f(z)$ (as, for instance, the successive Taylor approximations of f around 0) give rise to strongly ill-posed problems. However, thanks to the well-known Kreiss theory (the so-called normal mode analysis [26], [22]), the stability theory of higher order ABCs is more or less completely understood. In particular, necessary and sufficient conditions were given in [27] about the rational approximations of $f(z)$ in order to ensure the strong well-posedness of the corresponding IBVP (of course, the approximations f_n given by (1.9) fulfill these conditions, as already observed in [12]) and energy estimates (giving rise to stability results, i.e., a priori estimates independent of N) were obtained in [8].

Concerning the accuracy of ABCs, the simplest analysis consists in analyzing the reflection of plane waves, which amounts to studying particular solutions of the following form ($k \in \mathbb{R}$ and $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ are parameters, K is the wave number, while θ represents the angle of incidence of the incident plane wave):

$$u_\theta(x,t) = \exp ik(x_1 \sin\theta - x_2 \cos\theta - ct) + \mathbf{R} \exp ik(x_1 \sin\theta + x_2 \cos\theta - ct), \tag{1.11}$$

where

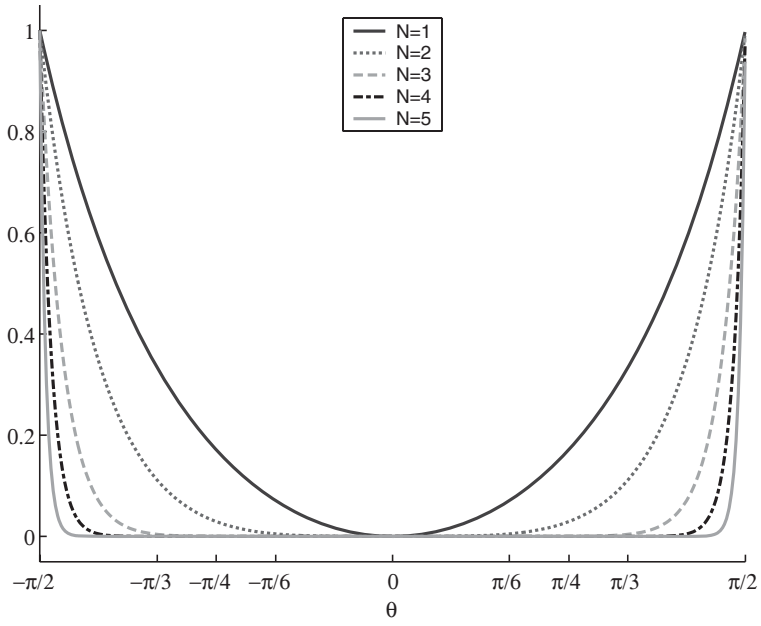


FIG. 1. *The reflection coefficient.*

(i) $\exp ik(x_1 \sin\theta - x_2 \cos\theta - ct)$ is the incident wave,

(ii) $\exp ik(x_1 \sin\theta + x_2 \cos\theta - ct)$ is the reflected wave, $\mathbf{R}$ being the reflection coefficient.

By construction, (1.11) is a solution of the wave equation (1.1). It remains to determine $\mathbf{R}$ in order to satisfy the boundary condition (1.2). The computations show that $\mathbf{R}$ depends only on the angle of incidence θ:

$$\mathbf{R} = \mathbf{R}_N(\theta) \equiv \frac{(f_N - f)(\sin\theta)}{(f_N + f)(\sin\theta)} = (-1)^N \left(\frac{1-\cos\theta}{1+\cos\theta}\right)^N . \tag{1.12}$$

In particular one sees that for any $\theta \in]-\frac{\pi}{2}, \frac{\pi}{2}[$, $\mathbf{R}_N(\theta)$ tends (exponentially fast) to 0 when $N \to +\infty$ while $|\mathbf{R}_N(\pm\frac{\pi}{2})| = 1$ (see also Figure 1). There are much fewer results about convergence and error estimates. In fact, there was no real progress since the initial result of Engquist and Majda, which we are recalling now. They were addressing the following 2D model problem:

$$\begin{cases} \text{Find } v : \mathbb{R}^2_- \times \mathbb{R} \mapsto \mathbb{R} & \text{such that} \\ \dfrac{1}{c^2}\dfrac{\partial^2 v}{\partial t^2} - \Delta v = 0 & \text{in } \mathbb{R}^2_- \times \mathbb{R}^+, \\ v(x_1, 0, t) = g(x_1, t) & \text{on } x_2 = 0, \\ v(x,t) = 0 & \text{for } t < 0. \end{cases} \tag{1.13}$$

One wishes to get a good approximation of v in a domain $\Omega_b = \{x / -b < x_2 < 0\}$ for

given $b > 0$ by putting an ABC on the line $x_2 = -a$, with $a > b$:

$$\begin{cases} \text{Find } v^N : \Omega_a \times \mathbb{R} \mapsto \mathbb{R} & \text{such that} \\ \dfrac{1}{c^2}\dfrac{\partial^2 v^N}{\partial t^2} - \Delta v^N = 0 & \text{in } \Omega_a \times \mathbb{R}^+, \\ v^N(x_1, 0, t) = g(x_1, t) & \text{on } x_2 = 0, \\ \mathcal{B}^N v^N = 0 & \text{on } x_2 = -a, \\ v^N(x, t) = 0 & \text{for } t < 0. \end{cases} \tag{1.14}$$

In (1.14) there are two important parameters: the order N of the boundary condition and the distance a from the source g to the interface. One assumes that the function g is square integrable in both space and time:

$$\int_0^{+\infty} \int_{\mathbb{R}} |g(x_1, t)|^2 \, dx_1 dt < +\infty. \tag{1.15}$$

THEOREM 1.3 (see [12]). *For any $\varepsilon > 0$ and any arbitrarily large integer M, there exist $N_0 = N_0(\varepsilon, M)$ and $a_0 = a_0(\varepsilon, M)$ such that, for any $N \geq N_0$ and $a \geq a_0$,*

$$\int_0^T \int_{\Omega_b} |(v - v^N)(x, t)|^2 \, dx \, dt < \varepsilon \quad \forall T \leq Ma. \tag{1.16}$$

(i) This result is only a convergence result and does not provide an error estimate. Thus it is not a guide for choosing in practice N and a.

(ii) The fact that the result is valid for any time interval of the form $[0, Ma]$ indicates that the result takes into account an arbitrary large number of reflections on the absorbing boundary.

(iii) What is not satisfactory with Theorem 1.3 is the fact that the estimate (1.16) requires a to be sufficiently large. In particular, this does not provide a convergence result when $N \to +\infty$ for fixed a.

(iv) Looking at the proof of the theorem enlightens the need for a sufficiently large. It is not our purpose to reproduce here the proof, but it seems useful to emphasize some points. The idea is to use the Fourier transform in space and time:

$$v(x_1, x_2, t) \to \widetilde{\mathbf{v}}(k, x_2, \omega) = \widetilde{v}(k, x_2, i\omega).$$

One can get an explicit solution for both $\widetilde{\mathbf{v}}$ and $\widetilde{\mathbf{v}}^N$. In particular, we have

$$\left| \begin{array}{l} \widetilde{\mathbf{v}}(k, x_2, \omega) = \widetilde{\mathbf{g}}(k, \omega) \exp(k^2 - \omega^2/c^2)^{\frac{1}{2}} \cdot x_2, \\ (k^2 - \omega^2/c^2)^{\frac{1}{2}} = \begin{cases} \sqrt{k^2 - \omega^2/c^2} & \text{if } k^2 \geq \omega^2/c^2, \\ i\sqrt{\omega^2/c^2 - k^2} & \text{if } k^2 \leq \omega^2/c^2. \end{cases} \end{array} \right.$$

In particular,

(i) if $k^2 < \omega^2/c^2$, the function $x_2 \to \widetilde{\mathbf{v}}(k, x_2, \omega)$ is oscillating: this is the region of propagative modes;

(ii) if $k^2 > \omega^2/c^2$, the function $x_2 \to \widetilde{\mathbf{v}}(k, x_2, \omega)$ is exponentially decaying when $x_2 \to -\infty$: this is the region of evanescent modes.

When one looks at the error $e^N = v^N - v$, its Fourier transform appears as a sum (the sum is a priori infinite but becomes finite if one is interested in times less than Ma) over $j \geq 1$ of terms of the form

$$\mathcal{R}_N \left(\frac{ck}{\omega} \right)^j \cdot \widetilde{\mathbf{g}}(k, \omega) \cdot \exp[(k^2 - \omega^2/c^2)^{\frac{1}{2}} (\pm x_2 + 2ja)],$$

where the *reflection coefficient* $\mathcal{R}_N$ is given by

$$\mathcal{R}_N(\nu) = \frac{(f_N - f)(\nu)}{(f_N + f)(\nu)} \tag{1.17}$$

and satisfies

$$\begin{cases} \mathcal{R}_N(\sigma) \leq 1 & \text{(stability result)}, \\ \mathcal{R}_N(\sigma) \to 0 & \text{for } |\sigma| < 1 \text{ (cf. Remark 1.2)}. \end{cases}$$

Up to technical details (this is in particular where the assumption (1.15) intervenes), the idea of the proof is the following:

(i) In the propagative region $k^2 < \omega^2/c^2$, $|\mathcal{R}_N(\sigma)|^j$ can be made arbitrarily small by choosing N large enough.

(ii) In the evanescent region $k^2 > \omega^2/c^2$, $|\exp[(k^2 - \omega^2/c^2)^{\frac{1}{2}} (\pm x_2 + 2ja)]|$ can be made arbitrarily small by choosing a large enough.

One then concludes with Plancherel's theorem.

Physically, the fact that $f_n(z)$ has nothing to do with $f(z)$ for $|z| > 1$ means that the evanescent modes are not correctly taken into account by the absorbing condition. This is why one needs to have a large enough in order to "kill" the amplitude of the evanescent modes at the boundary $x_2 = a$.

In 1988, Halpern and Rauch proposed a high-frequency analysis in [20]. More recently, an advance was achieved by Hagström (see [17] and [16], [18]), who derived an approximation theory for the approximation of (a class of) pseudodifferential operators, with the aim of applying it to ABCs, based on a new reinterpretation of (1.2) and standard quadrature theory. He obtained error estimates and convergence results only by making N go to $+\infty$ (i.e., without touching the position of the boundary). However, its results were nonuniform in time.

The history of the present work is the following. The Cagniard–De Hoop method is particularly well known in the physics and engineering communities for calculating analytical solutions of time-dependent wave propagation problems, especially in seismology (see [7], [25], [24]). This method permits one, moreover, to establish a link between time domain solutions and harmonic plane waves. Trying to learn something about this method (for a completely different problem), we immediately realized that it could easily be applied to the problem of ABCs and would probably help to get new error estimates. The computations are so simple that it is rather surprising to see that nobody did them before, to our knowledge. This article presents the results we have obtained with this method and may prove to be a useful tool in teaching this subject.

The outline of the paper is as follows. In section 2, we describe the model problem we are dealing with (the half-space problem with a point source) and state our two main results: Theorem 2.1, which provides an explicit solution of the corresponding fundamental solution, and Theorem 2.4, which provides error estimates in the case

of a general source function. These two results show that one can get a convergence result only by letting N go to $+\infty$. In some sense, this shows that the need for large a in Theorem 1.3 is due to the technique used in the proof but does not correspond to a necessity. However, our results in Theorems 2.1 and 2.4 show that increasing the distance from the source to the absorbing boundary helps to get better error estimates. We also pay attention to large time behavior of the error, which has already been the subject of previous research works (see [10], [9], [4]). Sections 3 and 4 are devoted to the proofs of Theorems 2.1 and 2.4. In section 5, we analyze our results in more detail and make the comparison between numerical results and (quasi-)analytical results.

2. Main results. The first result of this paper is an explicit expression of the fundamental solution of the 2D wave equation in the half-space $\mathbb{R}^+_2 = \{x_2 > 0\}$ with higher order ABCs on $\Gamma = \{x_2 = 0\} (= \partial\mathbb{R}^+_2)$. Since the problem is invariant under translations in the x_1-direction, we can restrict ourselves to the case where the source point is

$$x_S = (0, h) \quad \text{with } h > 0. \tag{2.1}$$

The problem we want to solve is

$$\begin{cases} \text{Find } u : \mathbb{R}^2_+ \times \mathbb{R} \mapsto \mathbb{R} & \text{such that} \\ \dfrac{1}{c^2}\dfrac{\partial^2 u}{\partial t^2} - \Delta u = \delta(x - x_S) \times \delta(t) & \text{in } \mathbb{R}^2_+, \\ \mathcal{B}^N u = 0 & \text{on } \Gamma, \\ u(x,t) = 0 & \text{for } t < 0. \end{cases} \tag{2.2}$$

To state our result, it is useful to introduce some notation. Let us define the image source point x^*_S by

$$x^*_S = (0, -h) \tag{2.3}$$

and let us set (see Figure 2)

$$r(x) = |x - x_S|, \qquad r^*(x) = |x - x^*_S|. \tag{2.4}$$

We also define the function $\theta(x)$, $x \in \mathbb{R}^2_+$, by

$$\theta(x) \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[, \quad x - x^*_S = (r^*(x)\sin\theta(x), r^*(x)\cos\theta(x))^t, \tag{2.5}$$

and finally the function $\Phi(x,t)$, $x \in \mathbb{R}^2_+$, $t > 0$, by

$$\Phi(x,t) = \frac{r^*(x)^2\sin^2\theta(x) - (c^2t^2 - r^*(x)^2)}{r^*(x)^2\sin^2\theta(x) + (c^2t^2 - r^*(x)^2)} = \frac{x_1^2 - (c^2t^2 - r^*(x)^2)}{x_1^2 + (c^2t^2 - r^*(x)^2)}. \tag{2.6}$$

We can notice that

$$ct > r^*(x) \Longrightarrow |\Phi(x,t)| < 1.$$

Finally, we recall that the Chebyshev polynomials $P_N(x)$, $N \geq 0$, are defined by

$$P_0(x) = 1, \qquad P_1(x) = x, \qquad P_{N+1}(x) - 2xP_N(x) + P_{N-1}(x) = 0 \tag{2.7}$$

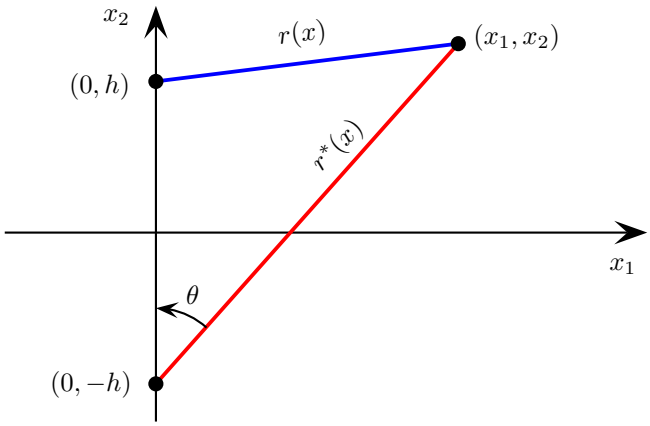


FIG. 2. *Illustration of the notation*

and satisfy

$$\forall x \in [-1,1], \quad P_N(x) = \cos(N \arccos(x)). \tag{2.8}$$

In particular, we see that

$$\forall x \in [-1,1], \quad |P_N(x)| \leq 1.$$

THEOREM 2.1. *The solution $u(x,t) = G^N(x,t)$ of problem (2.2) is given by*

$$G^N(x,t) = G_i(x,t) + G_r^N(x,t), \tag{2.9}$$

where, if H denotes the Heaviside function,

$$\left| \begin{array}{l} G_i(x,t) = \dfrac{1}{2\pi\sqrt{t^2 - \frac{r(x)^2}{c^2}}} H(ct - r(x)), \\ G_r^N(x,t) = -\dfrac{P_N\left(\Phi(x,t)\right)}{2\pi\sqrt{t^2 - \frac{r^*(x)^2}{c^2}}} \left[\dfrac{ct - (x_2+h)}{ct + (x_2+h)}\right]^N H(ct - r^*(x)). \end{array} \right. \tag{2.10}$$

Remark 2.2. The function $G_i(x,t)$, which does not depend on N, is nothing but the restriction to the half-space $\mathbb{R}^2_+$ of the fundamental solution of the 2D wave equation in the whole space. That is why it is called the incident field. Conversely, the field $G_r^N(x,t)$, due to the presence of the boundary Γ, is called the reflected field, which does depend on N.

Remark 2.3. The presence of the factor $H(ct - r^*(x))$ indicates that the reflected field $G_r^N(\cdot,t)$ is compactly supported in the set $\Omega(t) = \Omega_1(t) \cup \Omega_2(t)$ (see Definition 4.6).

Let us now consider the approximation in the upper half-space of the solution u of the 2D wave equation with a "smooth" point source:

$$\begin{cases} \text{Find } u : \mathbb{R}^2 \times \mathbb{R}^+ \mapsto \mathbb{R} & \text{such that} \\ \dfrac{1}{c^2}\dfrac{\partial^2 u}{\partial t^2} - \Delta u = \delta(x - x_S) \times f(t) & \text{in } \mathbb{R}^2 \times \mathbb{R}^+, \\ u(x,0) = 0, \qquad \dfrac{\partial u}{\partial t}(x,0) = 0 & \text{in } \mathbb{R}^2, \end{cases} \tag{2.11}$$

where we assume that the source function $f(t)$ is bounded and has support $[0,T]$ (T can be equal to $+\infty$, which includes the case of a permanent source term) by the solution u^N of the boundary value problem

$$\begin{cases} \text{Find } u^N : \mathbb{R}^2 \times \mathbb{R}^+ \mapsto \mathbb{R} & \text{such that} \\ \dfrac{1}{c^2}\dfrac{\partial^2 u^N}{\partial t^2} - \Delta u^N = \delta(x - x_S) \times f(t) & \text{in } \mathbb{R}^2 \times \mathbb{R}^+, \\ \mathcal{B}^N u^N = 0 & \text{on } \Gamma, \\ u(x,0) = 0, \qquad \dfrac{\partial u}{\partial t}(x,0) = 0 & \text{in } \mathbb{R}^2. \end{cases} \tag{2.12}$$

THEOREM 2.4. *At each point $x \in \mathbb{R}^2_+$, one has the following pointwise estimates:*

(i) *For $\frac{r^*(x)}{c} \le t \le \frac{r^*(x)}{c} + T$ ($\Leftrightarrow x \in \Omega_1(t)$—see (4.6)),*

$$\begin{aligned} &|u(x,t) - u^N(x,t)| \\ &\quad \le \frac{1}{2\pi}\left(\frac{ct - (x_2+h)}{ct + (x_2+h)}\right)^N \mathrm{Log}\left(\frac{ct + \sqrt{c^2t^2 - r^*(x)^2}}{r^*(x)}\right) \|f\|_{L^\infty}. \end{aligned} \tag{2.13}$$

(ii) *For $t > \frac{r^*(x)}{c} + T$ ($\Leftrightarrow x \in \Omega_2(t)$—see (4.6)),*

$$\begin{aligned} &|u(x,t) - u^N(x,t)| \\ &\quad \le \frac{1}{2\pi}\left(\frac{ct - (x_2+h)}{ct + (x_2+h)}\right)^N \mathrm{Log}\left(\frac{ct + \sqrt{c^2t^2 - r^*(x)^2}}{c(t-T) + \sqrt{c^2(t-T)^2 - r^*(x)^2}}\right) \|f\|_{L^\infty}. \end{aligned} \tag{2.14}$$

Moreover, one has the following uniform estimates:

(i) *For $\frac{h}{c} \le t \le \frac{h}{c} + T$,*

$$\|(u - u^N)(\cdot,t)\|_{L^\infty(\mathbb{R}^2_+)} \le \frac{1}{2\pi}\left(\frac{ct-h}{ct+h}\right)^N \mathrm{Log}\left(\frac{t + \sqrt{t^2 - (h/c)^2}}{(h/c)}\right) \|f\|_{L^\infty}. \tag{2.15}$$

(ii) *For $t > \frac{h}{c} + T$,*

$$\|(u - u^N)(\cdot,t)\|_{L^\infty(\mathbb{R}^2_+)} \le \frac{1}{2\pi}\left(\frac{ct-h}{ct+h}\right)^N \mathrm{Log}\left(\frac{t + \sqrt{t^2 - (t-T)^2}}{t - T}\right) \|f\|_{L^\infty}. \tag{2.16}$$

These results lead to the following comments:

(i) The error converges spectrally to 0 (in the uniform norm) when N goes to infinity.

(ii) For given t, the upper bounds in the estimates (2.15) and (2.16) diminish when the distance h from the source to the absorbing boundary increases. This is coherent with the physical intuition and numerical observations.

(iii) Concerning the behavior of the error for large t, if we assume that $T < +\infty$, we observe that the right-hand side in the estimate (2.16) behaves for large t as

$$\frac{1}{2\pi}\sqrt{\frac{2T}{t}},$$

which shows that, for all N and h, the error converges uniformly to 0 when t tends to $+\infty$. On the other hand, when $T = +\infty$, the right-hand side in the estimate (2.15) behaves as

$$\frac{1}{2\pi}\text{Log } t,$$

which a priori authorizes a logarithmic growth on the error when t tends to $+\infty$. This is what happens if $f(t)$ is, for instance, the Heaviside function.

Remark 2.5. We have chosen here to analyze the approximation of a problem associated to a point source. It would not be difficult to adapt Theorem 2.4 (or more precisely its proof) to treat the case of a distributed source term $f(x,t)$ or nonzero initial data u_0 and u_1. In the same way, we have chosen to present L^∞ estimates, which seemed to us more pertinent in practice. However, once again, it is easy to adapt the proof in order to get L^p or energy estimates.

3. Proof of Theorem 2.1. As we have already stated, the formula (2.9), (2.10) results directly from the application of the Cagniard–De Hoop method to problem (2.2). In order to make this paper easily understandable to a reader who is not familiar with this technique, we detail the proof (only some explicit calculations will be omitted). Let us decompose the solution u of (2.2) as

$$u = G_i + u^r,$$

where G_i, given by (2.10), is the fundamental solution of the 2D wave equation. By linearity, it is clear that u^r satisfies

$$\text{(3.1)}\qquad \begin{cases} \text{Find } u^r : \mathbb{R}^2_+ \times \mathbb{R} \mapsto \mathbb{R} & \text{such that} \\ \dfrac{1}{c^2}\dfrac{\partial^2 u^r}{\partial t^2} - \Delta u^r = 0 & \text{in } \mathbb{R}^+_2, \\ \mathcal{B}^N u^r = -\mathcal{B}^N G_i & \text{on } \Gamma, \\ u(x,t) = 0 & \text{for } t < 0. \end{cases}$$

We apply the following successively to u^r:

(i) The Laplace transform in time (s is the dual variable of t):

$$\text{(3.2)}\qquad \widetilde{u}^r(x_1,x_2,s) = \int_0^{+\infty} u^r(x_1,x_2,t)e^{-st}\,dt.$$

(ii) The Fourier transform in the tangential space variable x_1 (k is the dual variable of x_1):

$$\text{(3.3)}\qquad \widehat{u}^r(k,x_2,s) = \int_{-\infty}^{+\infty} \widetilde{u}^r(x_1,x_2,s)e^{ikx_1}\,dx_1.$$

The algorithm for applying the Cagniard–De Hoop method is the following:

1. Compute explicitly $\widehat{u}^r(k,x_2,s)$.
2. Apply the inverse Fourier transform in x_1:

$$\text{(3.4)}\qquad \widetilde{u}^r(x_1,x_2,s) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \widehat{u}^r(k,x_2,s)e^{-ikx_1}\,dx_1.$$

3. Transform (by means of complex analysis methods) the integral (3.4) into an integral of the form

$$\widetilde{u}^r(x_1,x_2,s)=\int_0^{+\infty}F(x_1,x_2,t)e^{-st}\,dt. \tag{3.5}$$

Then, by surjectivity of the Laplace transform, we shall have identified the solution (compare (3.2) and (3.5))

$$u^r(x_1,x_2,t)\equiv F(x_1,x_2,t). \tag{3.6}$$

The first step is straightforward. From the wave equation, we deduce that the function $x_2\mapsto\widehat{u}^r(k,x_2,s)$ satisfies

$$-\frac{d^2\widehat{u}^r}{dx_2^2}+\left(k^2+\frac{s^2}{c^2}\right)\widehat{u}^r=0.$$

Retaining only the solutions that decay when $x_2\to+\infty$ for $\mathcal{R}e(s)\geq 0$, we deduce the existence of a complex-valued function $A(k,s)$ such that

$$\widehat{u}^r(k,x_2,s)=A(k,s)e^{-(k^2+\frac{s^2}{c^2})^{\frac{1}{2}}x_2}, \tag{3.7}$$

where we have chosen to use the determination of the complex square root corresponding to

$$\forall z\in\mathbb{C},\quad \mathcal{R}e\ z^{\frac{1}{2}}\geq 0, \tag{3.8}$$

which corresponds to making the branch cut of $z^{\frac{1}{2}}$ coincide with the semireal axis $\mathcal{I}m\ z=0$, $\mathcal{R}e\ z<0$ (see Figure 3). Since u^r+G_i is smooth for $y<h$, we can use the fact that

$$\mathcal{B}^N(u^r+G_i)=0\Longleftrightarrow\left(\frac{1}{c}\frac{\partial}{\partial t}-\frac{\partial}{\partial x_2}\right)^N(u^r+G_i)=0\quad\text{for } x_2=0. \tag{3.9}$$

On the other hand, it is well known that the Laplace–Fourier transform of G_i is given by

$$\widehat{G}_i(k,x_2,s)=\frac{e^{-(k^2+\frac{s^2}{c^2})^{\frac{1}{2}}|x_2-h|}}{2(k^2+\frac{s^2}{c^2})^{\frac{1}{2}}}. \tag{3.10}$$

Taking into account the form (see (3.7) and (3.10)) of $\widehat{u}^r$ and $\widehat{G}_i$, the boundary condition (3.9) leads to

$$\left(\frac{s}{c}+\left(k^2+\frac{s^2}{c^2}\right)^{\frac{1}{2}}\right)^N A(k,s)+\frac{1}{2}\left(\frac{s}{c}-\left(k^2+\frac{s^2}{c^2}\right)^{\frac{1}{2}}\right)^N\frac{e^{-(k^2+\frac{s^2}{c^2})^{\frac{1}{2}}h}}{(k^2+\frac{s^2}{c^2})^{\frac{1}{2}}}=0.$$

This permits us to compute $A(k,s)$ and finally to get

$$\widehat{u}^r(k,x_2,s)=-R^N\left(k,\frac{s}{c}\right)\frac{e^{-(k^2+\frac{s^2}{c^2})^{\frac{1}{2}}(x_2+h)}}{2(k^2+\frac{s^2}{c^2})^{\frac{1}{2}}}, \tag{3.11}$$

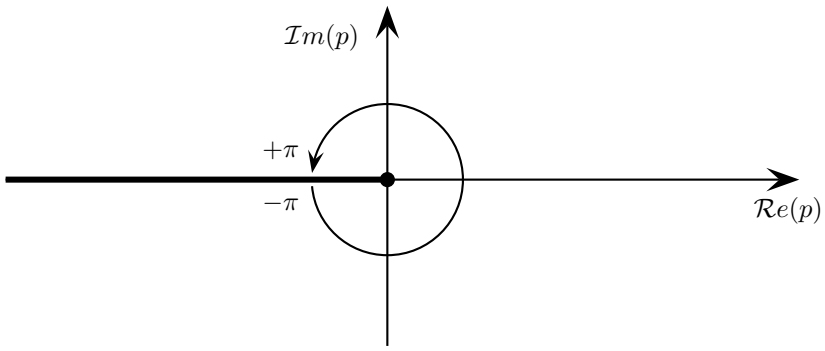


FIG. 3. *The branch cut for the determination of the complex square root.*

where we have set

$$R^N(k,\sigma) = \mathcal{R}_N\left(\frac{k}{\sigma}\right) = \left[\frac{\sigma - (k^2+\sigma^2)^{\frac{1}{2}}}{\sigma + (k^2+\sigma^2)^{\frac{1}{2}}}\right]^N. \tag{3.12}$$

Therefore (this is step 2) we have

$$\widetilde{u}^r(x_1,x_2,s) = -\frac{1}{4\pi}\int_{-\infty}^{+\infty} R^N\left(k,\frac{s}{c}\right)\frac{e^{-(k^2+\frac{s^2}{c^2})^{\frac{1}{2}}(x_2+h)}}{(k^2+\frac{s^2}{c^2})^{\frac{1}{2}}}e^{-ikx_1}\,dk, \tag{3.13}$$

which we would like to transform into an integral of the form (3.5) (this is step 3). This is the part of the approach which is specific to the Cagniard–De Hoop method. We are helped by the following facts:

(i) The integrand in (3.13) is homogeneous in s and k.

(ii) The dependence with respect to x_2 of this integrand is exponential.

First, exploiting the homogeneity property, we apply the change of variable $k = ps/c$ and obtain

$$\begin{aligned}&\widetilde{u}^r(x_1,x_2,s)\\ &= -\frac{1}{4\pi}\int_{-\infty}^{+\infty} R^N(p,1)\frac{e^{-s[(1+p^2)^{\frac{1}{2}}(\frac{x_2+h}{c})+ip\frac{x_1}{c}]}}{(1+p^2)^{\frac{1}{2}}}dp, \quad \left(\equiv \int_{-\infty}^{+\infty}\Psi(p)dp\right).\end{aligned} \tag{3.14}$$

In what follows we fix $(x_1,x_2) \in \mathbb{R}^2_+$ with $x_1 > 0$ (which is not restrictive since the solution we are looking for is clearly even in x_1). We introduce $r^* = r^*(x)$ and $\theta = \theta(x)$ ($\in [0,\pi/2]$ since $x_1 \geq 0$) according to the definitions (2.4) and (2.5). We thus have

$$\begin{cases} x_1 = r^*\sin\theta, \\ x_2 + h = r^*\cos\theta. \end{cases} \tag{3.15}$$

Now the idea is to consider the variable p as a complex variable and to see the integral (3.14) as a contour integral, the contour coinciding with the real axis. If one is able, by a contour deformation, to transform this integral into a contour integral

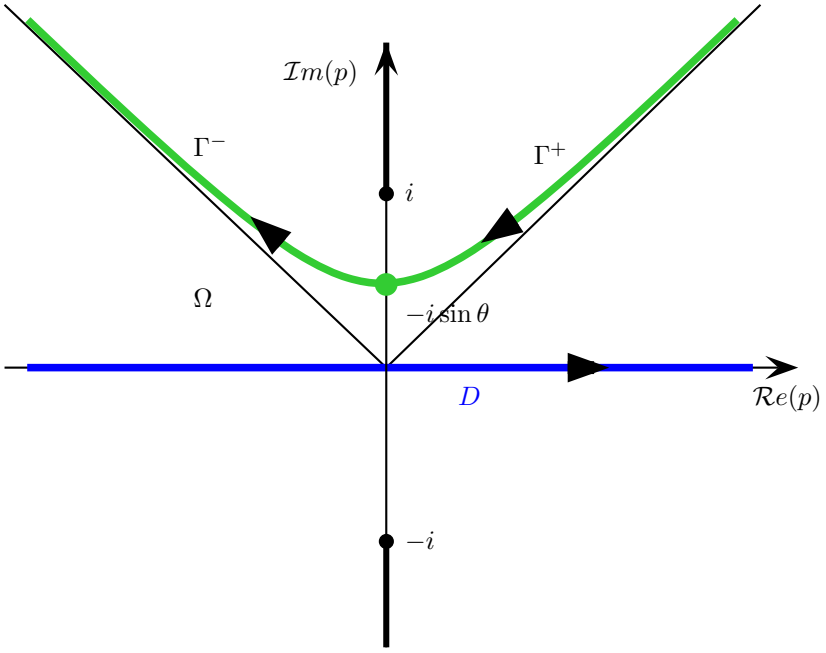


FIG. 4. *The contours* Γ *and* D.

on some curve Γ along which one can use a parametric representation of the form

$$(1+p^2)^{\frac{1}{2}}\left(\frac{x_2+h}{c}\right)+ip\frac{x_1}{c}=t \quad \text{for } t>0, \tag{3.16}$$

we shall have reached our goal. To achieve this, we first remark that the integrand $\Psi(p)$ in (3.14) is an analytic function of p if one excludes the two branch cuts constituted of the two half-lines of purely imaginary numbers whose modulus is greater than 1 (see Figure 4). Then we introduce the so-called Cagniard–De Hoop contour Γ, defined as

$$\begin{cases} \Gamma=\Gamma^+\cup\Gamma^-, \\ \Gamma^\pm=\left\{p=\gamma^\pm(t)\equiv -i\dfrac{ct}{r^*}\sin\theta\pm\cos\theta\sqrt{\dfrac{c^2t^2}{r^{*2}}-1},\ t\geq\dfrac{r^*}{c}\right\}. \end{cases} \tag{3.17}$$

It is clear that the two curves $\Gamma^\pm$ are symmetric with respect to the imaginary axis, and meet at point $-i\sin\theta$ (for $t=r^*/c$). More precisely, it is easy to check that Γ is nothing but the branch of the hyperbola of

$$\frac{Y^2}{\sin^2\theta}-\frac{X^2}{\cos^2\theta}=1 \quad (p=X+iY,\ (X,Y)\in\mathbb{R}^2),$$

which is located in the upper half-space $Y=\mathcal{I}m\ p>0$. Note that this hyperbola does not intersect the two branch cuts of Ψ. All this information is summarized in Figure 4.

In fact to understand where (3.17) comes from, it suffices to remark that $\gamma^\pm(t)$ are nothing but the two roots of (3.16), considered as an equation in p (the calculations are left to the reader).

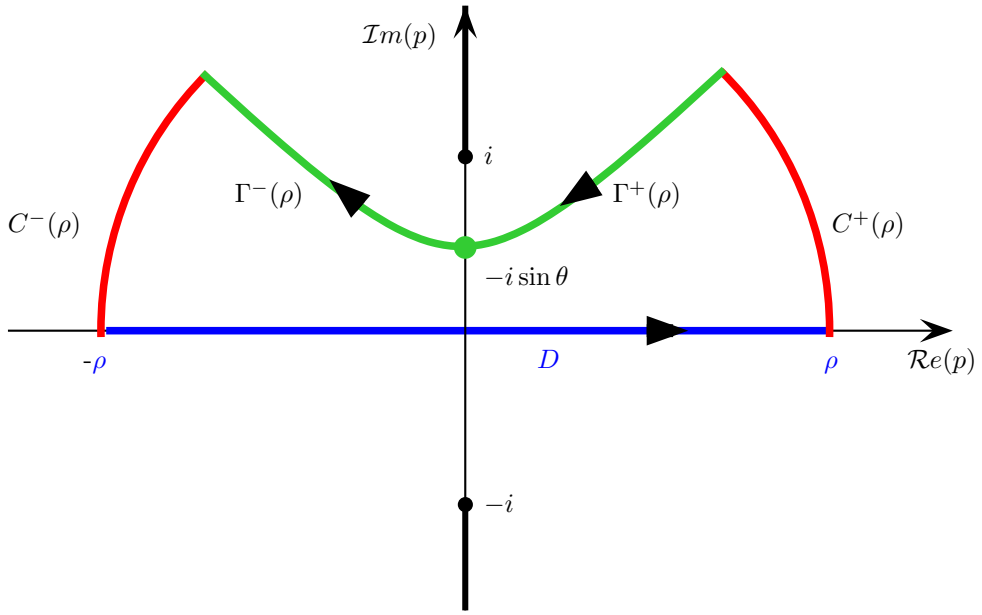


FIG. 5. *The closed contour* $D_\rho \cup C_\rho \cup \Gamma_\rho$.

Let us denote by D the real line and by Ω the connected part of the complex plane delimited by D and Γ. Let $\rho > 0$ be a parameter that is meant to tend to $+\infty$. We set

$$\left| \begin{array}{ll} D_\rho = \{p \in D / |p| \leq \rho\}, & \Gamma_\rho = \{p \in \Gamma / |p| \leq \rho\}, \\ C_\rho = \{p \in \Omega / |p| = \rho\}. & \end{array} \right.$$

Note that C_ρ is made of two arcs of the circle of center 0 and radius ρ that join D_ρ to Γ_ρ in such a way that $D_\rho \cup C_\rho \cup \Gamma_\rho$ is a closed curve. Since $\Psi(p)$ is analytic in Ω, the integral of Ψ along $D_\rho \cup C_\rho \cup \Gamma_\rho$ (we choose the orientation of this path such that the real segment is described in the sense of increasing values—see Figure 5) is identically 0:

$$\int_{D_\rho} \Psi(p)\, dp + \int_{C_\rho} \Psi(p)\, dp + \int_{\Gamma_\rho} \Psi(p)\, dp = 0.$$

Thanks to the choice of the square root, and since $x_2 + h > 0$, the function $\Psi(p)$ decays exponentially to 0 when $\mathcal{I}m\ p$ goes to $+\infty$. As a consequence, it is easy to show that (Jordan's lemma)

$$\lim_{\rho \to +\infty} \int_{C_\rho} \Psi(p)\, dp = 0.$$

Therefore, from (3.14), we deduce

$$\widetilde{u}^r(x_1, x_2, s) = -\frac{1}{4\pi} \int_\Gamma R^N(p, 1) \frac{e^{-s[(1+p^2)^{\frac{1}{2}}(\frac{x_2+h}{c}) + ip\frac{x_1}{c}]}}{(1+p^2)^{\frac{1}{2}}}\, dp.$$

We use the parametrizations $p = \gamma^+(t)$ and $p = \gamma^-(t)$ for $t \geq \frac{r^*(x)}{c}$, respectively, along Γ^+ and Γ^- and remark that

$$\left| \begin{array}{ll} \bullet\ (1+p^2)^{\frac{1}{2}} \left(\dfrac{x_2+h}{c} \right) + ip \dfrac{x_1}{c} = t & \text{(by construction)}, \\ \bullet\ \dfrac{dp}{(\frac{1}{c^2}+p^2)^{\frac{1}{2}}} = \pm \dfrac{dt}{(t^2 - \frac{r^{*2}}{c^2})^{\frac{1}{2}}} & \text{on } \Gamma^{\pm}. \end{array} \right.$$

Therefore, since t goes from $+\infty$ to $\frac{r^*}{c}$ on Γ^+ and from $\frac{r^*}{c}$ to $+\infty$ on Γ^-,

$$\widetilde{u}^r(x_1, x_2, s) = -\frac{1}{4\pi} \int_{\frac{r^*}{c}}^{+\infty} \left[R^N(\gamma^+(t), 1) + R^N(\gamma^-(t), 1) \right] \frac{e^{-st}}{(t^2 - \frac{r^{*2}}{c^2})^{\frac{1}{2}}}\, dt.$$

Finally, observing that $\gamma^-(t)^2 = \overline{\gamma^+(t)}^2$, and using the fact that $\sqrt{\overline{z}} = \overline{\sqrt{z}}$, we deduce that

$$R^N(\gamma^-(t), 1) = \overline{R^N(\gamma^+(t), 1)} \Longrightarrow R^N(\gamma^+(t), 1) + R^N(\gamma^-(t), 1) = 2\mathcal{R}e \left[R^N(\gamma^+(t), 1) \right],$$

which yields

$$\widetilde{u}^r(x_1, x_2, s) = -\frac{1}{2\pi} \int_{\frac{r^*}{c}}^{+\infty} \mathcal{R}e \left[R^N(\gamma^+(t), 1) \right] \frac{e^{-st}}{(t^2 - \frac{r^{*2}}{c^2})^{\frac{1}{2}}}\, dt. \tag{3.18}$$

Thanks to formula (3.12), one has

$$R^N(\gamma^+(t), 1) = \left[R^1(\gamma^+(t), 1) \right]^N, \tag{3.19}$$

while one easily computes that

$$R^1(\gamma^+(t), 1) = \frac{r^* - ct\cos\theta + i\sin\theta\sqrt{c^2t^2 - r^{*2}}}{r^* + ct\cos\theta - i\sin\theta\sqrt{c^2t^2 - r^{*2}}} = \rho_1(t) e^{i\alpha(t)}.$$

Setting $\Phi = \Phi(x, t)$ (cf. (2.6)), one finds that (the calculations—rather tedious but straightforward—are left to the reader)

$$\begin{cases} \rho_1(t) = \dfrac{ct - r^*\cos\theta}{ct + r^*\cos\theta} = \dfrac{ct - (x_2+h)}{ct + (x_2+h)}, \\ \cos\alpha(t) = \dfrac{r^{*2}\sin^2\theta - (c^2t^2 - r^{*2})}{r^{*2}\sin^2\theta + (c^2t^2 - r^{*2})} = \Phi. \end{cases}$$

Therefore, according to (3.19),

$$\mathcal{R}e \left[R^N(\gamma^+(t), 1) \right] = \rho_1(t)^N \cos(N\alpha(t)),$$

that is to say, since $\alpha(t) = \arccos\Phi$,

$$\mathcal{R}e \left[R^N(\gamma^+(t), 1) \right] = \left[\frac{ct - (x_2+h)}{ct + (x_2+h)} \right]^N P_N(\Phi). \tag{3.20}$$

It is then easy to conclude the proof of Theorem 2.1 from (3.18) and (3.20).

4. Proof of Theorem 2.4. Let u and u^N be the respective solutions of (2.11) and (2.12). We introduce the error (or reflected field) e^N defined as

$$e^N = u^N - u. \tag{4.1}$$

To get the pointwise estimates (2.13) and (2.14), we fix $x \in \mathbb{R}^2_+$ and set $r^* = r^*(x)$ and $\theta = \theta(x)$. Obviously $e^N(x,t) = 0$ for $t \le r^*/c$, while for $t > r^*/c$ we deduce from Theorem 2.1 that

$$e^N(x,t) = \int_{\max(\frac{r^*}{c}, t-T)}^{t} G_r^N(x,\tau) f(t-\tau)\, d\tau, \quad t \ge \frac{r^*}{c}, \tag{4.2}$$

using the fact that f is supported in $[0,T]$ and $G_r^N(x,\cdot)$ in $[0, \frac{r^*}{c}]$. We deduce that

$$\begin{cases} |e^N(x,t)| \le \|f\|_{L^\infty(0,t)} \cdot \|G_r^N(x,\cdot)\|_{L^1(\frac{r^*}{c},t)} & \text{if } \dfrac{r^*}{c} \le t \le \dfrac{r^*}{c} + T, \\ |e^N(x,t)| \le \|f\|_{L^\infty(0,T)} \cdot \|G_r^N(x,\cdot)\|_{L^1(t-T,t)} & \text{if } t > \dfrac{r^*}{c} + T. \end{cases} \tag{4.3}$$

We thus have to estimate the L^1-norm of the functions $t \mapsto G_r^N(\cdot,t)$. Using the fact that

$$|P_N(\Phi(x,t))| \le 1$$

(estimate which is quasi-optimal for a range of values of t) we get

$$|G_r^N(x,t)| \le \frac{1}{2\pi\sqrt{t^2 - \frac{r^{*2}}{c^2}}} \left(\frac{ct - (x_2+h)}{ct + (x_2+h)}\right)^N.$$

We remark that, for $t \ge \frac{r^*}{c}$, the function $ct \mapsto (\frac{ct-(x_2+h)}{ct+(x_2+h)})$ is increasing. Therefore,

$$\begin{aligned} \|G_r^N(x,\cdot)\|_{L^1(\frac{r^*}{c},t)} &\le \frac{1}{2\pi} \left(\frac{ct - (x_2+h)}{ct + (x_2+h)}\right)^N \int_{\frac{r^*}{c}}^{t} \frac{d\tau}{\sqrt{\tau^2 - \frac{r^{*2}}{c^2}}} \\ &= \frac{1}{2\pi} \left(\frac{ct - (x_2+h)}{ct + (x_2+h)}\right)^N \operatorname{Log}\left(\frac{ct + \sqrt{c^2t^2 - r^{*2}}}{r^*}\right), \end{aligned} \tag{4.4}$$

while, as soon as $t > \frac{r^*}{c} + T$,

$$\begin{aligned} &\|G_r^N(x,\cdot)\|_{L^1(t-T,t)} \\ &\quad \le \frac{1}{2\pi} \left(\frac{ct - (x_2+h)}{ct + (x_2+h)}\right)^N \int_{t-T}^{t} \frac{d\tau}{\sqrt{\tau^2 - \frac{r^{*2}}{c^2}}} \\ &\quad = \frac{1}{2\pi} \left(\frac{ct - (x_2+h)}{ct + (x_2+h)}\right)^N \operatorname{Log}\left(\frac{ct + \sqrt{c^2t^2 - r^{*2}}}{c(t-T) + \sqrt{c^2(t-T)^2 - r^{*2}}}\right). \end{aligned} \tag{4.5}$$

It is easy to deduce (2.13) and (2.14) from (4.3), (4.4), and (4.5).

We now move to the proof of the uniform estimates (2.15) and (2.16). Let us introduce the two disjoint sets

$$\begin{aligned} &\Omega_1(t) = \{x \in R^2_+ / c(t-T) < r^*(x) \le ct\}, \\ &\Omega_2(t) = \{x \in R^2_+ / r^*(x) \le c(t-T)\}. \end{aligned} \tag{4.6}$$

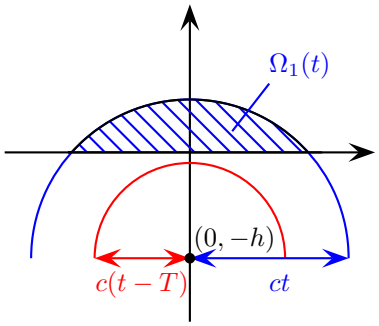


FIG. 6. *The set* $\Omega_1(t)$, $\frac{h}{c} \le t < \frac{h}{c} + T$.

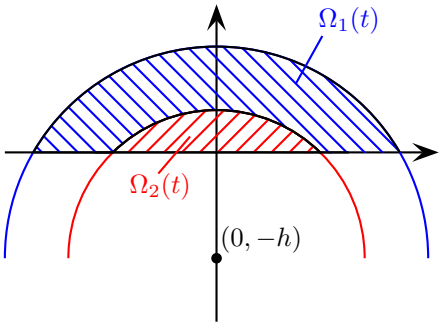


FIG. 7. $\Omega_1(t)$ *and* $\Omega_2(t)$, $t \ge \frac{h}{c} + T$.

These two sets are represented in Figures 6 and 7 for two values of t. Note that $\Omega_1(t)$ is not empty as soon as $t > \frac{h}{c}$, while $\Omega_2(t)$ is not empty as soon as $t > \frac{h}{c} + T$. According to (4.3), in order to derive an L^∞ estimate of $e^N(\cdot,t)$, we need an upper bound for the quantity

$$\sup_{x\in\Omega_1(t)} \|G_r^N(x,\cdot)\|_{L^1(\frac{r^*}{c},t)} \quad \text{when } t > \frac{h}{c}$$

and for the quantity

$$\sup_{x\in\Omega_2(t)} \|G_r^N(x,\cdot)\|_{L^1(t-T,t)} \quad \text{when } t > \frac{h}{c} + T.$$

We remark that for each $x \in \Omega(t) = \Omega_1(t) \cup \Omega_2(t)$, $h \le x_2 + h \le ct$. Therefore, noticing that the function

$$x \mapsto \frac{ct - x}{ct + x}, \quad x \in [0, ct],$$

is decreasing, we get

$$\sup_{x\in\Omega_1(t)} \left(\frac{ct - (x_2+h)}{ct + (x_2+h)}\right) = \sup_{x\in\Omega_2(t)} \left(\frac{ct - (x_2+h)}{ct + (x_2+h)}\right) = \frac{ct-h}{ct+h}. \tag{4.7}$$

On the other hand, using the fact that the two functions

$$\begin{cases} r \mapsto \dfrac{ct}{r} + \sqrt{\dfrac{c^2t^2}{r^2} - 1}, & r \in [0, ct], \\ r \mapsto \dfrac{ct + \sqrt{c^2t^2 - r^2}}{c(t-T) + \sqrt{c^2(t-T)^2 - r^2}}, & r \in [0, c(t-T)] \ (t > T), \end{cases}$$

are, respectively, decreasing and increasing, we deduce that

$$\left|\begin{array}{l} \displaystyle\sup_{x\in\Omega_1(t)} \operatorname{Log}\left|\frac{ct}{r^*}+\sqrt{\frac{c^2t^2}{r^{*2}}-1}\right| = \operatorname{Log}\left(\frac{t+\sqrt{t^2-(h/c)^2}}{(h/c)}\right) \quad \text{if } \frac{h}{c}<t<\frac{h}{c}+T, \\ \\ \displaystyle\sup_{x\in\Omega_1(t)} \operatorname{Log}\left|\frac{ct}{r^*}+\sqrt{\frac{c^2t^2}{r^{*2}}-1}\right| = \operatorname{Log}\left(\frac{t+\sqrt{t^2-(t-T)^2}}{t-T}\right) \quad \text{if } t>\frac{h}{c}+T, \end{array}\right.$$

$$\sup_{x\in\Omega_2(t)} \operatorname{Log}\left|\frac{ct+\sqrt{c^2t^2-r^{*2}}}{c(t-T)+\sqrt{c^2(t-T)^2-r^{*2}}}\right| = \operatorname{Log}\left(\frac{t+\sqrt{t^2-(t-T)^2}}{t-T}\right)$$

$$\text{if } t>\frac{h}{c}+T$$

These last three equalities, together with (4.3), (4.4), and (4.5), permit us to show the inequalities

$$\begin{aligned} &\sup_{x\in\Omega_1(t)} \left\|G_r^N(x,\cdot)\right\|_{L^1(\frac{r^*}{c},t)} \\ &\qquad \le \frac{1}{2\pi}\left(\frac{ct-h}{ct+h}\right)^N \operatorname{Log}\left(\frac{ct}{h}+\sqrt{\frac{c^2t^2}{h^2}-1}\right) \quad \left(t>\frac{h}{c}\right), \end{aligned} \tag{4.8}$$

$$\begin{aligned} &\sup_{x\in\Omega_2(t)} \left\|G_r^N(x,\cdot)\right\|_{L^1(t-T,t)} \\ &\qquad \le \frac{1}{2\pi}\left(\frac{ct-h}{ct+h}\right)^N \operatorname{Log}\left(\frac{t+\sqrt{t^2-(t-T)^2}}{t-T}\right) \quad \left(t>\frac{h}{c}+T\right). \end{aligned} \tag{4.9}$$

It is then easy to conclude the proof of Theorem 2.4 from (4.8), (4.9), and (4.3).

Remark 4.1. In formula (2.10), the function G_r^N naturally appears as the product of three terms. In the proof above, in order to estimate G_r^N we have estimated independently, for the sake of simplicity, each of these factors. In particular, our final estimates are not necessarily sharp.

5. Illustration and analysis of the results.

5.1. Analysis of the 2D fundamental solutions.

Relative error analysis. One of the difficulties in representing numerically the reflected field G_r^N given by (2.10) is the presence of the singularity of the circle $r^*(x) = ct$. To overcome this difficulty, the idea is to compare this reflected field to what it would be with the Dirichlet boundary condition (which corresponds to $N = 0$). That is why we introduce the relative error field defined as

$$\gamma_r^N(x,t) = \frac{G_r^N(x,t)}{G_r^0(x,t)} = P_N\left(\Phi(x,t)\right)\left[\frac{ct-(x_2+h)}{ct+(x_2+h)}\right]^N, \quad x\in\Omega(t). \tag{5.1}$$

(Note that $G_r^0(\cdot,t)$ does not vanish inside the disk $r^*(x) < ct$.) In the following experiments we choose $h = 1$ and $c = 1$. On Figures 8 to 12 we represent, at three different times—namely, $t = 3, 5$, and 7 from top to bottom—the level lines of $x \mapsto \gamma_r^N(x,t)$. Each figure corresponds to one value of N ($N = 1, 2, 5, 10, 20$).

We clearly observe the following:

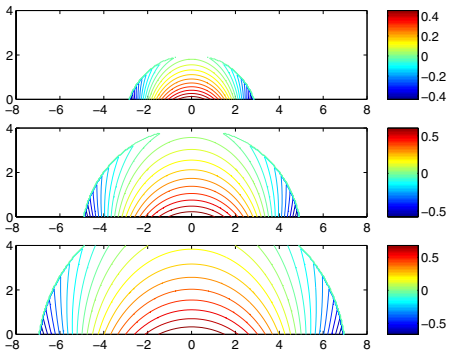


FIG. 8. $x \mapsto \gamma_r^1(x,t)$, $t = 3, 5,$ *and* 7.

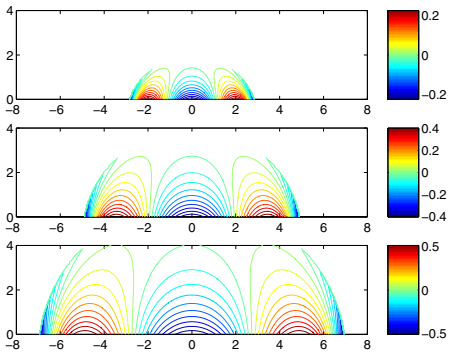


FIG. 9. $x \mapsto \gamma_r^2(x,t)$, $t = 3, 5,$ *and* 7.

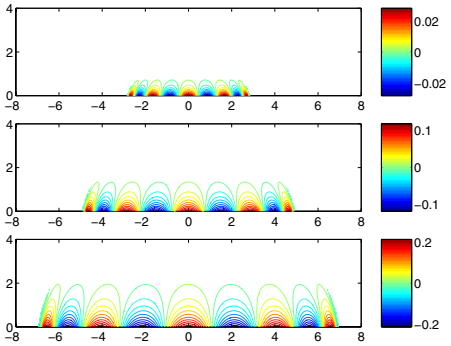


FIG. 10. $x \mapsto \gamma_r^5(x,t)$, $t = 3, 5,$ *and* 7.

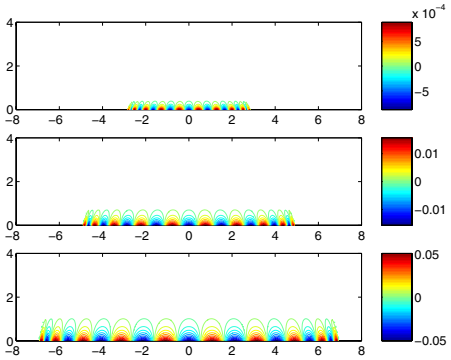


FIG. 11. $x \mapsto \gamma_r^{10}(x,t)$, $t = 5, 7,$ *and* 9.

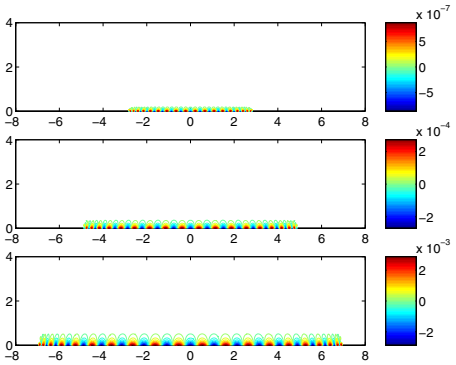


FIG. 12. $x \mapsto \gamma_r^{20}(x,t)$, $t = 3, 5,$ *and* 7.

(i) The amplitude of the error strongly decays with N (take care of the scales). For instance, at $t = 3$, the error level is 0.4 for $N = 1, 0.2$ for $N = 2, 0.02$ for $N = 5, 7.10^{-4}$ for $N = 10$, and 6.10^{-7} for $N = 20$.

(ii) As expected, the amplitude of the error also increases in time. By a homogeneity argument, it is obvious that looking at different t's for fixed h is equivalent to

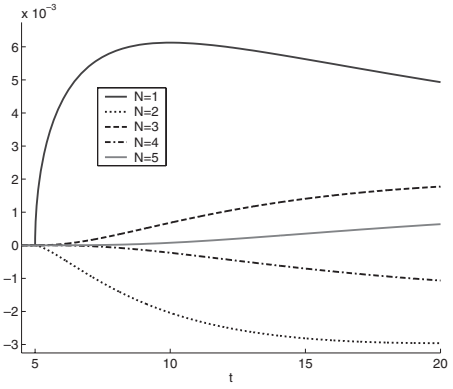


FIG. 13. $t \mapsto G_r^N(x,t)$, $r^* = 5$, $\theta = 0$.

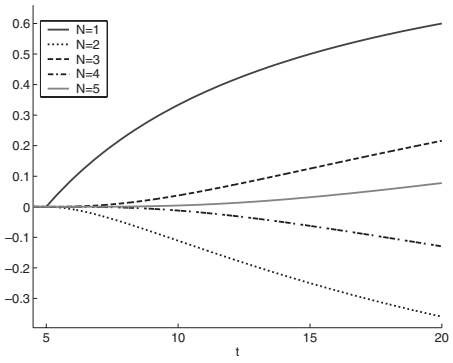


FIG. 14. $t \mapsto \gamma_r^N(x,t)$, $r^* = 5$, $\theta = 0$.

looking at different h's for fixed t. Therefore, our results also illustrate the influence of h on the reflected field.

(iii) When N increases, the relative error concentrates more and more on the neighborhood of the absorbing boundary. Moreover, its dependence with respect to the space variable is more and more complicated (this is the effect of the Chebyshev polynomials).

Study of the error as a function of time. Here we wish to study the evolution of the reflected field at a given point x as a function of time. All the points we observe are located on the circle $r^*(x) = 5$ so that the reflected field arrives at these points at time $t = 5$.

(i) *The case of the point* $\theta(x) = 0$. Contrary to what the plane wave analysis might suggest, the reflected field is not identically 0 for $\theta(x) = 0$, i.e., on the x_2 axis. However, the function $t \mapsto G_r^N(x,t)$ is not discontinuous (except for $N = 0$!) at time $t = \tau = \tau(x) = (x_2 + h)/c$, as shown by the formula

$$G_r^N(x,t) = \frac{(-1)^{N+1}}{2\pi} \frac{[ct - (x_2+h)]^{N-\frac{1}{2}}}{[ct + (x_2+h)]^{N+\frac{1}{2}}} \quad \text{for } t > \tau. \tag{5.2}$$

It becomes even less and less singular as N increases. Moreover, one sees that the function $t \mapsto G_r^N(x,t)$ is increasing from $t = \tau$ to $t = 2N\tau$, then decreasing for $t > 2N\tau$, and tends to 0 when $t \to +\infty$ as $1/2\pi ct$. The maximum of $t \mapsto G_r^N(x,t)$ is given by

$$\sup_{t \geq \tau} G_r^N(x,t) = \frac{1}{2N+1} \left(\frac{2N-1}{2N+1} \right)^{N-\frac{1}{2}} \sim \frac{1}{2Ne} \quad (N \to +\infty). \tag{5.3}$$

These properties are illustrated in Figures 13 and 14, where we represent (in Figure 13) the variations of $t \mapsto G_r^N(x,t)$, $t \in [0,20]$ for $N = 1$ to 5. Looking at the relative error, the formula

$$\gamma_r^N(x,t) = (-1)^N \frac{[ct - (x_2+h)]^{N-\frac{1}{2}}}{[ct + (x_2+h)]^{N+\frac{1}{2}}} \quad \text{for } t > \tau \tag{5.4}$$

shows that the function $t \mapsto \gamma_r^N(x,t)$, $t > \tau$, increases from 0 to 1. This is confirmed in Figure 14.

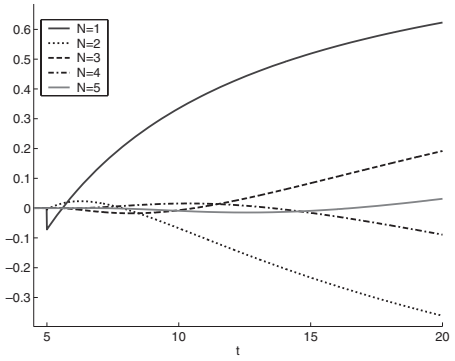


FIG. 15. $t \mapsto \gamma_r^N(x,t)$, $r^* = 5$, $\theta = \frac{\pi}{6}$.

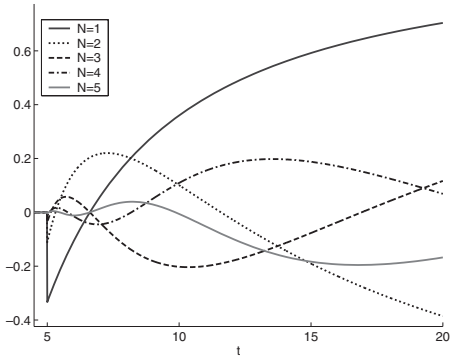


FIG. 16. $t \mapsto \gamma_r^N(x,t)$, $r^* = 5$, $\theta = \frac{\pi}{3}$.

(ii) *The case of points* $\theta(x) \neq 0$. In this case, the function $t \mapsto \gamma_r^N(x,t)$ is no longer continuous for $t = \tau$:

$$\lim_{t \to \tau} \gamma_r^N(x,t) = \mathbf{R}_N(\theta(x)) = (-1)^N \left(\frac{1 - \cos\theta}{1 + \cos\theta} \right)^N . \tag{5.5}$$

In Figures 15 and 16 we represent the variations of $t \mapsto \gamma_r^N(x,t)$ for $\theta = \pi/6$ and $\theta = \pi/3$. In each figure, one makes N vary from 1 to 5. Clearly, the higher N is, the more the function oscillates. Finally, for large times, one easily computes that

$$\lim_{t \to +\infty} \gamma_r^N(x,t) = (-1)^N \tag{5.6}$$

independently of the value of N.

Study of the error as a function of the distance to the image source. We consider here the spatial variation of the reflected field along a ray, namely, the part of a half-line starting from the image source point S^* included in the half-space $\mathbb{R}^2_+$. For a given direction $\theta \in]-\pi/2, \pi/2[$, this ray is also defined as

$$D_\theta = \left\{ x \in \mathbb{R}^2_+ / \theta(x) = \theta \right\} = \{ (r^* \sin\theta, r^* \cos\theta),\ r^* \geq h/\cos\theta \}.$$

In the following figures we represent the variations of the reflected field G_r^N, for fixed θ as a function of $r^* \geq h/\cos\theta$, for different values of t and N.

For $\theta = 0$, $r^* \leq h$. In Figures 17 to 19, we represent the variations of G_r^N along D_0 for three values of t, $t = 3, 5, 8$. Each figure corresponds to one value of N, and therefore the scale varies a lot from one picture to another. Once again, one observes that the reflected field is smoother and smoother as N increases.

For $\theta = \pi/6$, $r^* \leq 2h/\sqrt{3}$. In Figures 20 to 22, we represent the variations of G_r^N along $D_{\pi/6}$ for $t = 3, 5, 8$. This time, the functions are singular for $r^* = ct$. However, one observes that the region where G_r^N takes very large values becomes more and more confined close to the point $r^* = ct$ as N increases.

For $\theta = \pi/3$, $r^* \leq 2h$. In Figures 23 to 25, we represent the variations of G_r^N along $D_{\pi/3}$ for $t = 3, 5, 8$. The functions are still singular for $r^* = ct$. The shape of the reflected wave is more complicated than for $\theta = \pi/6$, especially for large N.

Angular variation of the reflected wave. Our previous results have already illustrated the dependence of the reflected field with respect to $\theta(x)$. Here, let us consider the relative error $\gamma_r^N(x,t)$ along the "reflected" wave front defined as

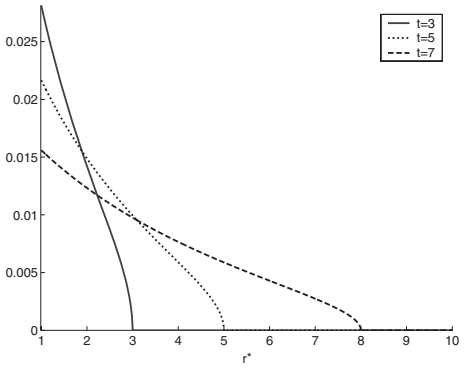


FIG. 17. $r^* \mapsto G_r^1(r^*, \theta = 0, t)$.

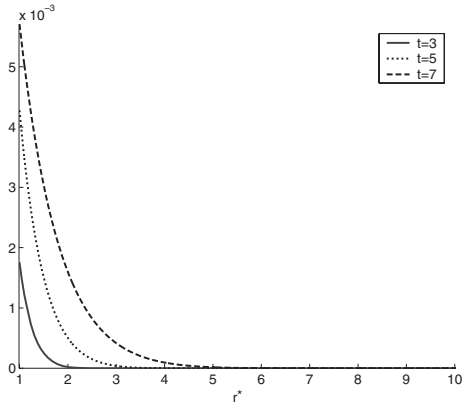


FIG. 18. $r^* \mapsto G_r^5(r^*, \theta = 0, t)$.

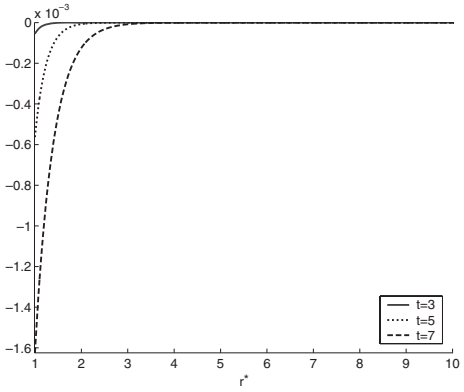


FIG. 19. $r^* \mapsto G_r^{10}(r^*, \theta = 0, t)$.

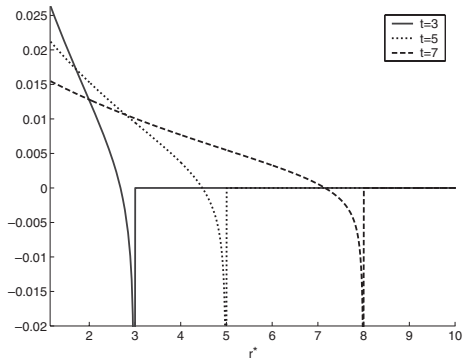


FIG. 20. $r^* \mapsto G_r^1(r^*, \theta = \pi/6, t)$.

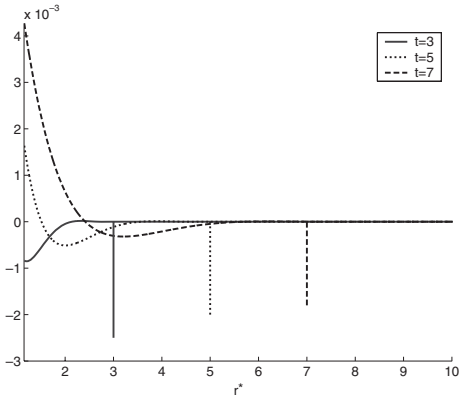


FIG. 21. $r^* \mapsto G_r^5(r^*, \theta = \pi/6, t)$.

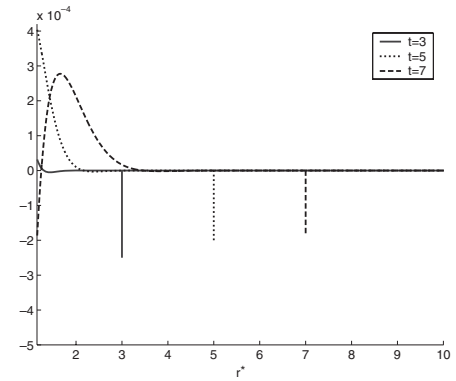


FIG. 22. $r^* \mapsto G_r^{10}(r^*, \theta = \pi/6, t)$.

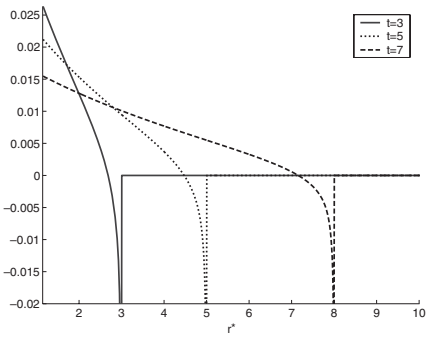


FIG. 23. $r^* \mapsto G_r^1(r^*, \theta = \pi/3, t)$.

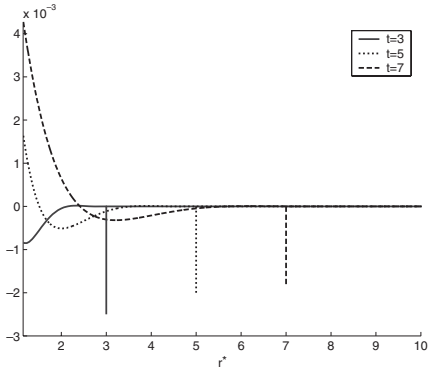


FIG. 24. $r^* \mapsto G_r^5(r^*, \theta = \pi/3, t)$.

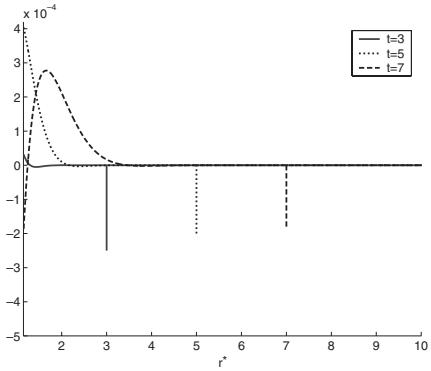


FIG. 25. $r^* \mapsto G_r^{10}(r^*, \theta = \pi/3, t)$.

$$WF_r(t) = \partial\Omega(t) = \{x \in \mathbb{R}^2_+ / r^*(x) = ct\} \quad \left(\neq \emptyset \text{ for } t > \frac{h}{c}\right).$$

Let $M_\theta(t) = (ct\sin\theta, ct\cos\theta) \in WF_r(t)$ (note that $M_\theta(t)$ describes $WF_r(t)$ when θ varies from $-\arccos\frac{h}{ct}$ to $+\arccos\frac{h}{ct}$); one easily deduces from (2.10) that

$$\lim_{x \to M_\theta(t),\ x \in \Omega(t)} \gamma_r^N(x,t) = \mathbf{R}_N(\theta) = (-1)^N \left(\frac{1-\cos\theta}{1-\cos\theta}\right)^N.$$

In other words, the curve representing, as a function of the direction θ, the variations of the relative error $\gamma_r^N(x,t)$ along the "reflected" wave front $WF_r(t)$ is nothing but the portion of the curve of Figure 1 that corresponds to $-\arccos\frac{h}{ct} \le \theta \le \arccos\frac{h}{ct}$.

5.2. The case of a source term.

Comparison with numerical experiments. We have implemented a MATLAB code for the computation of the convolution integral (4.2). To validate our "exact" solution(!), we have compared our results with those obtained with a finite difference code written by F. Collino. In our experiment, the source function is a truncated first derivative of a Gaussian:

$$f(t) = \frac{d}{dt}\{e^{-2\pi f_0 (t-t_0)^2}\} H(2t_0 - t), \quad f_0 = 10,\ t_0 = 1/f_0. \tag{5.7}$$

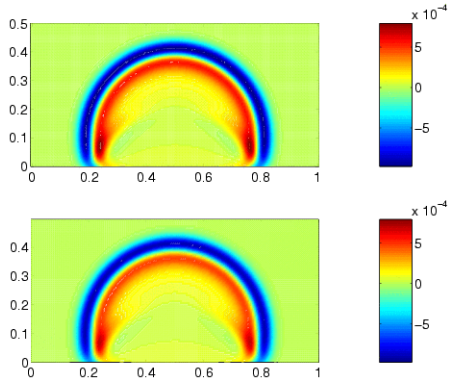


FIG. 26. *Total field.* $N = 1$.

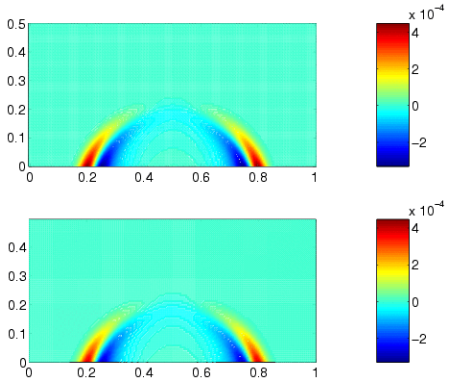


FIG. 27. *Reflected field.* $N = 1$.

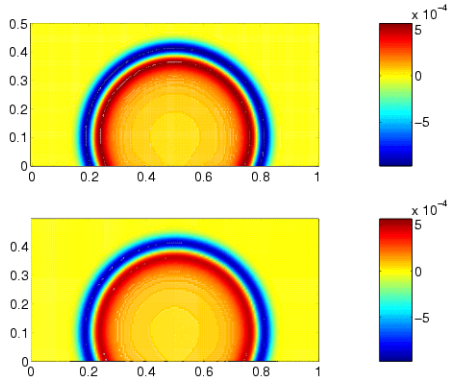


FIG. 28. *Total field.* $N = 5$.

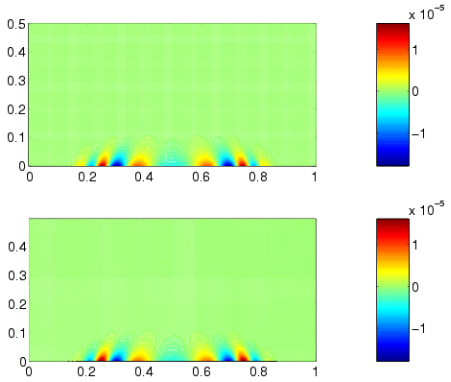


FIG. 29. *Reflected field.* $N = 5$.

In Figures 26 to 29 we have compared the "analytical" solution (top picture in each panel) to the numerical one (bottom picture in each panel) for two values of N: $N = 1$ and 5. In each picture we represent the level lines of the solution at time $t = 0.4$. The left pictures represent the total field while the right pictures represent the reflected field (the error). In each case, for the representation, the reflected field has been amplified by a factor which depends on N: 1.3 for $N = 1$ and 25 for $N = 5$. In each case, the results reveal a very good agreement between the two solutions. In Figures 30 to 33 we have compared both solutions at point $(0.9, 0.1)$ as functions of time. The solid curves represent the "analytical" solution and the dashed curves the numerical one for two values of N: $N = 1$ and 5. As before the left pictures represent the total field while the right pictures represent the reflected field.

L^∞ error estimates. In Figures 34 and 35 we have compared the L_∞-norm of the reflected field (the solid curves) to the uniform estimates (2.15) and (2.16) given by Theorem 2.4 (the dashed curves) for $N = 1, 2, 5, 10$. The source is a step function in time: $f(t) = 1$ if $0 \le t \le 2$ and $f(t) = 0$ otherwise. Our estimate appears to be very sharp for $N = 1$ and becomes less accurate (although quite acceptable) as N increases. Since we used the L^∞-norm of the source function to establish our error estimates, one could imagine that this estimate is not very sharp for more complicated source

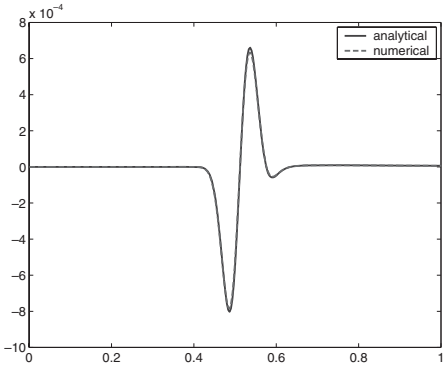


FIG. 30. *Total field.* $N = 1$.

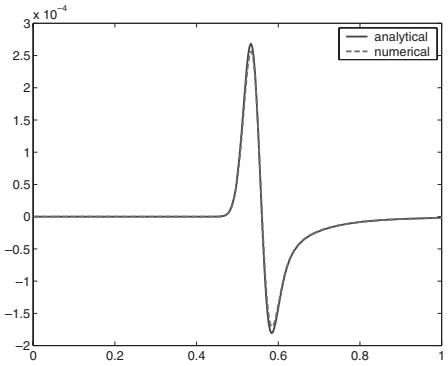


FIG. 31. *Reflected field.* $N = 1$.

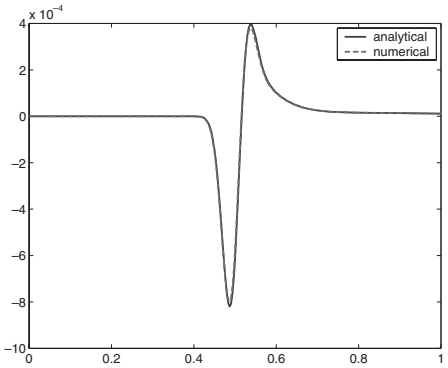


FIG. 32. *Total field.* $N = 5$.

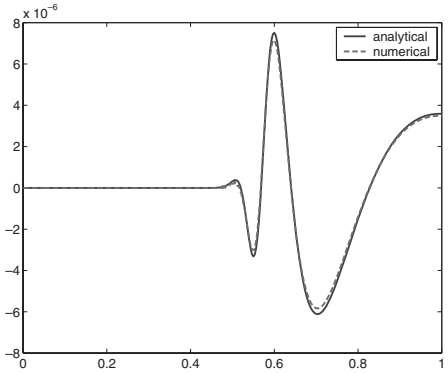


FIG. 33. *Reflected field.* $N = 5$.

functions. To check this, we have repeated the previous experiment when the source is still given by (5.7) with $f_0 = 1$. Figures 36 and 37 illustrate these experiments for $N = 1$ and 5. The estimate is obviously less accurate than in the case of the step source function but still gives a reasonable bound.

6. Conclusion and perspectives. The use of the Cagniard–De Hoop method has enabled us to obtain a quasi-analytical representation of the field reflected by Engquist–Majda higher order ABCs. This permits us to obtain new theoretical estimates for the time-dependent problem.

Of course, the method can be applied to other boundary conditions (we give in the appendix the example of Higdon's boundary conditions). It would also be interesting to treat other equations such as Maxwell's equations or elastodynamics equations. One also might think that the Cagniard–De Hoop method could be a new tool for analyzing the stability of boundary conditions.

In a forthcoming work, we wish to treat the case of the PMLs for ABCs. This should give some insights about the quantitative comparison between ABCs and PMLs.

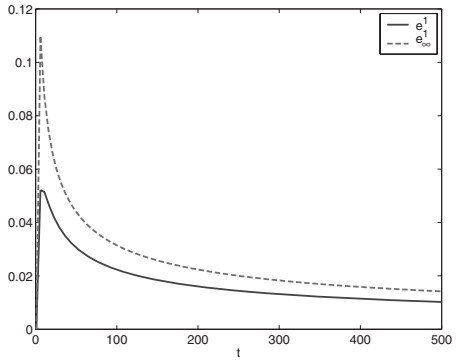


FIG. 34. *Error estimates.* $N = 1$.

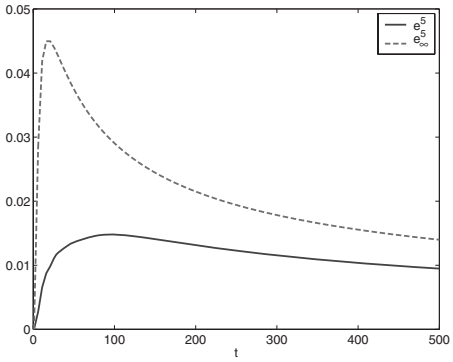


FIG. 35. *Error estimates.* $N = 5$.

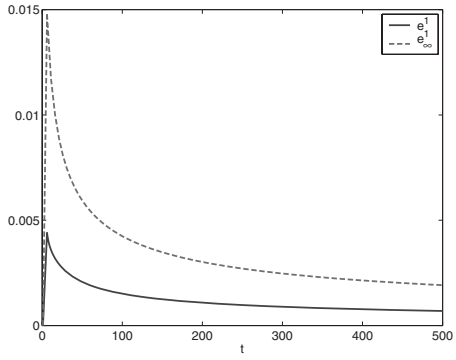


FIG. 36. *Error estimates.* $N = 1$.

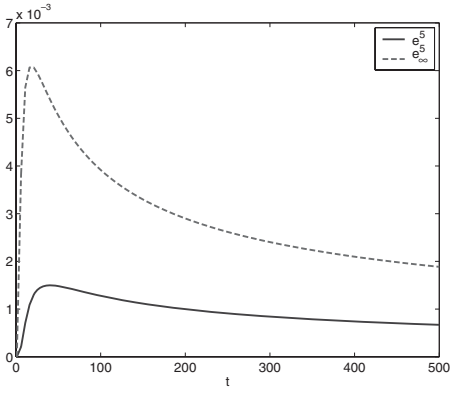


FIG. 37. *Error estimates.* $N = 5$.

Appendix. Extension to Higdon's boundary conditions. In 1986 Higdon [23], [21] proposed another approximation of the condition (1.2) of the form

$$B_{Hig}^N u = \prod_{j=1}^{N} \left(\cos\alpha_j \frac{\partial}{\partial t} - c\frac{\partial}{\partial x} \right) u = 0. \tag{A.1}$$

These conditions are a generalization of the condition (1.5) (obtained with $\alpha_j = 0$ for all j) and have the property to be exactly satisfied by any linear combination of plane waves whose angle of incidence is α_j.

Using the same method as in section 3, it can be shown that the solution of the problem

$$\begin{cases} \text{Find } u : \mathbb{R}^2_+ \times \mathbb{R} \mapsto \mathbb{R} & \text{such that} \\ \dfrac{1}{c^2}\dfrac{\partial^2 u}{\partial t^2} - \Delta u = \delta(x - x_S) \times \delta(t) & \text{in } \mathbb{R}^2_+, \\ \mathcal{B}_{Hig}^N u = 0 & \text{on } \Gamma, \\ u(x,t) = 0 & \text{for } t < 0 \end{cases} \tag{A.2}$$

is given by

$$u(x,t) = G^N_{Hig}(x,t) = G_i(x,t) + G^N_{Hig,r}(x,t), \tag{A.3}$$

where

$$\left| \begin{array}{l} G_i(x,t) = \dfrac{1}{2\pi\sqrt{t^2 - \frac{r(x)^2}{c^2}}} H(ct - r(x)), \\ G^N_{Hig,r}(x,t) = \dfrac{1}{2\pi\sqrt{t^2 - \frac{r^*}{c^2}}} \left[\prod_{j=1}^{n} \rho_j(x,t)\right] \cos\left[\sum_{j=1}^{n} \psi_i(x,t)\right] H(ct - r^*(x)), \end{array} \right. \tag{A.4}$$

where $\rho_j(x,t)$ and $\psi_i(x,t)$ are given by

$$\rho_j(x,t) = \sqrt{\frac{(ct-a_j)^2 - b_j^2}{(ct+a_j)^2 - b_j^2}} \tag{A.5}$$

and

$$\psi_j(x,t) = \arccos\left[\frac{r^*(x,t)^2 - c^2t^2 + r^*(x,t)^2\cos^2\alpha_j - r^*(x,t)^2\cos^2\theta}{\sqrt{((ct-a_j)^2 - b_j^2)((ct+a_j)^2 - b_j^2)}}\right], \tag{A.6}$$

$$\psi_j(x,t) = \arccos\left[\frac{r^*(x,t)^2 - c^2t^2 + r^*(x,t)^2\cos^2\alpha_j - (x_2+h)^2}{\sqrt{((ct-a_j)^2 - b_j^2)((ct+a_j)^2 - b_j^2)}}\right] \tag{A.7}$$

with

$$a_j = r^*(x)\cos\alpha_j\cos\theta = \cos\alpha_j(x_2+h)$$

and

$$b_j = r^*(x)\sin\theta\sin\alpha_j = x_1\sin\alpha_j.$$

Thanks to (A.5), one can see that the function $x \mapsto G^N_{Hig,r}(x,t)$ is singular on the circle $r^*(x) = ct$ except in the directions α_j.

REFERENCES

[1] B. Alpert, L. Greengard, and T. Hagstrom, *Rapid evaluation of nonreflecting boundary kernels for time-domain wave propagation*, SIAM J. Numer. Anal., 37 (2000), pp. 1138–1164.

[2] B. Alpert, L. Greengard, and T. Hagström, *Nonreflecting boundary conditions for the time-dependent wave equation*, J. Comput. Phys., 180 (2002), pp. 270–296.

[3] G. Baker, *Essentials of Padé Approximants*, Academic Press, New York, 1975.

[4] H. Barucq, *A new family of first-order boundary conditions for the Maxwell system: Derivation, well-posedness and long-time behavior*, J. Math. Pures Appl., 9 (2003), pp. 67–88.

[5] J. Bérenger, *Three-dimensional perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 127 (1996), pp. 363–379.

[6] J. P. Bérenger, *A perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 114 (1994), pp. 185–200.

[7] L. Cagniard, *Réflexion et Réfraction des Ondes Sismiques Progressives*, Gauthier-Villars, Paris, 1939.

[8] T. H. Duong and P. Joly, *On the stability analysis of boundary conditions for the wave equation by energy methods. Part* I: *The homogeneous case*, Math. Comp., 62 (1994), pp. 539–563.

[9] B. Engquist and L. Halpern, *Far field boundary conditions for computation over long time*, Appl. Numer. Math., 4 (1988), pp. 21–45.

[10] B. Engquist and L. Halpern, *Long-time behaviour of absorbing boundary conditions*, Math. Methods Appl. Sci., 13 (1990), pp. 189–203.

[11] B. Engquist and A. Majda, *Absorbing boundary conditions for the numerical simulation of waves*, Math. Comp., 31 (1977), pp. 629–651.

[12] B. Engquist and A. Majda, *Radiation boundary conditions for acoustic and elastic wave calculations*, Comm. Pure Appl. Math., 32 (1979), pp. 314–358.

[13] D. Givoli, *Exact and high order non-reflecting computational boundaries*, in Mathematical and Numerical Aspects of Wave Propagation, G. Cohen, E. Heikkola, P. Joly, and P. Neittanmaki, eds., Springer, Berlin, 2003, pp. 26–31.

[14] M. J. Grote and J. B. Keller, *Exact nonreflecting boundary conditions for the time dependent wave equation*, SIAM J. Appl. Math., 55 (1995), pp. 280–297.

[15] M. J. Grote and J. B. Keller, *Nonreflecting boundary conditions for time dependent scattering*, J. Comput. Phys., 127 (1996), pp. 52–81.

[16] T. Hagström, *On the convergence of local approximations to pseudodifferentials operators*, in Proceedings of the 3rd International Conference on Mathematical and Numerical Aspects of Wave Propagation, E. Bécache, G. Cohen, P. Joly, and J. Roberts, eds., SIAM, Philadelphia, 1995, pp. 474–482.

[17] T. Hagström, *On high-order radiation boundary conditions*, in Computational Wave Propagation (Minneapolis, MN, 1994/1995), B. Engquist and G. A. Kriegsmann, eds., IMA Vol. Math. Appl. 86, Springer, New York, 1997, pp. 1–21.

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

[19] T. Hagström, *New results on absorbing layers and radiation boundary conditions*, in Topics in Computational Wave Propagation: Direct and Inverse Problems, M. Ainsworth, P. Davies, D. Duncan, P. Martin, and B. Rynne, eds., Lect. Notes Comput. Sci. Eng. 31, Springer, Berlin, 2003, pp. 1–42.

[20] L. Halpern and J. Rauch, *Error analysis for absorbing boundary conditions*, Numer. Math., 51 (1987), pp. 459–467.

[21] R. L. Higdon, *Absorbing boundary conditions for difference approximations to the multidimensional wave equation*, Math. Comp., 47 (1986), pp. 437–459.

[22] R. L. Higdon, *Initial-boundary value problems for linear hyperbolic systems*, SIAM Rev., 28 (1986), pp. 177–217.

[23] R. L. Higdon, *Numerical absorbing boundary conditions for the wave equation*, Math. Comp., 4 (1987), pp. 65–90.

[24] J. V. D. Hijden, *Propagation of Transient Elastic Waves in Stratified Anisotropic Media*, North-Holland Ser. Appl. Math. Mech. 32, North-Holland, Amsterdam, 1987.

[25] A. T. D. Hoop, *The surface line source problem*, Appl. Sci. Res. B, 8 (1959), pp. 349–356.

[26] H. O. Kreiss, *Initial boundary value problems for hyperbolic systems*, Comm. Pure Appl. Math., 23 (1970), pp. 277–298.

[27] L. N. Trefethen and L. Halpern, *Well-posedness of one-way wave equations and absorbing boundary conditions*, Math. Comp., 47 (1986), pp. 421–435.

SIAM J. APPL. MATH.
Vol. 65, No. 5, pp. 1576–1596

STRUCTURAL BIFURCATION OF 2-D NONDIVERGENT FLOWS WITH DIRICHLET BOUNDARY CONDITIONS: APPLICATIONS TO BOUNDARY-LAYER SEPARATION*

MICHAEL GHIL†, TIAN MA‡, AND SHOUHONG WANG§

Abstract. This article addresses transitions in the topological structure of a family of divergence-free vector fields $u(\cdot, t)$ with Dirichlet boundary conditions. We show that structural bifurcation—i.e., change in topological-equivalence class—occurs at t_0 if $u(\cdot, t_0)$ has an isolated degenerate ∂-singular point $\bar{x} \in \partial M$ such that $\partial^2 u(\bar{x}, t_0)/\partial n \partial t \neq 0$. The main results are proved by classifying orbit structures of u near such a point $\bar{x} \in \partial M$ of $u(\cdot, t_0)$. The condition of $\bar{x}$ being a ∂-singular point is equivalent to the one originally postulated by Prandtl for boundary-layer separation. Our analysis and classification do contribute, in fact, to a rigorous characterization of boundary-layer separation in 2-D incompressible fluid flows.

Key words. structural bifurcation, boundary layer separation, 2-D incompressible flows

AMS subject classifications. 37G, 76M, 34D30, 35Q30, 37E

DOI. 10.1137/S0036139903438818

1. Introduction.

This article is part of a research program on the use of topological ideas to study the spatio-temporal structure of 2-D incompressible fluid flows in physical space, along with its stability and bifurcations. This program consists of research in two areas: (a) the study of the topological structure of divergence-free vector fields, and its evolution in time or with respect to an arbitrary parameter; and (b) the study of the structure and evolution of velocity fields for 2-D incompressible fluid flows governed by a class of equations that comprises the Navier–Stokes equations, the Euler equations, and the quasi-geostrophic equations of rotating flows.

Mathematically speaking, there are two general methods for describing a fluid flow: the Euler representation and the Lagrange representation; see [1, 2, 3, 6, 9, 13, 17]. In the Euler representation, the motion of a fluid is described by a set of partial differential equations (PDEs)—such as the Euler equations or the Navier–Stokes equations, supplemented with proper boundary conditions—that govern the velocity field at every point in the (2-D or 3-D) flow domain. The Lagrange representation of a fluid flow, on the other hand, amounts to studying the trajectories of fluid particles

*Received by the editors December 15, 2003; accepted for publication (in revised form) October 24, 2004; published electronically May 12, 2005.
http://www.siam.org/journals/siap/65-5/43881.html

†Département Terre-Atmosphère-Océan, Ecole Normale Supérieure, F-75231 Paris, Cedex 05, France, and Institute of Geophysics and Planetary Physics, University of California, Los Angeles, CA 90095. This author was supported in part by the National Science Foundation under grants from the Atmospheric Sciences and the Oceanic Sciences Divisions.

‡Department of Mathematics, Sichuan University, Chengdu, Sichuan 610064, People's Republic of China, and Department of Mathematics, Indiana University, Bloomington, IN 47405. The research of this author was supported in part by the Office of Naval Research under a grant from the Mathematical, Computer, and Information Sciences Division, by the National Science Foundation under a grant from the Division of Mathematical Sciences (DMS), and by the National Science Foundation of China.

§Department of Mathematics, Indiana University, Bloomington, IN 47405 (showang@indiana.edu, http://www.indiana.edu/~fluid). The research of this author was supported in part by the Office of Naval Research under a grant from the Mathematical, Computer, and Information Sciences Division, by the National Science Foundation under a grant from the Division of Mathematical Sciences (DMS), and by the National Science Foundation of China.

as a function of initial position in the flow domain, subject to the ordinary differential equations (ODEs) that govern the change in position given the velocity. Of course the velocities of the particles satisfy the PDEs we just mentioned.

Our approach is to classify the topological structure of the *instantaneous* velocity field, treating the time variable as a parameter, and the changes in this structure with respect to time. The aforementioned two areas of our program draw inspiration from and are relevant to both the Eulerian and the Lagrangian approaches to fluid flows.

The study in area (a) is kinematic in nature, and the results and methods developed can naturally be applied to other problems of mathematical physics that involve divergence-free vector fields. These include, for instance, problems in electromagnetism in which the magnetic field is necessarily divergence-free. The main topics in this area include structural classification, structural stability, and structural bifurcation, as well as their applications to fluid dynamics in general and to geophysical fluid dynamics in particular. The study in area (b) involves specific connections between the solutions of the evolution equations—whether Navier–Stokes, Euler, or quasi-geostrophic—and the flow structure in the physical space.

The main objective of this paper is to contribute to a rigorous characterization of boundary-layer separation in 2-D incompressible fluid flows. This is a long-standing problem in fluid mechanics that goes back to the pioneering work of Prandtl [15] in 1904. Classical boundary-layer theory is presented in [2, 9, 16]. The Prandtl equation represents an approximation of the Navier–Stokes equations inside the boundary layer in the absence of separation; this equation is rigorously analyzed in a recent textbook [14] and several articles [4, 5, 18].

Basically, the boundary layer is a narrow region of sharp velocity gradients between a no-slip wall, where the velocity has to vanish, and the interior of the fluid. This layer of high shear can detach from the boundary, generating vortices and leading to more complicated turbulent behavior near the wall as well as in the interior of the flow domain [9]. It is important, therefore, to characterize, if at all possible, the conditions for separation. Experimentally one observes that the normal derivative of the velocity field vanishes at or near separation points. Chorin and Marsden [2] note that there is no known theorem that can be applied to determine the separation reliably. This article, along with [8, 10], is an attempt to derive a rigorous characterization of streamline detachment from the boundary for 2-D divergence-free vector fields. These results are applied in the companion papers [7, 12] to the actual problem of boundary-layer separation in solutions of the 2-D incompressible Navier–Stokes equations.

In the present article, we address structural transitions for a family of divergence-free vector fields $u(\cdot, t)$ with Dirichlet boundary conditions. We show that structural bifurcation—i.e., change in their topological-equivalence class—occurs at t_0 if $u(\cdot, t_0)$ has an isolated degenerate ∂-singular point $\bar{x} \in \partial M$, such that $\partial^2 u(\bar{x}, t_0)/\partial n \partial t \neq 0$. The condition that $\bar{x} \in \partial M$ is a ∂-singular point is related, as we shall see in section 3, to the Prandtl condition [15] for boundary-layer separation.

Our main results are based on a complete classification of orbit structures near an isolated degenerate ∂-singular point. These results extend over several papers and rely on a delicate analysis of the flow structure near the boundary for both free-slip and Dirichlet boundary conditions. The first step was to classify the flow structure and its transitions near the boundary for flows subject only to boundary conditions of zero normal flow, often called free-slip conditions in fluid dynamics [8]. Second, Ma and Wang [10] analyzed the case of Dirichlet boundary conditions for a 2-D divergence-free vector field; in fluid mechanics the Dirichlet condition on the velocity is often called

the no-slip condition.

Technically speaking, homogeneous Dirichlet boundary conditions for $u(\cdot, t_0)$ imply that all points on ∂M are singular points in the usual sense. Hence, to analyze and to classify the structure of u near the boundary, including the separation point, we need to use the concept of a ∂-singular point, introduced in [10], which corresponds to singular points of the normal derivative of u in the usual sense. Finally, in the present paper we make the connection between the structure of the original velocity fields and the structure of the normal derivative of the velocity field.

The paper is organized as follows. In section 2, we summarize our previous results, including a structural stability theorem, necessary conditions on structural bifurcation, and a singularity classification theory for 2-D divergence-free vector fields. Section 3 states the structural-bifurcation theorems near a flat boundary for such fields, which are proved in section 4. Section 5 addresses structural bifurcations near a curved boundary, and section 6 applies the theory to streamline detachment from the boundary. It is the results of section 6 that are applied in the companion papers [7, 12] to boundary-layer separation in the 2-D Navier–Stokes equations for incompressible flows.

2. Preliminaries. Let $M \subset \mathbb{R}^2$ be a closed and bounded domain with C^{r+1} $(r \geq 2)$ boundary ∂M, and let TM be the tangent bundle of M. Let $C_n^r(TM)$ be the space of all C^r vector fields on M that satisfy a boundary condition of no normal flow (or no penetration), let $D^r(TM)$ be the subspace of $C_n^r(TM)$ that is divergence-free, and let $B_0^r(TM)$ be the subspace of $D^r(TM)$ that satisfies a homogeneous Dirichlet boundary condition:

$$\begin{aligned} C_n^r(TM) &= \{u \in C^r(TM) \mid u_n|_{\partial M} = 0\}, \\ D^r(TM) &= \{u \in C^r(TM) \mid u_n|_{\partial M} = 0,\ \mathrm{div} u = 0\}, \\ B_0^r(TM) &= \{u \in D^r(TM) \mid u|_{\partial M} = 0\}. \end{aligned}$$

Here $u_n = u \cdot n$ and $u_\tau = u \cdot \tau$, while n and τ are the unit normal and tangent vector on ∂M, respectively. It is easy to see that

$$B_0^r(TM) \subset D^r(TM) \subset C_n^r(TM) \subset C^r(TM).$$

We start with some basic concepts. Let $X = D^r(TM)$ or $B_0^r(TM)$ in the following four definitions.

DEFINITION 2.1. *Two vector fields $u, v \in X$ are called topologically equivalent in X if there exists a homeomorphism of $\varphi : M \to M$, which takes the orbits of u to the orbits of v and preserves their orientation.*

DEFINITION 2.2. *Let $u \in C^1([0,T], X)$ be a one-parameter family of vector fields in X. The vector field $u_0 = u(\cdot, t_0)$ $(0 < t_0 < T)$ is called a bifurcation point of u at time t_0 if, for any $t^- < t_0$ and $t_0 < t^+$ with t^- and t^+ sufficiently close to t_0, the vector field $u(\cdot, t^-)$ is not topologically equivalent to $u(\cdot, t^+)$. In this case, we say that $u(x,t)$ has a bifurcation at t_0 in its global structure.*

DEFINITION 2.3. *Let $u \in C^1([0,T], X)$. We say that $u(x,t)$ has a bifurcation in its local structure in a neighborhood $U \subset M$ of x_0 at t_0 $(0 < t_0 < T)$ if, for any $t^- < t_0$ and $t_0 < t^+$ with t^- and t^+ sufficiently close to t_0, the vector fields $u(\cdot, t^-)$ and $u(\cdot, t^+)$ are not topologically equivalent locally in $U \subset M$.*

We remark here that bifurcation in the vector field's local structure does not imply bifurcation in its global structure. In fact, one can easily construct examples

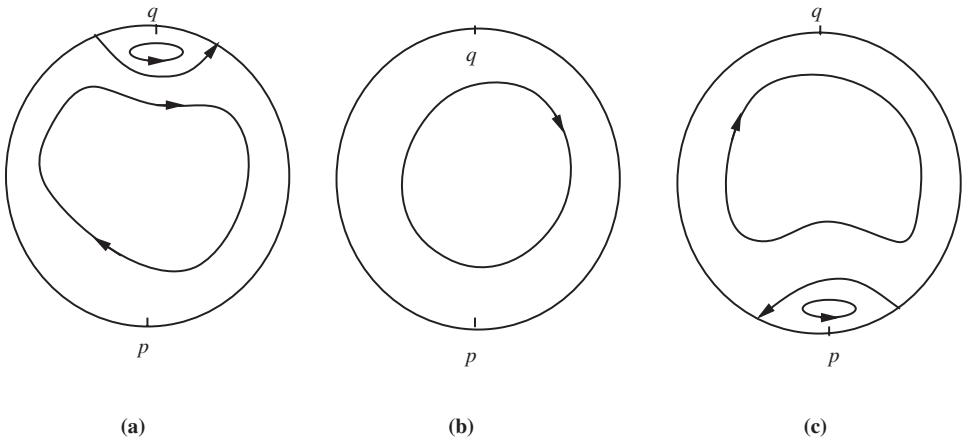


FIG. 2.1. *Flow structure* (a) *for* $t = t^- < t_0$, (b) *for* $t = t_0$, *and* (c) *for* $t = t^+ > t_0$. *Bifurcations in local structure occur at* $t = t_0$ *near both* p *and* q, *but there is no bifurcation in global structure when going from* (a) *to* (c).

showing that flow structure changes in some local area $U \subset M$, but not on the whole manifold M; see, for instance, the flow transitions shown in Figure 2.1.

DEFINITION 2.4. *A vector field* $v \in X$ *is called structurally stable in* X *if there exists a neighborhood* $\mathcal{O} \subset X$ *of* v *such that for any* $u \in \mathcal{O}$, u *and* v *are topologically equivalent.*

A point $p \in M$ is called a singular point of $u \in C_n^r(TM)$ if $u(p) = 0$; a singular point p of u is called nondegenerate if the Jacobian matrix $Du(p)$ is invertible; u is called regular if all singular points are nondegenerate.

By definition, in the case where homogeneous Dirichlet boundary conditions are satisfied, all points on the boundary are singular points. In order to classify the structure of the divergence-free vector fields near the boundary in this case, and infer the possible bifurcations in this structure, we need to distinguish between different types of singular points on the boundary. The concepts of ∂-regular and ∂-singular points introduced in [10] are crucial in order to study the topological structure of divergence-free vector fields with homogeneous Dirichlet boundary conditions.

Let $p \in \partial M$, and let U be a neighborhood of p. Then on $\partial M \cap U$ there exist unit tangent and normal vector fields τ and n. For U sufficiently small, we can extend these two vector fields to U so that the orbits of n in U are tangent to λn with n restricted to $\partial M \cap U$; here $0 \le \lambda \le 1$. Note that when U is sufficiently small, for any two points $x, y \in \partial M \cap U$, λn_x and λn_y do not intersect within U. The extension of τ in U is taken to be orthogonal to n.

DEFINITION 2.5. *Let* $u \in B_0^r(TM)$ $(r \ge 2)$.

(i) *A point* $p \in \partial M$ *is called a* ∂*-regular point of* u *if* $\partial u_\tau(p)/\partial n \ne 0$; *otherwise,* $p \in \partial M$ *is called a* ∂*-singular point of* u.

(ii) *A* ∂*-singular point* $p \in \partial M$ *of* u *is called nondegenerate if*

$$\det \begin{pmatrix} \dfrac{\partial^2 u_\tau(p)}{\partial \tau \partial n} & \dfrac{\partial^2 u_\tau(p)}{\partial n^2} \\ \dfrac{\partial^2 u_n(p)}{\partial \tau \partial n} & \dfrac{\partial^2 u_n(p)}{\partial n^2} \end{pmatrix} \ne 0.$$

A nondegenerate ∂-singular point of u is also called a ∂-saddle point of u.

DEFINITION 2.6. *$u \in B_0^r(TM)$ $(r \geq 2)$ is called D-regular if*

(i) *u is regular in $\overset{\circ}{M}$, the interior of M, and*

(ii) *all ∂-singular points of u on ∂M are nondegenerate.*

For a D-regular divergence-free vector field v on M, an interior nondegenerate singular point of v can be either a center or a saddle, and saddles of v must be connected to saddles. An interior saddle $p \in \overset{\circ}{M}$ is called *self-connected* if p is connected only to itself, i.e., p occurs in a graph whose topological form is that of the number 8.

The following structural stability theorem in the presence of homogeneous Dirichlet boundary conditions was proved in [10].

THEOREM 2.7 (Ma and Wang [10]). *Let $u \in B_0^r(TM)$ $(r \geq 2)$. Then u is structurally stable in $B_0^r(TM)$ if and only if*

1. *u is D-regular;*
2. *all interior saddle points of u are self-connected; and*
3. *each ∂-saddle point of u on ∂M is connected to a ∂-saddle point on the same connected component of ∂M.*

Moreover, the set of all structurally stable vector fields is open and dense in $B_0^r(TM)$.

Based on Theorem 2.7, the following theorem gives some necessary conditions for structural bifurcation.

THEOREM 2.8. *Let $u \in C^1([0,T], B_0^r(TM))$ $(r \geq 2)$.*

1. *If $u(x,t)$ has a bifurcation in its local structure in an arbitrarily small neighborhood $U \subset M$ of x_0 at t_0 $(0 < t_0 < T)$, then $x_0 \in \overset{\circ}{M}$ (respectively, $x_0 \in \partial M$) must be a degenerate singular point (respectively, degenerate ∂-singular point) of $u(x,t)$ at t_0.*
2. *If $u(x,t)$ has a bifurcation in its global structure at t_0 $(0 < t_0 < T)$, then $u(x,t_0)$ does not satisfy at least one of the conditions* (1)–(3) *in Theorem* 2.7.

To proceed, we need to recall the definition of indices of singular points of a vector field. Let $p \in M$ be an isolated singular point of $v \in C_n^r(TM)$; then

$$\operatorname{ind}(v,p) = \deg(v,p),$$

where $\deg(v,p)$ is the Brouwer degree of v at p.

Let $p \in \partial M$ be an isolated singular point of v, and let $\widetilde{M} \subset \mathbb{R}^2$ be an extension of M, i.e., $M \subset \widetilde{M}$ such that $p \in \widetilde{M}$ is an interior point of $\widetilde{M}$. In a neighborhood of p in $\widetilde{M}$, v can be extended by reflection to $\tilde{v}$ such that p is an interior singular point of $\tilde{v}$, thanks to the no normal flow condition, i.e., $v \cdot n|_{\partial M} = 0$. Then we define the index of v at $p \in \partial M$ by

$$\operatorname{ind}(v,p) = \frac{1}{2}\operatorname{ind}(\tilde{v},p).$$

Let $p \in M$ be an isolated singular point of $v \in C_n^r(TM)$. An orbit γ of v is said to be a stable orbit (respectively, an unstable orbit) connected to p if the limit set $\omega(x) = p$ (respectively, $\alpha(x) = p$) for any $x \in \gamma$.

We now introduce a singularity classification theorem for incompressible vector fields, which will be useful in our discussion of structural bifurcation.

THEOREM 2.9 (Ghil, Ma, and Wang [8]). *Let $p \in M$ be an isolated singular point of $v \in D^r(TM)$, $r \geq 1$. Then p is connected only to a finite number of orbits, and the stable and unstable orbits connected to p alternate when tracing a closed curve around p. Furthermore*

1. *when $p \in \mathring{M}$, p has $2n$ $(n \geq 0)$ orbits, n of which are stable, and the other n unstable, while the index of p is*
$$\mathrm{ind}(v, p) = 1 - n;$$
2. *when $p \in \partial M$, p has $n + 2$ $(n \geq 2)$ orbits, two of which are on the boundary ∂M, and the index of p is*
$$\mathrm{ind}(v, p) = -\frac{n}{2}.$$

No confusion should arise between the integer n, used for counting orbits, and the notation n for the normal direction to the boundary ∂M.

3. Structural bifurcations near a flat boundary. In this section, we assume that the boundary ∂M contains a flat part $\Gamma \subset \partial M$, and consider structural bifurcation near a ∂-singular point $\bar{x} \in \Gamma$. For simplicity, we take a coordinate system (x_1, x_2) with $\bar{x}$ at the origin and with Γ given by
$$\Gamma = \{(x_1, 0) \,\big|\, |x_1| \leq \delta\}$$
for some $\delta > 0$. Obviously, the tangent and normal vectors on Γ are the unit vectors in the x_1- and x_2-directions, respectively.

Let $u \in C^1([0, T], B_0^r(TM))$ $(r \geq 2)$ be a one-parameter family of divergence-free vector fields subject to homogeneous Dirichlet boundary conditions. In a neighborhood $U \subset M$ of $\bar{x} \in \Gamma$, $u(x, t)$ can be expressed near $x = 0$ by
$$u(x, t) = x_2 v(x, t). \tag{3.1}$$
It is easy to see that u and v have topologically equivalent streamlines in an interior neighborhood of $x = 0$. To proceed, we consider the Taylor expansions of both $u(x, t)$ and $v(x, t)$ at t_0 $(0 < t_0 < T)$:
$$\begin{cases} u(x,t) = u^0(x) + (t - t_0)u^1(x) + o(|t - t_0|^2), \\ u^0(x) = u(x, t_0), \\ u^1(x) = \dfrac{\partial u(x, t_0)}{\partial t}, \end{cases} \tag{3.2}$$
$$\begin{cases} v(x,t) = v^0(x) + (t - t_0)v^1(x) + o(|t - t_0|)^2, \\ v^0(x) = v(x, t_0), \\ v^1(x) = \dfrac{\partial v(x, t_0)}{\partial t}. \end{cases} \tag{3.3}$$

Let $u^i = (u_1^i, u_2^i)$, $v^i = (v_1^i, v_2^i)$, $i = 0, 1$. We start with the following conditions for structural bifurcation.

Assumption (H). Let $\bar{x} = 0 \in \Gamma$ be an isolated degenerate ∂-singular point of $u^0(x)$, $u^0 \in C^{k+1}$ near $\bar{x} \in \Gamma$ for some $k \geq 2$. Assume that
$$\frac{\partial u^0(0)}{\partial n} = 0, \tag{3.4}$$
$$\mathrm{ind}(v^0, 0) \neq -\frac{1}{2}, \tag{3.5}$$
$$\frac{\partial u^1(0)}{\partial n} \neq 0, \tag{3.6}$$
$$\frac{\partial^{k+1} u_1^0(0)}{\partial^k \tau \partial n} \neq 0. \tag{3.7}$$

Some remarks are now in order.

Remark 3.1. Condition (3.4) says that $\bar{x} = 0 \in \Gamma$ is a ∂-singular point of $u^0(x)$. In a 2-D incompressible flow governed by either the Euler or the Navier–Stokes equations, this condition is equivalent to the leading-order vorticity vanishing at $\bar{x}$. The latter is the so-called Prandtl condition, which Prandtl suggested might identify boundary-layer separation points in incompressible flows [15].

Remark 3.2. Condition (3.5) amounts to saying that there exist a number $n \neq 1$ of interior orbits of v^0 connected to $\bar{x} \in \Gamma$. Since $u^0 = x_2 v^0$, the number of interior orbits of u^0 connected to $\bar{x} \in \Gamma$ is exactly $n \neq 1$ as well. This shows that $\bar{x} \in \Gamma$ is a degenerate ∂-singular point of $u^0(x)$, which is necessary for structural bifurcation, according to our structural stability and bifurcation theorems; see Theorems 2.7 and 2.8 here, or [8, 10].

Remark 3.3. Condition (3.6) states that the first-order term u^1 of the Taylor expansion for the normal derivative of u is different from zero. This is just the simplest necessary condition of such a type; if it does not hold, we need to work on a higher-order Taylor expansion, and the corresponding results proved in this article will be true as well. In fluid-mechanics applications, condition (3.6) is equivalent to the vorticity associated with u^1 not vanishing at the boundary. In addition, it is easy to see that (3.6) is equivalent to

$$\frac{\partial u_1^1(0)}{\partial x_2} = \frac{\partial u_1^1(0)}{\partial n} \neq 0,$$

which shows that the acceleration of the fluid in the tangential direction near $\bar{x}$ is nonzero.

Remark 3.4. Condition (3.7) is a technical condition and amounts to saying that the tangential component u_1^0 of the leading-order term has a nontrivial Taylor expansion. Furthermore, let k be the smallest integer satisfying condition (3.7). It is easy to show that $k \geq 2$. In fact, $u^0(x)$ has the Taylor expansion at $x = 0$,

$$u^0(x) = \begin{cases} cx_2 + 2ax_1x_2 + bx_2^2 + x_2h_1(x), \\ \qquad\qquad -ax_2^2 + x_2h_2(x), \end{cases} \tag{3.8}$$

with $h_i(x) = o(|x|)$ $(i = 1, 2)$. Since $\bar{x} \in \Gamma$ is a degenerate ∂-singular point of $u^0(x)$, it follows that $c = 0$, $a = 0$, which implies that $k \geq 2$.

The structural bifurcation of $u(x, t)$ near a degenerate ∂-singular point on a flat boundary segment is described by the following theorems.

THEOREM 3.5. *Let $u \in C^1([0, T], B_0^r(TM))$ $(r \geq 2)$ satisfy Assumption* (H). *Then there exist a neighborhood*

$$\Gamma_0 = \{(x_1, 0) \,|\, |x_1| \leq \delta_0\} \subset \Gamma$$

of $\bar{x} = 0$ and an $\varepsilon_0 > 0$ such that all ∂-singular points of $u(\cdot, t_0 \pm \varepsilon)$ are nondegenerate for any $0 < \varepsilon \leq \varepsilon_0$. Moreover,

1. *if the index* $\mathrm{ind}(v^0, 0)$ *is an integer, then one of $u(x, t_0 \pm \varepsilon)$ has exactly two ∂-singular points on Γ_0, and the other has no ∂-singular points on Γ_0; and*
2. *if the index* $\mathrm{ind}(v^0, 0)$ *is not an integer, then each of $u(x, t_0 \pm \varepsilon)$ has exactly one ∂-singular point on Γ_0.*

THEOREM 3.6 (structural bifurcation theorem). *Let $u \in C^1([0, T], B_0^r(TM))$ $(r \geq 2)$ satisfy Assumption* (H). *Then*

1. *the vector field u has a bifurcation in its local structure at $(\bar{x}, t_0)$; and*

2. *if $\bar{x} \in \partial M$ is a unique ∂-singular point of u with the same index as* $\mathrm{ind}(v^0, 0)$ *on ∂M, then $u(x,t)$ has a bifurcation in its global structure at $t = t_0$.*

The proofs of these two theorems are based on analyzing the orbits of u to provide a complete classification of the local structure of u near $(\bar{x}, t_0)$. The bifurcation results follow immediately from the classification.

4. Proofs of the main theorems.

4.1. Proof of Theorem 3.5. The proof of the theorem is a direct consequence of the following three propositions.

PROPOSITION 4.1. *There exist a neighborhood $\Gamma_0 \subset \Gamma$ of $\bar{x} = 0$ and an $\varepsilon_0 > 0$ such that all ∂-singular points of $u(\cdot, t_0 \pm \varepsilon)$ are nondegenerate for any $0 < \varepsilon \le \varepsilon_0$.*

Proof. Let $x^* \in \Gamma$, $x^* = (x_1^*)$, and $0 < |x_1^*| \le \delta_0$ with $0 < \delta_0$ sufficiently small, to be chosen later. Then

$$\frac{\partial u}{\partial n}(x^*, t_0 \pm \varepsilon) = \frac{\partial u}{\partial x_2}(x^*, t_0 \pm \varepsilon) = 0. \tag{4.1}$$

By the Taylor expansion (3.2) and condition (3.6) it suffices to consider only the first-order approximation of (3.2). Hence,

$$\frac{\partial u_i^0(x_1^*, 0)}{\partial x_2} \pm \varepsilon \frac{\partial u_i^1(x_1^*, 0)}{\partial x_2} = 0 \qquad (i = 1, 2). \tag{4.2}$$

We need to show that

$$\det \begin{pmatrix} \dfrac{\partial^2 u_1^0}{\partial x_1 \partial x_2} \pm \varepsilon \dfrac{\partial^2 u_1^1}{\partial x_1 \partial x_2} & \dfrac{\partial^2 u_1^0}{\partial x_2^2} \pm \varepsilon \dfrac{\partial^2 u_1^1}{\partial x_2^2} \\ \dfrac{\partial^2 u_2^0}{\partial x_1 \partial x_2} \pm \varepsilon \dfrac{\partial^2 u_2^1}{\partial x_1 \partial x_2} & \dfrac{\partial^2 u_2^0}{\partial x_2^2} \pm \varepsilon \dfrac{\partial^2 u_2^1}{\partial x_2^2} \end{pmatrix}_{x=(x_1^*,0)} \neq 0. \tag{4.3}$$

For $u \in C^1([0,T], B_0^r(TM))$, $u(x_1, 0, t) = 0$ for all $(x_1, 0) \in \Gamma$ and $0 \le t \le T$. Thus we obtain

$$\frac{\partial u_1}{\partial x_1}(x_1, 0, t) = 0 \qquad \forall\, |x_1| \le \bar{\delta}.$$

Thanks to the fact that u is divergence-free, we have

$$\frac{\partial u_2}{\partial x_2}(x_1, 0, t) = -\frac{\partial u_1}{\partial x_1}(x, 0, t) = 0 \qquad \forall\, |x_1| \le \bar{\delta}.$$

Consequently

$$\frac{\partial u_2^0(x_1, 0)}{\partial x_2} \pm \varepsilon \frac{\partial u_2^1(x_1, 0)}{\partial x_2} = 0 \qquad \forall\, |x_1| \le \bar{\delta},$$

which yields

$$\frac{\partial^2 u_2^0(x_1^*, 0)}{\partial x_1 \partial x_2} \pm \varepsilon \frac{\partial^2 u_2^1(x_1^*, 0)}{\partial x_1 \partial x_2} = 0.$$

To verify (4.3), it suffices to prove that

$$\frac{\partial^2 u_1^0(x_1^*, 0)}{\partial x_1 \partial x_2} \pm \varepsilon \frac{\partial^2 u_1^1(x_1^*, 0)}{\partial x_1 \partial x_2} \neq 0,$$

since the sum of the diagonal terms in (4.3) is zero thanks to the fact that $\partial u/\partial x_2$ is divergence-free.

From conditions (3.7) and (3.6) we get

$$\begin{cases} \dfrac{\partial u_1^0(x_1,0)}{\partial x_2} = \alpha x_1^k + o(|x_1|^k), \quad \alpha \neq 0, \\ \dfrac{\partial u_1^1(x_1,0)}{\partial x_2} = \beta + o(|x_1|), \quad \beta \neq 0. \end{cases} \tag{4.4}$$

Therefore it follows from (4.2) that

$$\varepsilon = \pm \frac{\alpha}{\beta}\, x_1^{*^k} + o(|x_1^*|^k). \tag{4.5}$$

Thus from (4.4) and (4.5) we obtain that

$$\frac{\partial^2 u_1^0(x_1^*,0)}{\partial x_1 \partial x_2} \pm \varepsilon\, \frac{\partial^2 u_1^1(x_1^*,0)}{\partial x_1 \partial x_2} = \alpha k x_1^{*^{k-1}} + o\left(|x_1^*|^{k-1}\right),$$

which is different from zero for $0 < |x_1^*| \leq \delta_0$, provided that δ_0 is sufficiently small. Hence (4.3) follows, and the proof of the proposition is complete. □

PROPOSITION 4.2. *If* $\operatorname{ind}(v^0, 0) =$ *integer, then one of* $u(x, t_0 \pm \varepsilon)$ *has no* ∂*-singular points on* Γ_0*, and the other one has exactly two* ∂*-singular points on* Γ_0 *with one* ∂*-singular point on each side of* $\bar{x} = 0$*.*

Proof. From (3.1) it is easy to see that the zero points of $v(x,t)$ on Γ are equivalent to the ∂-singular points of $u(x,t)$. Hence we only have to prove the assertion for the vector field $v(x, t_0 \pm \varepsilon)$.

According to (3.8) we have, for the component v_2 that is normal to Γ,

$$v_2(x_1, 0, t) = 0 \quad \forall\, x_1 \in \Gamma,\ t \geq 0.$$

By (3.6) and (3.7) we infer from (3.1) that

$$\begin{cases} v_1^0(x_1,0) = \alpha x_1^k + o(|x_1|^k), & \alpha \neq 0, \\ v_1^1(x_1,0) = \beta + g(x_1), & \beta \neq 0,\ g(0) = 0. \end{cases} \tag{4.6}$$

On the other hand, if $\operatorname{ind}(v^0, 0)$ is an integer, then by Remark 3.2 the number n of interior orbits of v^0 connected to $\bar{x} = 0$ is even. Hence, one of the two boundary orbits of v^0 connected to $\bar{x} = 0$ is stable, and the other one is unstable. It follows that the exponent k in (4.6) is even, i.e., $k = 2m$ $(m \geq 1)$.

Consider the two equations

$$\begin{aligned} 0 &= v_1(x_1, 0, t_0 + \varepsilon) = v_1^0(x_1, 0) + \varepsilon v_1^1(x_1, 0) + o(|\varepsilon|) \\ &= \alpha x_1^{2m} + \varepsilon\beta + \varepsilon g(x_1) + o(|\varepsilon|, |x_1|^{2m}), \end{aligned} \tag{4.7}$$

$$\begin{aligned} 0 &= v_1(x_1, 0, t_0 - \varepsilon) = v_1^0(x_1, 0) - \varepsilon v_1^1(x_1, 0) + o(|\varepsilon|) \\ &= \alpha x_1^{2m} - \varepsilon\beta - \varepsilon g(x_1) + o(|\varepsilon|, |x_1|^{2m}), \end{aligned} \tag{4.8}$$

where $m \geq 1$. Without loss of generality, assume that $\alpha, \beta > 0$. Then there is a $\delta_0 > 0$ such that for any $\varepsilon > 0$ sufficiently small, (4.8) has only two solutions $x_1^{\pm}$ of opposite sign, $x_1^-(\varepsilon) < 0 < x_1^+(\varepsilon)$, in the interval $(-\delta_0, \delta_0)$, and (4.7) has no solutions in $(-\delta_0, \delta_0)$. The claim is verified. □

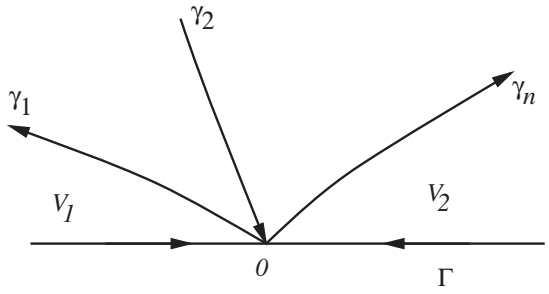


FIG. 4.1. *Case of n interior orbits connected to $\bar{x} = 0 \in \Gamma$, where $\Gamma \subset \partial M$ is flat, and $n \geq 3$ is odd.*

PROPOSITION 4.3. *If* $\operatorname{ind}(v^0, 0) =$ *noninteger, then for any* $\varepsilon > 0$ *sufficiently small each* $u(x, t_0 \pm \varepsilon)$ *has exactly one* ∂*-singular point on* Γ.

Proof. Indeed, when $\operatorname{ind}(v^0, 0) = -n/2$ is a fraction, then n is odd. Hence the exponent k in (4.6) is odd, i.e., $k = 2m + 1$ $(m \geq 1)$. Therefore each of the two equations

$$\alpha x_1^{2m+1} + \varepsilon\beta + \varepsilon g(x_1) + o\left(|\varepsilon|, |x_1|^{2m+1}\right) = 0,$$
$$\alpha x_1^{2m+1} - \varepsilon\beta - \varepsilon g(x_1) + o\left(|\varepsilon|, |x_1|^{2m+1}\right) = 0$$

has exactly one solution in $(-\delta_0, \delta_0)$. □

4.2. Proof of Theorem 3.6. As mentioned earlier, u and v have topologically equivalent streamlines in an interior neighborhood of $x = 0$; hence bifurcation in the local structure of u at $\bar{x} = 0$ is equivalent to that of v. As a result, we only have to consider the local bifurcation for the vector field v. We divide the proof into two steps.

Step 1. *The case where* $\operatorname{ind}(v^0, 0) =$ *integer.* By Theorem 3.5, the number of boundary saddle points of $v^0 + \varepsilon v^1$ in a small neighborhood $\Gamma_0 \subset \Gamma \subset \partial M$ of $\bar{x}$ is not the same as that for $v^0 - \varepsilon v^1$. Therefore, $v^0 + \varepsilon v^1$ and $v^0 - \varepsilon v^1$ are not topologically equivalent locally near $\bar{x} \in \Gamma \subset \partial M$.

Step 2. *The case where* $\operatorname{ind}(v^0, 0) =$ *fraction.* Without loss of generality, we assume that the two orbits connected to $\bar{x} = 0$ on $\Gamma_0 \subset \partial M$ are stable (i.e., $\alpha < 0$ in (4.6)), and $v_1^1(0) > 0$ (i.e., $\beta > 0$ in (4.6)). Let $x^+ = (x_1^+, 0)$ and $x^- = (x_1^-, 0) \in \Gamma_0 \subset \partial M$ be the singular points of $v^0 + \varepsilon v^1$ and $v^0 - \varepsilon v^1$, respectively. Hence, $x_1^- < 0 < x_1^+$ as in Theorem 3.5. The nondegeneracy of both x^- and x^+ implies that

$$\operatorname{ind}(v^0 \pm \varepsilon v^1, x^\pm) = -\frac{1}{2},$$

and there is exactly one orbit $\gamma^+(\varepsilon)$ of $v^0 + \varepsilon v^1$ in $\mathring{M}$ connected to x^+ (respectively, exactly only one orbit $\gamma^-(\varepsilon)$ of $v^0 - \varepsilon v^1$ in $\mathring{M}$ connected to x^-).

By (3.5) and Remark 3.2, there are n ($n \geq 3$ and odd) orbits γ_i $(1 \leq i \leq n)$ of v^0 in $\mathring{M}$ connected to $\bar{x} \in \Gamma$; see Figure 4.1. Let $V_1 \subset \mathring{M}$ (respectively, $V_2 \subset \mathring{M}$) be the domain near $\bar{x} = 0 \in \Gamma_0 \subset \partial M$ enclosed by ∂M and γ_1 (respectively, by ∂M and γ_n).

Since $v_1^1(0) > 0$, the flow of $v^0 + \varepsilon v^1$ crosses γ_n transversally and enters into V_2 (respectively, the flow of $u^0 - \varepsilon u^1$ crosses γ_1 transversally and enters into V_1).

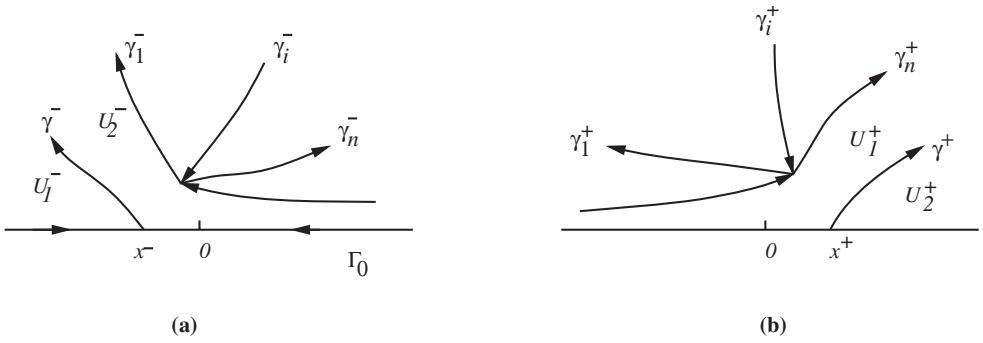


FIG. 4.2. *Sketch illustrating the proof of assertion* (1) *in Theorem* 3.6*:* (a) *orbits at* $t^- < t_0$*, and* (b) *orbits at* $t^+ > t_0$

Therefore, we have

$$\gamma^+(\varepsilon) \subset V_2, \quad \gamma^-(\varepsilon) \subset V_1. \tag{4.9}$$

Obviously,

$$\lim_{\varepsilon\to 0} \gamma^-(\varepsilon) = \gamma_1, \quad \lim_{\varepsilon\to 0} \gamma^+(\varepsilon) = \gamma_n.$$

Moreover, there are n orbits $\gamma_i^+ = \gamma_i^+(\varepsilon)$ of $v^0 + \varepsilon v^1$ in $V - V_1$ (respectively, n orbits $\gamma_i^- = \gamma_i^-(\varepsilon)$ of $v^0 - \varepsilon v^1$ in $V - V_2$), which are connected to singular points of $v^0 + \varepsilon v^1$ (respectively, $v^0 - \varepsilon v^1$) such that

$$\lim_{\varepsilon\to 0} \gamma_i^+(\varepsilon) = \gamma_i, \qquad \lim_{\varepsilon\to 0} \gamma_i^-(\varepsilon) = \gamma_i,$$

where $V \subset \mathring{M}$ is a neighborhood of $\bar{x} = 0$.

Let U_1^- (respectively, U_2^+) be the domain enclosed by ∂M and γ^- (respectively, by ∂M and γ^+), and $U_2^- = V - U_1^-$ (respectively, $U_1^+ = V - U_2^+$); see Figures 4.2(a) and 4.2(b), respectively. We infer from (4.9) that

$$\gamma_i^+ \subset U_1^+, \quad \gamma_i^- \subset U_2^- \quad (1 \le i \le n). \tag{4.10}$$

We know that, under topological equivalence, the orbits of $v^0 + \varepsilon v^1$ connected to a singular boundary point are mapped, preserving orientation, to the orbits of $v^0 - \varepsilon v^1$ connected to a singular boundary point. Since $v^0 + \varepsilon v^1$ and $v^0 - \varepsilon v^1$ are topologically equivalent locally at $\bar{x} \in \Gamma$, the restriction of $v^0 + \varepsilon v^1$ to U_1^+ would have to be topologically equivalent to $v^0 - \varepsilon v^1$ in U_1^-. This is in contradiction with (4.10). Thus, assertion (1) of Theorem 3.6 is proven.

Assertion (2) of Theorem 3.6 is a corollary of assertion (1). Indeed, if $\bar{x} \in \partial M$ is a unique singular point of u^0 which has the same index as $\mathrm{ind}(v^0, 0)$ on ∂M, then the structural bifurcation of $u(x,t)$ locally at $(\bar{x}, t_0)$ implies the structural bifurcation in its global structure.

The proof of Theorem 3.6 is thus complete.

5. Structural bifurcations near a curved boundary.

5.1. Main theorems. We now generalize in this section the main bifurcation theorems in section 3 for the flat boundary case to the curved boundary case.

Consider structural bifurcation near a ∂-singular point $\bar{x} \in \partial M$ of $u(x,t)$ on a general C^{r+1} boundary ∂M ($r \geq 2$). Let $\bar{x} \in \partial M$ and (x_1, x_2) be an orthogonal coordinate system with origin $\bar{x}$, which has its x_1-axis tangent to ∂M at $\bar{x}$, and its x_2-axis oriented in the inward normal direction.

Let $u \in C^1([0,T], B_0^r(TM))$ ($r \geq 2$) have the Taylor expansion at t_0 ($0 < t_0 < T$) as in (3.2). In particular, let

$$u^0 = (u_1^0, u_2^0) = u(x, t_0),$$
$$u^1 = (u_1^1, u_1^2) = \frac{\partial u(x, t_0)}{\partial t}.$$

In addition, let n be the number of interior orbits of $u^0(x)$ connected to $\bar{x}$. Then we can restate Assumption (H) as (H′) below, with condition (3.5) on the index there replaced by a geometrical condition $n \neq 1$, i.e., (5.2) below.

Assumption (H′). Let $\bar{x} \in \partial M$ be an isolated degenerate ∂-singular point of $u^0(x) = u(x, t_0)$, with $u^0 \in C^{k+1}$ near $\bar{x} \in \Gamma$ for some $k \geq 2$. Assume that

$$\frac{\partial u^0(\bar{x})}{\partial n} = 0, \tag{5.1}$$

$$n \neq 1, \tag{5.2}$$

$$\frac{\partial u_1^1(\bar{x})}{\partial n} \neq 0, \tag{5.3}$$

$$\frac{\partial^{k+1} u_1^0(\bar{x})}{\partial^k \tau \partial n} \neq 0. \tag{5.4}$$

We have then the following structural bifurcation theorems, as in section 3.

THEOREM 5.1. *Let $u \in C^1([0,T], B_0^r(TM))$ satisfy Assumption* (H′) *and let $r \geq 2$. Then in a neighborhood $\Gamma \subset \partial M$ of $\bar{x}$, the ∂-singular points of $u(x, t_0 \pm \varepsilon)$ are nondegenerate for any $\varepsilon > 0$ sufficiently small. Moreover,*

1. *if $n =$ even ($n \geq 0$), then one of $u(x, t_0 \pm \varepsilon)$ has two ∂-singular points on Γ, and the other one has no ∂-singular point on Γ; and*
2. *if $n =$ odd ($n \geq 3$), then each of $u(x, t_0 \pm \varepsilon)$ has only one ∂-singular point on Γ.*

THEOREM 5.2 (structural bifurcation theorem). *Let $u \in C^1([0,T], B_0^r(TM))$ be a one-parameter family of divergence-free vector fields satisfying Assumption* (H′) *and $r \geq 2$. Then the following assertions hold true:*

1. *$u(x,t)$ has a bifurcation in its local structure at $(\bar{x}, t_0)$; and*
2. *if $\bar{x} \in \partial M$ is a unique degenerate ∂-singular point of $u^0(x) = u(x, t_0)$ on ∂M, then $u(x,t)$ has a bifurcation in its global structure at t_0.*

5.2. Coordinate transformation. The main ideas that prove Theorems 5.1 and 5.2 are as follows. First, we introduce a local coordinate transformation, which preserves the divergence-free character of the vector field and maps a neighborhood $\Gamma \subset \partial M$ of $\bar{x}$ to a flat boundary. This allows us to show that Assumption (H′) is equivalent to Assumption (H) for the new transformed vector field. Then Theorems 5.1 and 5.2 follow immediately from Theorems 3.5 and 3.6.

In the coordinate system (x_1, x_2) introduced at the beginning of this section, the boundary ∂M can be expressed locally near $\bar{x} = 0 \in \partial M$ by

$$x_2 = f(x_1), \quad f(0) = 0, \quad f'(0) = 0. \tag{5.5}$$

We make the local coordinate transformation

$$\begin{cases} \tilde{x}_1 = x_1, \\ \tilde{x}_2 = x_2 - f(x_1). \end{cases} \tag{5.6}$$

Obviously, the transformation (5.6) takes a neighborhood $U \subset M$ of $\bar{x}$ to a domain $\tilde{U} \subset \mathbb{R}^2_+ = \{(\tilde{x}_1, \tilde{x}_2) \in \mathbb{R}^2 \mid \tilde{x}_2 \geq 0\}$ and maps the boundary part $U \cap \partial M$ to a neighborhood of $x = 0$ on the $\tilde{x}_1$-axis.

Let $\varphi : U \to \tilde{U}$ be the transformation (5.6), and $\varphi^* : C^r(TU) \to C^r(T\tilde{U})$ the isomorphism induced by φ. It is easy to see that

$$\varphi^* = D\varphi = \begin{pmatrix} 1 & 0 \\ -f'(\tilde{x}_1) & 1 \end{pmatrix},$$

and, for any $u \in C^r(TU)$,

$$\tilde{u} = \varphi^* \circ u = \begin{pmatrix} u_1 \\ u_2 - f'(\tilde{x}_1)u_1 \end{pmatrix}. \tag{5.7}$$

Then it is a direct calculation to derive the following lemma.

LEMMA 5.3. *If $u \in C^r(TU)$ is divergence-free, then $\tilde{u} = \varphi^* \circ u$ is also divergence-free. Moreover, as $u|_{\partial M \cap U} = 0$, then $\tilde{u}(\tilde{x}_1, 0) = 0$, and*

$$\begin{cases} \tilde{u}(\tilde{x}) = \tilde{x}_2 \tilde{v}(\tilde{x}), \\ \tilde{v}_2(\tilde{x}_1, 0) = 0. \end{cases} \tag{5.8}$$

5.3. Proof of Theorems 5.1 and 5.2. According to Theorems 3.5 and 3.6, the proof of these two theorems will be achieved in a few lemmas, as follows.

LEMMA 5.4. *Let $u \in C^1([0,T], B_0^r(TM))$ satisfy Assumption* (H′). *Then the vector field $\tilde{u} = \varphi^* \circ u = \tilde{x}_2 \tilde{v}(\tilde{x}, t)$ satisfies Assumption* (H).

Proof. Notice that φ maps $\bar{x} = 0$ to $\tilde{x} = 0$. Since u and $\tilde{u}$ are topologically equivalent locally near $x = 0$ and $\tilde{x} = 0$, (5.2) implies that the number n of interior orbits connected to $\tilde{x} = 0$ of $\tilde{v}(x, t_0)$ is different from 1. Hence

$$\operatorname{ind}(\tilde{v}(x, t_0), 0) = -\frac{n}{2} \neq -\frac{1}{2}.$$

By (5.7), $\tilde{u}_1 = u_1$, and φ^* takes the inward normal vector n at $x = 0$ to the normal vector $\tilde{n} = (0, 1)$ at $\tilde{x} = 0$. Therefore, we have

$$\frac{\partial u_1^1(0)}{\partial n} \neq 0 \Longrightarrow \frac{\partial \tilde{u}_1^1(0)}{\partial x_2} \neq 0,$$

where $u_1^1(x) = \partial u(x, t_0)/\partial t$.

By tensor analysis, we know that, under a coordinate transformation $\varphi : U \to \tilde{U}$, the directional derivative of a function $f(x)$ satisfies $\partial f/\partial r = \partial f/\partial \tilde{r}$, where r is a vector and $\tilde{r} = (D\varphi)r$.

By (5.7) we see that $u_1^0 = \tilde{u}_1^0$, and $\tilde{n} = (0,1)$ at $\tilde{x} = 0$. Hence we have

$$\frac{\partial^k}{\partial \tau^k}\frac{\partial u_1^0(0)}{\partial n} = \frac{\partial^k}{\partial \tilde{\tau}^k}\cdot\frac{\partial \tilde{u}_1^0(0)}{\partial \tilde{n}} = \frac{\partial^k}{\partial x_1^k}\frac{\partial \tilde{u}_1^0(0)}{\partial x_2},$$

and the proof of the lemma is complete. ∎

LEMMA 5.5. *A point $x \in \partial M \cap U$ is a ∂-regular point of $u \in B_0^r(TM)$ if and only if $\tilde{x} = \varphi(x) = (\tilde{x}_1, 0)$ is a ∂-regular point of $\tilde{u} = D\varphi \cdot u$.*

Proof. By Definition 2.5, it suffices to show that the two vector fields $\partial u/\partial n$ and $\partial \tilde{u}/\partial \tilde{x}_2$ have the same singular points on the boundary in the sense of homeomorphism; here

$$\frac{\partial u}{\partial n} = (N\cdot\nabla)u = n_1\frac{\partial u}{\partial x_1} + n_2\frac{\partial n}{\partial x_2}$$

and the vector field N is defined in U with the unit modulus $|N| = 1$ such that the orbits of N are the normal lines λn in U. Note that, when U is properly chosen, for any $x, y \in U \cap \partial M$, $x \neq y$, the normal lines λn_x and λn_y do not intersect within U.

Equivalently, we proceed to check the desired result for the two vector fields $\partial \tilde{u}/\partial \tilde{n}$ and $\partial \tilde{u}/\partial \tilde{x}_2$, where $\partial \tilde{u}/\partial \tilde{n}$ is the transformation of $\partial u/\partial n$ that is expressed by

$$\begin{cases} \dfrac{\partial \tilde{u}}{\partial \tilde{n}} = \left(\tilde{N}\cdot\tilde{\nabla}\right)\tilde{u} = \tilde{n}_1\dfrac{\partial \tilde{u}}{\partial \tilde{x}_1} + \tilde{n}_2\dfrac{\partial \tilde{u}}{\partial \tilde{x}_2}, \\ \tilde{N} = \begin{pmatrix}\tilde{n}_1\\ \tilde{n}_2\end{pmatrix} = D\varphi \circ N = \begin{pmatrix}1 & 0\\ -f' & 1\end{pmatrix}\begin{pmatrix}n_1\\ n_2\end{pmatrix} = \begin{pmatrix}n_1\\ n_2 - f'n_1\end{pmatrix}. \end{cases} \tag{5.9}$$

From (5.8) we see that

$$\tilde{u}(\tilde{x}_1, 0) = 0 \quad \forall\ \tilde{x}_1 \in \partial\mathbb{R}^2_+ \cap \tilde{U}. \tag{5.10}$$

By (5.9) and (5.10) we deduce that

$$\frac{\partial \tilde{u}}{\partial \tilde{n}} = (n_2 - f'(x_1)n_1)\frac{\partial \tilde{u}}{\partial \tilde{x}_2} \quad \text{on } \partial\mathbb{R}^2_+ \cap \tilde{U}. \tag{5.11}$$

From (5.5) and $n_2 \neq 0$ for $x \in \partial M$ near $\bar{x}$, the result follows and the proof of the lemma is complete. ∎

LEMMA 5.6. *A point $x \in \partial M \cap U$ is a ∂-saddle point of $u \in B_0^r(TM)$ if and only if $\tilde{x} = \varphi(x) \in \partial\mathbb{R}^2_+ \cap \tilde{U}$ is a ∂-saddle point of $\tilde{u} = D\varphi \cdot u$.*

Proof. It suffices to prove that $\partial \tilde{u}/\partial \tilde{n}$ and $\partial \tilde{u}/\partial \tilde{x}_2$ have the same nondegenerate singular points on the boundary $\Gamma = \partial\mathbb{R}^2_+ \cap \tilde{U} = \{(\tilde{x}_1, 0) \mid |\tilde{x}_1| < \delta\}$, i.e., both Jacobian determinants

$$\det\begin{pmatrix} \dfrac{\partial^2 \tilde{u}_1}{\partial \tilde{x}_1 \partial \tilde{n}} & \dfrac{\partial^2 \tilde{u}_1}{\partial \tilde{x}_2 \partial \tilde{n}} \\ \dfrac{\partial^2 \tilde{u}_2}{\partial \tilde{x}_1 \partial \tilde{n}} & \dfrac{\partial^2 \tilde{u}_2}{\partial \tilde{x}_2 \partial \tilde{n}} \end{pmatrix} \quad \text{and} \quad \det\begin{pmatrix} \dfrac{\partial^2 \tilde{u}_1}{\partial \tilde{x}_1 \partial \tilde{x}_2} & \dfrac{\partial^2 \tilde{u}_1}{\partial \tilde{x}_2^2} \\ \dfrac{\partial^2 \tilde{u}_2}{\partial \tilde{x}_1 \partial \tilde{x}_2} & \dfrac{\partial^2 \tilde{u}_2}{\partial \tilde{x}_2^2} \end{pmatrix}$$

have the same nonzero points on Γ.

From (5.8) we see that

$$\frac{\partial \tilde{u}_2(\tilde{x}_1,0)}{\partial x_2} = 0 \quad \forall\ |x_1| < \delta \tag{5.12}$$

for some $\delta > 0$. By (5.11) and (5.12) one deduces that

$$\frac{\partial^2 \tilde{u}_2(\tilde{x}_1,0)}{\partial \tilde{x}_1 \partial \tilde{n}} = 0, \quad \frac{\partial^2 \tilde{u}_2(\tilde{x}_1,0)}{\partial \tilde{x}_1 \partial \tilde{x}_2} = 0.$$

Hence we only need to prove that

$$\frac{\partial^2 \tilde{u}_1(\tilde{x}_1,0)}{\partial \tilde{x}_1 \partial \tilde{n}} \neq 0 \Longleftrightarrow \frac{\partial^2 \tilde{u}_1(\tilde{x}_1,0)}{\partial \tilde{x}_1 \partial \tilde{x}_2} \neq 0, \tag{5.13}$$

$$\frac{\partial^2 \tilde{u}_2(\tilde{x}_1,0)}{\partial \tilde{x}_2 \partial \tilde{n}} \neq 0 \Longleftrightarrow \frac{\partial^2 \tilde{u}_2(\tilde{x}_1,0)}{\partial \tilde{x}_2^2} \neq 0. \tag{5.14}$$

From (5.11) and (5.12) we immediately derive (5.14).

Thanks to (5.8), for any integer $k \geq 0$,

$$\frac{\partial^k \tilde{u}_1(\tilde{x}_1,0)}{\partial \tilde{x}_1 k} = 0. \tag{5.15}$$

By assumption, $(\tilde{x}_1, 0)$ is a singular point of $\partial\tilde{u}/\partial x_2$, i.e., a ∂-saddle point of $\tilde{u}$:

$$\frac{\partial \tilde{u}_1(\tilde{x}_1,0)}{\partial x_2} = 0. \tag{5.16}$$

From (5.15) and (5.16) it follows that

$$\begin{aligned}\frac{\partial^2 \tilde{u}_1(\tilde{x}_1,0)}{\partial \tilde{x}_1 \partial \tilde{n}} &= \frac{\partial}{\partial \tilde{x}_1}\left[\tilde{n}_1 \frac{\partial \tilde{u}_1}{\partial \tilde{x}_1} + \tilde{n}_2 \frac{\partial \tilde{u}_1}{\partial \tilde{x}_2}\right]\Bigg|_{\tilde{x}=(\tilde{x}_1,0)} \\ &= \tilde{n}_2 \frac{\partial \tilde{u}_1(\tilde{x}_1,0)}{\partial \tilde{x}_1 \partial \tilde{x}_2}\ .\end{aligned}$$

By (5.9), $\tilde{n}_2 = n_2 - f'(x_1)n_1 \neq 0$ for $x = (x_1, x_2) \in \partial M$ near $\bar{x} = 0$. Thus we derive (5.13), and the proof is complete. □

6. Applications to boundary-layer separation.

6.1. Two examples. In order to understand intuitively the connection between structural bifurcation and boundary-layer separation, we proceed by discussing two typical examples that illustrate how structural bifurcations occur in some fluid flows. For simplicity, we consider in this section only bifurcation near flat boundaries.

Let $u \in C^1([0,T], B_0^r(TM))$, $\bar{x} \in \Gamma \subset \partial M$ be an isolated ∂-singular point of $u^0(x) = u(x, t_0)$, $0 < t_0 < T$, where Γ is a flat part of ∂M. We take a coordinate system (x_1, x_2) with $\bar{x}$ at the origin and $\Gamma = \{x_1, 0) \,\big|\, |x_1| < \delta\}$. Thus $u(x,t)$ can be expressed in a neighborhood $U \subset M$ of $\bar{x}$ by

$$u(x,t) = x_2 v(x,t), \tag{6.1}$$

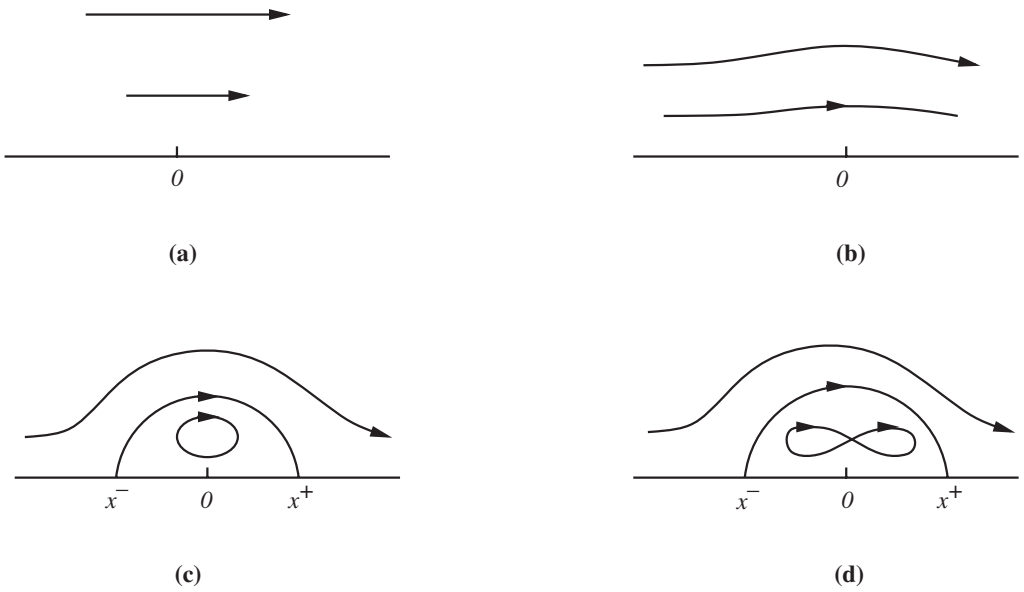


FIG. 6.1. *Boundary-layer separation and reattachment near a flat boundary.*

as in section 3. Let

$$v^0(x) = v(x, t_0) = (v_1^0(x), v_2^0(x)),$$
$$v^1(x) = \frac{\partial}{\partial t} v(x, t_0) = (v_1^1(x), v_2^1(x)).$$

Example 6.1. When the index of the vector field $v^0(x)$ at the singular point $\bar{x} = 0$ is zero, i.e., $\mathrm{ind}(v^0, 0) = 0$, structural bifurcation occurs, as shown in Figure 6.1. The figure corresponds to boundary-layer separation of an incompressible, initially parallel flow over a flat plate.

Figure 6.1(a) shows the flows structure of $u(x, t_0 - \varepsilon)$, a typical shear flow near the boundary. At the time instant $t_0 - \varepsilon$, with $\varepsilon > 0$ small, the flow exhibits no singular points near $\bar{x} = 0$. At the time instant t_0, the time at which structural bifurcation occurs, $u^0(x) = u(x, t_0)$ is given by Figure 6.1(b), which has an isolated ∂-singular point $\bar{x} = 0 \in \partial M$. At a later time, $u(x, t_0 + \varepsilon)$ is given by either Figure 6.1(c) or Figure 6.1(d). Even more complicated flow patterns are possible in the recirculation region, but in a real fluid the figure-eight streamline in Figure 6.1(d) will be affected by the viscosity and evolve into two separate, counterrotating vortices. On the boundary, there are exactly two ∂-saddle points on ∂M near $\bar{x} = 0$, denoted by x^- and x^+ in Figures 6.1(c) and 6.1(d). We shall prove hereafter that the pattern shown in Figure 6.1(c) is generic.

Example 6.2. When $\mathrm{ind}(v^0, 0) = -1$, global structural bifurcation of u may occur, due to a local transition near $\bar{x} \in \partial M$. An example of such a global bifurcation of an incompressible flow field is shown in Figure 6.2. The transition from Figure 6.2(a) through Figure 6.2(b) to Figure 6.2(c) is more idealized than in Figure 6.1 but is entirely consistent with our rigorous results, as well as physically plausible.

6.2. Boundary-layer separation. We now address the separation of streamlines and their reattachment in a 2-D divergence-free vector field, from the point of

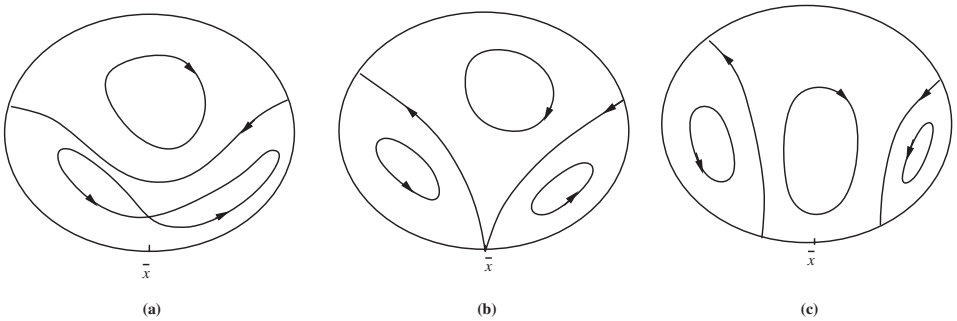


FIG. 6.2. *Global structural bifurcation of a 2-D incompressible flow in a compact domain.*

view of the rigorous results in sections 3 and 5. The connection to actual boundary-layer separation in an incompressible fluid governed by the Navier–Stokes equations is made in the companion papers [7, 12].

We start with Assumption (H) in the case where ind $(v^0, 0) = 0$, i.e., there are no interior orbits of u^0 connected to $\bar{x} = 0$. For convenience, we call it Assumption (H_0), and it reads as follows.

Assumption (H_0). Let $\bar{x} = 0 \in \Gamma$ be an isolated degenerate ∂-singular point of $u^0(x)$, $u^0 \in C^{k+1}$ near $\bar{x} \in \Gamma$ for some $k \geq 2$. Assume that

$$\frac{\partial u^0(0)}{\partial n} = 0, \tag{6.2}$$

$$\mathrm{ind}(v^0, 0) = 0, \tag{6.3}$$

$$\frac{\partial u^1(0)}{\partial n} \neq 0, \tag{6.4}$$

$$\frac{\partial^{k+1} u_1^0(0)}{\partial^k \tau \partial n} \neq 0. \tag{6.5}$$

THEOREM 6.3. *Let $u \in C^1([0,T], B_0^r(TM))$ be as given in* (6.1) *satisfying Assumption* (H_0). *Then the following conclusions hold:*

1. *There must be some closed orbits of u separated from $\bar{x} = 0 \in \partial M$, as shown schematically in either Figure* 6.1(c) *or Figure* 6.1(d).
2. *The flow structure after the separation enjoys the following properties:*
   (a) *there are exactly two ∂-saddle points x^- and x^+ of $u(\cdot, t)$ near $\bar{x} = 0$ with one on each side of $\bar{x}$, $x^- < \bar{x} < x^+$;*
   (b) *x^- and x^+ are connected by an extended interior orbit $\gamma(t)$ that consists of orbits of $u(\cdot, t)$; and*
   (c) *the closed orbits of conclusion* (1) *above are enclosed by the extended orbit $\gamma(t)$ and the portion of the boundary between x^- and x^+.*
3. *The whole extended orbit $\gamma(t)$ shrinks to $\bar{x}$ as $t \to t_0$.*

A few remarks are now in order.

Remark 6.4. The closed orbits in Figures 6.1(c), 6.1(d), and 6.2(a)–(c) correspond, in a real fluid, to isolated vortices or, in the case of figure-eight ones, to pairs of counterrotating vortices.

Remark 6.5. The separation occurs as t crosses the critical instant t_0 from either left to right or from right to left; this is dictated by the orientation of u^1 in comparison to that of u^0 in the expansion (3.2).

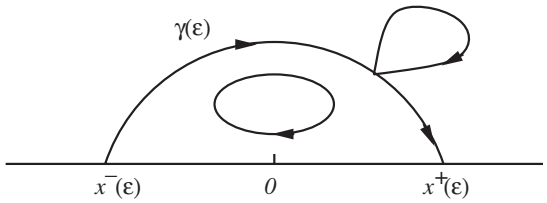


FIG. 6.3. *Convergence of an extended orbit* $\gamma(\varepsilon) \to \bar{x} = 0$.

Remark 6.6. The reattachment of a streamline $\gamma(t)$ to the boundary, as described in conclusion 2(b) above, differs from a particle's streakline (i.e., from its path in real time): as the vortex grows in time, particles are either trapped inside the vortex or pushed into the interior of the fluid, away from the vortex.

Proof of Theorem 6.3. By Theorem 3.5, without loss of generality, we assume that $u(x, t_0 + \varepsilon)$ has two ∂-saddle points $x^-(\varepsilon)$ and $x^+(\varepsilon)$, and $u(x, t_0 - \varepsilon)$ has none. Then both ∂-saddle points $x^-(\varepsilon)$ and $x^+(\varepsilon)$ of $u(x, t_0 + \varepsilon)$ tend to $x = 0$ as $\varepsilon \to 0$.

By assumption, the singular point $\bar{x} \in \partial M$ of $v^0(x)$ is isolated; therefore the stability lemma of extended orbits (Lemma 7.2) in the appendix ensures that both ∂-saddle points $x^-(\varepsilon)$ and $x^+(\varepsilon)$ must be connected by an extended orbit $\gamma(\varepsilon)$ with $\gamma(\varepsilon) \to \bar{x}$ as $\varepsilon \to 0$; see Figure 6.3. Because the sum of the indices of the singular points near $\bar{x} \in \partial M$ of $u(x, t_0 \pm \varepsilon)$ is zero, there exist centers of $v(x, t_0 + \varepsilon)$ near $\gamma(\varepsilon)$, which converge to $\bar{x} \in \partial M$ as $\varepsilon \to 0$. The other assertions are even easier to verify.

In the above theorem, there might be several centers that appear in the recirculation region. However, subject to an additional but generic assumption, the following theorem shows that there must be exactly one center that separates from the boundary near $\bar{x}$.

THEOREM 6.7. *Let* $u \in C^1([0, T], B_0^r(TM))$ *be as given in* (6.1)*; satisfy Assumption* (H_0) *with* $k = 2$*, and*

$$\frac{\partial^2 u_1^0(0)}{\partial x_2^2} \neq 0. \tag{6.6}$$

Then the center separated from $\bar{x} \in \partial M$ *is unique, as shown in Figure* 6.1(c).

Remark 6.8. Consider the Taylor expansion of the vorticity $\omega = -\partial u_2/\partial x_1 + \partial u_1/\partial x_2$ of the vector field u:

$$\begin{cases} \omega(x,t) = \omega^0(x) + \omega^1(x)(t - t_0) + o((t - t_0)^2), \\ \omega^0(x) = \omega(x, t_0), \qquad \omega^1(x) = \dfrac{\partial \omega}{\partial t}(x, t_0). \end{cases} \tag{6.7}$$

Conditions (6.2)–(6.4), and (6.6) are equivalent, respectively, to

$$\omega^0(0) = 0, \tag{6.8}$$

$$\frac{\partial^2 \omega^0(0)}{\partial x_1^2} \neq 0, \tag{6.9}$$

$$\omega^1(0) \neq 0, \tag{6.10}$$

$$\frac{\partial \omega^0(0)}{\partial n} \neq 0. \tag{6.11}$$

Proof of Theorem 6.7. First we observe that conditions (6.2), (6.4), and (6.5) are equivalent to the following conditions on v:

$$\frac{\partial^2 v_1^0(0)}{\partial x_1^2} \neq 0, \tag{6.12}$$

$$\frac{\partial v_1^0(0)}{\partial x_2} \neq 0, \tag{6.13}$$

$$v_1^1(0) \neq 0. \tag{6.14}$$

We only have to prove that the interior singular point of $v(x, t_0+\varepsilon)$ near $\bar{x} \in \partial M$ is unique. By the Taylor expansion, we have

$$\begin{cases} v_1(x, t_0+\varepsilon) = v_1^0(x) + \varepsilon v_1^1(x) + o(|\varepsilon|), \\ v_2(x, t_0, \varepsilon) = v_2^0(x) + \varepsilon v_2^1(x) + o(|\varepsilon|). \end{cases}$$

By the nondivergence of $u^0(x)$ and by (6.12) and (6.13), we derive

$$\begin{cases} v_1^0(x) = \lambda x_2 + \alpha x_1^2 + o(|x_1|^2, |x_2|), \quad \alpha \neq 0,\ \lambda \neq 0, \\ v_2^0(x) = -\alpha x_1 x_2 + x_2 \cdot o(|x|). \end{cases}$$

By (6.14) we have

$$\begin{cases} v_1^1(x) = \beta + O(|x|), \quad \beta < 0, \\ v_2^1(x) = x_2 \cdot O(|x|). \end{cases}$$

Hence the interior singular points $(\tilde{x}_1, \tilde{x}_2)$ of $v(x, t_0+\varepsilon)$ with $\tilde{x}_2 > 0$ satisfy the equations

$$\begin{cases} \alpha x_1^2 + \lambda x_2 + \varepsilon\beta + o(|\varepsilon|, |x_2|, |x_1|^2) = 0, \\ -\alpha x_1 + \varepsilon \cdot O(|x|) + o(|x|) = 0, \\ \alpha \neq 0,\ \lambda \neq 0,\ \beta < 0. \end{cases} \tag{6.15}$$

It follows from the implicit function theorem that a solution $(\tilde{x}_1, \tilde{x}_2)$ with $\tilde{x}_2 > 0$, if it exists, is unique for any $\varepsilon > 0$ sufficiently small. The existence of such a solution to (6.15) is derived by Theorem 6.3, and the proof is complete. □

Remark 6.9. The set of all vector fields u satisfying (6.12) and (6.13) is open and dense in the topological space

$$A = \left\{ u \in C^1([0,T], B_0^3(TM)) \,\middle|\, u \text{ satisfy Assumption } (\mathrm{H}_0) \right\}.$$

Hence, Theorem 6.3 shows that the separation from the boundary of a simple vortex is generic.

7. Appendix. Extended orbits and their stability. The purpose of this appendix is to recall a lemma on stability of extended orbits [11]. We start with a definition.

DEFINITION 7.1. *Let $v \in C^r(TM)$ be a vector field. A curve $\gamma \subset M$ is called an extended orbit of v if*

(i) *it is a union of curves*

$$\gamma = \bigcup_{i=1} \gamma_i;$$

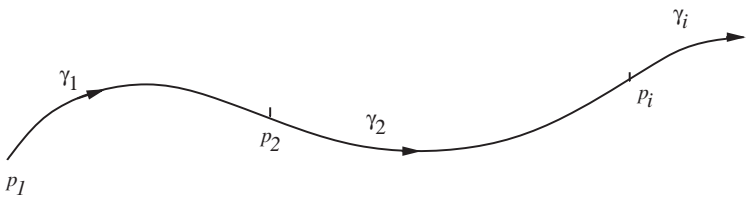


FIG. 7.1. *An extended orbit.*

(ii) *either γ_i is an orbit of v, or γ_i consists of both orbits and singular points of v; and*

(iii) *the ω-limit set of γ_i is the α-limit set of γ_{i+1},*

$$\omega(\gamma_i) = \alpha(\gamma_{i+1}),$$

whenever γ_i and γ_{i+1} are orbits of v; namely, the end points of γ_i are singular points of v, and the starting end point of γ_{i+1} is the finishing end point of γ_i; see Figure 7.1.

The point $p_1 = \alpha(\gamma_1)$ is called the starting point of the extended orbit γ.

The following stability lemma for extended orbits has been proved by Ma and Wang [11] in Step 2 of their proof of Lemma 4.5. We restate it here as a separate lemma since it is quite useful in analyzing the orbits of families of vector fields, and thus in solving some problems in 2-D incompressible fluid flows.

LEMMA 7.2 (stability of extended orbits [11]). *Let $v^n \in C^r(TM)$ be a sequence of 2-D vector fields with $\lim_{n\to\infty} v^n = v \in C^r(TM)$. Suppose that $\gamma^n \subset M$ is an extended orbit of v^n and the starting points p_1^n of γ^n converge to p_1. Then the extended orbits γ^n of v^n converge to an extended orbit γ of v with starting point p_1.*

Acknowledgments. The authors are grateful to an anonymous referee for providing insightful comments. Preliminary results of this investigation were presented at the the Fourth International Conference on Dynamical Systems and Differential Equations held in Wilmington, North Carolina, in May 2002, and at the SIAM Conference on Applications of Dynamical Systems (DS03) held in Snowbird, Utah, in May 2003.

REFERENCES

[1] L. CAFFARELLI, R. KOHN, AND L. NIRENBERG, *On the regularity of the solutions of Navier-Stokes equations*, Comm. Pure Appl. Math., 35 (1982), pp. 771–831.

[2] A. J. CHORIN AND J. E. MARSDEN, *A Mathematical Introduction to Fluid Mechanics*, Springer-Verlag, New York, 1997.

[3] P. CONSTANTIN AND C. FOIAS, *The Navier-Stokes Equations*, University of Chicago Press, Chicago, 1988.

[4] W. E, *Boundary layer theory and the zero-viscosity limit of the Navier-Stokes equation*, Acta Math. Sin. (Engl. Ser.), 16 (2000), pp. 207–218.

[5] W. E AND B. ENGQUIST, *Blow-up of solutions of the unsteady Prandtl's equation*, Comm. Pure Appl. Math., 50 (1997), pp. 1287–1293.

[6] M. GHIL AND S. CHILDRESS, *Topics in Geophysical Fluid Dynamics: Atmospheric Dynamics, Dynamo Theory, and Climate Dynamics*, Springer-Verlag, New York, 1987.

[7] M. GHIL, J.-G. LIU, C. WANG, AND S. WANG, *Boundary-layer separation and adverse pressure gradient for 2-D viscous incompressible flow*, Phys. D, 197 (2004), pp. 149–173.

[8] M. GHIL, T. MA, AND S. WANG, *Structural bifurcation of 2-D incompressible flows*, Indiana Univ. Math. J., 50 (2001), pp. 159–180.

[9] S. Goldstein, *Modern Developments in Fluid Dynamics, Vols.* I *and* II, Dover Publications, New York, 1965.
[10] T. Ma and S. Wang, *Structure of* 2*D incompressible flows with the Dirichlet boundary conditions*, Discrete Contin. Dyn. Syst. Ser. B, 1 (2001), pp. 29–41.
[11] T. Ma and S. Wang, *Structural classification and stability of incompressible vector fields*, Phys. D, 171 (2002), pp. 107–126.
[12] T. Ma and S. Wang, *Boundary layer separation and structural bifurcation for* 2*-D incompressible fluid flows*, Discrete Contin. Dyn. Syst. Ser. A, 10 (2003), pp. 459–472.
[13] A. J. Majda and A. L. Bertozzi, *Vorticity and Incompressible Flow*, Cambridge University Press, Cambridge, UK, 2002.
[14] O. Oleinik and V. N. Samokhin, *Mathematical Models in Boundary Layer Theory*, Chapman and Hall/CRC, Boca Raton, FL, 1999.
[15] L. Prandtl, *Über Flüssigkeitbewegung bei sehr kleiner Reibung*, in Verhandlungen des III. Internationalen Mathematiker-Kongress (Heidelberg, 1904), Leipzig, 1905, pp. 484–491.
[16] H. Schlichting, *Boundary Layer Theory*, 8th ed., Springer-Verlag, Berlin, Heidelberg, 2000.
[17] R. Temam, *Navier-Stokes Equations, Theory and Numerical Analysis*, 3rd ed., North-Holland, Amsterdam, 1984.
[18] X. Wang and R. Temam, *Asymptotic analysis of Oseen type equations in a channel at high Reynolds number*, Indiana Univ. Math. J., 45 (1996), pp. 863–916.

SIAM J. APPL. MATH.
Vol. 65, No. 5, pp. 1597–1614

AN AGE-STRUCTURED EPIDEMIC MODEL IN A PATCHY ENVIRONMENT*

WENDI WANG† AND XIAO-QIANG ZHAO‡

Abstract. A time-delayed epidemic model is proposed to describe the dynamics of disease spread among patches. An age structure is incorporated in order to simulate the phenomenon that some diseases only occur in the adult population. Sufficient conditions are established for global extinction and uniform persistence of the disease. One example shows that the disease could undergo persistence-extinction-persistence switches as the migration rate of juveniles increases, although juveniles are immune and cannot transmit the disease. The second example indicates that this switching phenomenon also happens when juveniles migrate with susceptible adults.

Key words. epidemic model, population dispersal, stage structure, persistence and extinction of disease

AMS subject classifications. 92D30, 34K20, 37N25

DOI. 10.1137/S0036139903431245

1. Introduction.

Population dispersal plays an important role in the dynamics of epidemic diseases. There has been much evidence that diseases spread from one region to other regions due to the immigration of infective individuals. For example, the arrival of new infectives has been demonstrated as being important in the outbreaks of measles observed in Iceland (see [5]); in the 19th century, cholera spread from its ancestral site in the Orient to other parts of the world, producing a pandemic in Europe; in the 14th century, Bubonic plague was transmitted to Europe and killed perhaps one third to one half of Europe's population, which was probably caused by trading ships moving from the East to Europe. Hence, it is important to use mathematical models to understand the effect of population dispersal on the spread of a disease.

Brauer and van den Driessche [4] proposed an epidemic model with population dispersal by adding an immigration term, where infective individuals enter the system at a constant rate. A constant immigration term has a stabilizing effect on the dynamics and tends to increase the minimum number of infective individuals in the models (see [3]).

In [19] we proposed an epidemic model for many patches where immigration rates and emigration rates of infective individuals depend on their numbers, and showed that population dispersal has a significant effect on a disease outbreak. Other researchers [1, 2] have studied multicity epidemic models and their basic reproduction numbers. For the models in [19, 1, 2], it is assumed that all the individuals in one patch have the same ability to transmit a disease and the same risk of being infected by a disease. For some diseases, such as sexual diseases, it is reasonable to consider

*Received by the editors July 10, 2003; accepted for publication (in revised form) September 1, 2004; published electronically May 12, 2005.

http://www.siam.org/journals/siap/65-5/43124.html

†Department of Mathematics, Southwest Normal University, Chongqing, 400715, People's Republic of China (wendi@swnu.edu.cn). The research of this author was supported by the NSF of China (grant 10271096).

‡Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, NF A1C 5S7, Canada (xzhao@math.mun.ca). The research of this author was supported by the NSERC of Canada.

the disease transmission in adult population and neglect transmission in juveniles. In order to incorporate these points into epidemic models with population dispersal, we may divide a population into two stages: juvenile stage and adult stage (see, e.g., [6]). For simplicity, we assume that

(A1) disease transmission occurs only in adult individuals, and juvenile individuals are immune to the disease;

(A2) juvenile individuals do not have the ability to reproduce, and adults are responsible for the reproduction of the population.

In the case of population dispersal among n patches, which may simulate n cities or n countries, let J_i be the number of juvenile individuals in patch i and A_i be the number of adult individuals in patch i. When the patches are isolated, we assume that J_i satisfies the following equation:

$$\frac{dJ_i}{dt} = B_i(A_i)A_i - \mu_i J_i, \tag{1.1}$$

where $B_i(A_i)$ is the per capita birth rate of adult individuals in patch i and μ_i is the per capita death rate of juveniles in patch i. Following [6, 22], we assume that $B_i(A_i)$ satisfy the following basic assumptions:

(A3) $B_i(A_i) > 0$ for all $A_i > 0, i = 1, 2, \ldots, n$;

(A4) $B_i(A_i)$ is continuously differentiable for $A_i > 0$, and $B_i'(A_i) < 0$ for all $A_i > 0$, $i = 1, 2, \ldots, n$.

As mentioned in [6], the following three types of birth functions $B_i(A_i)$ can be found in the biological literature:

(B1) $B_i(A_i) = b_i e^{-a_i A_i}$ with $a_i > 0, b_i > 0$;

(B2) $B_i(A_i) = \frac{p_i}{q_i + A_i^m}$ with $p_i, q_i, m > 0$;

(B3) $B_i(A_i) = \frac{k_i}{A_i} + l_i$ with $k_i > 0, l_i > 0$.

We consider a disease transmission of SIS type. The adult population is divided into two classes: susceptible individuals and infective individuals. Susceptible individuals become infective after contact with infective individuals. Infective individuals return to the susceptible class when they are recovered. Gonorrhea and other sexually transmitted diseases or bacterial infections exhibit this phenomenon. We denote the number of susceptible individuals in patch i by S_i and the number of infective individuals in patch i by I_i. Therefore, $A_i = S_i + I_i$. When the patches are isolated, we assume that S_i and I_i obey the following system:

$$\begin{cases} \dfrac{dS_i}{dt} = R_i(t) - d_i S_i - \beta_i S_i I_i + \gamma_i I_i, \\ \dfrac{dI_i}{dt} = \beta_i S_i I_i - (d_i + \gamma_i) I_i, \end{cases} \tag{1.2}$$

where $R_i(t)$ is the transition rate of juvenile individuals from juvenile stage to adult stage in patch i, d_i is the death rate of adults in patch i, β_i is the contact rate of susceptible individuals with infectious individuals in patch i, and γ_i is the recovery rate of infectives in patch i.

When the patches are connected, we suppose that the immigration rate of susceptible individuals from the jth patch to the ith patch is a_{ij}, the immigration rate of infective individuals from the jth patch to the ith patch is b_{ij}, $-a_{ii} > 0$ is the emigration rate of susceptible individuals in the ith patch, $-b_{ii} > 0$ is the emigration rate of infective individuals in the ith patch, c_{ij} is the immigration rate of juvenile individuals from the jth patch to the ith patch, and $-c_{ii} > 0$ is the emigration rate

of juvenile individuals in the ith patch. Under the above assumptions, we have the following model:

$$\begin{cases} \dfrac{dJ_i}{dt} = B_i(A_i(t))A_i(t) - \mu_i J_i(t) - R_i(t) + \sum\limits_{j=1}^{n} c_{ij}J_j(t), \\ \dfrac{dS_i}{dt} = R_i(t) - d_i S_i - \beta_i S_i I_i + \gamma_i I_i + \sum\limits_{j=1}^{n} a_{ij}S_j, \\ \dfrac{dI_i}{dt} = \beta_i S_i I_i - (d_i + \gamma_i)I_i + \sum\limits_{j=1}^{n} b_{ij}I_j, \\ \quad i = 1, \ldots, n. \end{cases} \tag{1.3}$$

Here we have neglected the death rates and birth rates of individuals during the dispersal process. Thus, we have

$$\sum_{j=1}^{n} a_{ji} = 0, \quad \sum_{j=1}^{n} b_{ji} = 0, \quad \sum_{j=1}^{n} c_{ji} = 0, \quad i = 1, \ldots, n. \tag{1.4}$$

We further assume that all of a_{ij} and b_{ij}, $i \neq j$, are positive so that the $n \times n$ matrix (a_{ij}) and (b_{ij}) are irreducible. Note that system (1.3) indicates that the population can have different demographic structures and different infection forces among different patches.

We now derive a formula for $R_i(t)$ in terms of the parameters and the variables in the model. For simplicity, we define r as the age at which a juvenile in each patch becomes an adult host. Let $J(t,a) := (J_1(t,a), \ldots, J_n(t,a))^T$, where $J_i(t,a)$ is the number of juveniles in patch i at time t with age a. Clearly, $R(t) := (R_1(t), \ldots, R_n(t))^T = J(t,r)$. As in [13, 16], the age-space dynamics of the population is described by

$$\begin{aligned} (\partial_t + \partial_a)J_i(t,a) &= \sum_{j=1}^{n} c_{ij}J_j(t,a) - \left(\sum_{j=1}^{n} c_{ji} + \mu_i\right) J_i(t,a) \\ &= \sum_{j=1}^{n} c_{ij}J_j(t,a) - \mu_i J_i(t,a), \end{aligned} \tag{1.5}$$

with the birth law given by

$$J(t,0) = B(A(t)) := (B_1(A_1(t))A_1(t), \ldots, B_n(A_n(t))A_n(t))^T.$$

Define $V(t,a) := J(t,t-a)$ for all $t \geq a \geq 0$. Then $V(t,a)$ satisfies

$$\frac{\partial V(t,a)}{\partial t} = C_J V(t,a), \quad t \geq a, \tag{1.6}$$

where

$$C_J := \begin{bmatrix} -\mu_1 + c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & -\mu_2 + c_{22} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & \cdots & -\mu_n + c_{nn} \end{bmatrix}.$$

Integrating (1.6) from a to t, we have

$$V(t,a) = \exp(C_J(t-a))V(a,a) = \exp(C_J(t-a))B(A(a)) \quad \forall t \geq a,$$

and hence

$$J(t,s) = V(t,t-s) = \exp(C_J s)B(A(t-s)) \quad \forall t \geq s \geq 0.$$

It then follows that

$$R(t) = J(t,r) = \exp(C_J r)B(A(t-r)). \tag{1.7}$$

By arguments similar to those in [18], we can interpret the (i,j)-element of $\exp(C_J r)$ as the probability that a juvenile individual born in the jth patch will be found at age r in the ith patch.

Motivated by three types of birth functions (see (B1), (B2), and (B3)), we always assume that birth rate $G_i(A_i) := B_i(A_i)A_i$ satisfies $G_i \in C^1([0,\infty), R)$ for each i. Clearly, $G_i(A_i) \geq 0$ for all $A_i \geq 0$. Define $G(A) := (G_1(A_1), \ldots, G_n(A_n))$ for all $A = (A_1, \ldots, A_n) \in R^n_+$.

Note that $R(t)$ does not depend on the variables of juveniles. The S_i and I_i equations can be decoupled from the J_i equations in (1.3) to obtain the following reduced model:

$$\begin{cases} \dfrac{dS_i}{dt} = R_i(t) - d_i S_i - \beta_i S_i I_i + \gamma_i I_i + \displaystyle\sum_{j=1}^{n} a_{ij} S_j, \\ \dfrac{dI_i}{dt} = \beta_i S_i I_i - (d_i + \gamma_i) I_i + \displaystyle\sum_{j=1}^{n} b_{ij} I_j, \\ (R_1(t), \ldots, R_n(t))^T = \exp(C_J r) G(A(t-r)), \\ S(\theta) = \phi(\theta), I(\theta) = \psi(\theta), \ \forall\, \theta \in [-r, 0], \ (\phi, \psi) \in C^2_+, \ i = 1, \ldots, n, \end{cases} \tag{1.8}$$

where $C_+ := C([-r,0], R^n_+)$.

By [12, Theorem 5.2.1], it follows that for any $(\phi, \psi) \in C^2_+$ there is a unique solution $(S(t,\phi,\psi), I(t,\phi,\psi))$ of (1.8), and $S(t,\phi,\psi) \geq 0$, $I(t,\phi,\psi)) \geq 0$ for all $t \geq 0$ in its maximal interval of existence. The purpose of this paper is to study the long-term behavior of the model system (1.8) and the effects of population dispersal on the spread of the disease.

The remaining parts of this paper are organized as follows. Section 2 presents the discussion of the disease free equilibrium and reproduction number for model (1.8). In section 3, we establish sufficient conditions for the global extinction and persistence, respectively, of the disease. Section 4 contains the application of the general results to a special case of the model with two patches. A discussion section completes the paper.

2. Disease free equilibrium. In this section, we establish sufficient conditions for the existence and uniqueness of a disease free equilibrium so that we can define the reproduction number for the model (1.8). We start with the introduction of cooperative systems of delay differential equations.

For $x, y \in R^n$, we write $x \geq y$ if $x - y \in R^n_+$, $x > y$ if $x - y \in R^n_+ \setminus \{0\}$, $x \gg y$ if $x - y \in \text{int}(R^n_+)$. Let r be a nonnegative real number and $x(t)$ be a continuous function from $[-r, \sigma)$ to R^n $(\sigma > 0)$. For each $t \in [0, \sigma)$, we define $x_t \in C([-r,0], R^n)$ by $x_t(\theta) = x(t+\theta)$ for all $\theta \in [-r, 0]$. Consider an autonomous delay system

$$\frac{dx(t)}{dt} = f(x_t), \tag{2.1}$$

where $f : D \to R^n$ is Lipschitz continuous and D is an open subset of $C([-r,0], R^n)$. For any $\phi \in D$, let $x(t,\phi)$ be the unique solution of (2.1) satisfying $x(\theta,\phi) = \phi(\theta)$

for all $\theta \in [-r, 0]$. For $\phi, \psi \in C([-r, 0], R^n)$ we write $\phi \leq \psi$ if $\phi(\theta) \leq \psi(\theta)$ for all $\theta \in [-r, 0]$. System (2.1) is said to be cooperative if the following quasi-monotone condition holds:

(Q) Whenever $\phi \leq \psi$ and $\phi_i(0) = \psi_i(0)$ holds for some i, then $f_i(\phi) \leq f_i(\psi)$.

In particular, (2.1) with $r = 0$, an autonomous ordinary differential system, is cooperative if off-diagonal elements of the Jacobian matrix of f at any point in the domain are nonnegative. It is well known that the comparison principle holds for cooperative systems. For the basic theory of cooperative systems, we refer to [12].

Recall that the stability modulus of an $n \times n$ matrix M, denoted by $s(M)$, is defined by

$$s(M) := \max\{\mathrm{Re}\lambda : \ \lambda \text{ is an eigenvalue of } M\}.$$

If M has nonnegative off-diagonal elements and is irreducible, then $s(M)$ is a simple eigenvalue of M with a (componentwise) positive eigenvector (see, e.g., [14, Theorem A.5]).

Let $(S_1^*, \ldots, S_n^*, 0, \ldots, 0)$ be a disease free equilibrium of (1.8). Then it is easy to see that $(S_1^*, \ldots, S_n^*)$ is an equilibrium of the following ordinary differential system:

$$\frac{dS_i}{dt} = \overline{R}_i(S) - d_i S_i + \sum_{j=1}^{n} a_{ij} S_j, \quad i = 1, \ldots, n, \tag{2.2}$$

where $S = (S_1, \ldots, S_n) \in R_+^n$, $(\overline{R}_1(S), \ldots, \overline{R}_n(S))^T := \exp(C_J r) G(S)$. Set $\exp(C_J r) = (p_{ij})$. Since (1.6) is cooperative, it follows that $p_{ij} \geq 0$ for all $1 \leq i, j \leq n$. Further, we have

$$\overline{R}_i(S) = \sum_{j=1}^{n} p_{ij} G_j(S_j). \tag{2.3}$$

Define $H = (H_1, \ldots, H_n) : R_+^n \to R^n$ by

$$H_i(S) = \overline{R}_i(S) - d_i S_i + \sum_{j=1}^{n} a_{ij} S_j, \quad \forall S \in R_+^n, \ 1 \leq i \leq n.$$

Clearly, $H(0) \geq 0$. If $G_i'(A_i) \geq 0$ for all $A_i \geq 0$, $1 \leq i \leq n$, then $\frac{\partial H_i(S)}{\partial S_j} > 0$ for all $S \in R_+^n$, $1 \leq i \neq j \leq n$.

For any $\alpha \in (0, 1)$ and $S \gg 0$, by (2.3) and (A4), we have

$$\overline{R}_i(\alpha S) - d_i \alpha S_i + \sum_{j=1}^{n} a_{ij} \alpha S_j > \alpha \left[\overline{R}_i(S) - d_i S_i + \sum_{j=1}^{n} a_{ij} S_j \right]$$

for each $i = 1, 2, \ldots, n$, and hence $H(\alpha S) \gg \alpha H(S)$, which means that H is strongly sublinear on R_+^n (see, e.g., [21]).

For $S \gg 0$, let $M(S)$ denote the following matrix:

$$\begin{bmatrix} p_{11}B_1(S_1) - d_1 + a_{11} & a_{12} + p_{12}B_2(S_2) & \cdots & a_{1n} + p_{1n}B_n(S_n) \\ p_{21}B_1(S_1) + a_{21} & p_{22}B_2(S_2) - d_2 + a_{22} & \cdots & p_{2n}B_n(S_n) + a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1}B_1(S_1) + a_{n1} & p_{n2}B_2(S_2) + a_{n2} & \cdots & p_{nn}B_n(S_n) - d_n + a_{nn} \end{bmatrix}.$$

In what follows we use notation $K = (k, \ldots, k) \in R^n$ for each $k \in R$, and let $H'(S)$ be the Jacobian matrix of H at $S \in R_+^n$. Assume that the following hold:

(A5) $\frac{dG_i(A_i)}{dA_i} \geq 0$ for all $A_i \in (0,\infty)$, $i = 1,\dots,n$;
(A6) $s(M(\infty)) < 0$;
(A7) $s(H'(0)) > 0$ if $H(0) = 0$.

Biologically, (A5) means that the birth rate of juveniles in each patch increases as the number of adult individuals in the same patch increases. Roughly speaking, $M(\infty)$ is the per capita reproduction matrix for susceptible hosts in all patches when their numbers are sufficiently large in every patch. (A6) implies that the sizes of susceptible host populations decrease as long as they are large enough, which prevents unbounded susceptible populations. Similarly, (A7) implies that the sizes of susceptible host populations increase as long as they are small, which prevents the extinction of susceptible hosts. These can be seen from the proof of the following theorem.

THEOREM 2.1. *Let* (A1)–(A7) *hold. Then* (1.8) *has a unique disease free equilibrium* $E_0 = (S_1^*, S_2^*, \dots, S_n^*, 0, \dots, 0)$.

Proof. By (A6), we have $s(M(K)) < 0$ if k is large enough. Let $\bar{v} = (\bar{v}_1, \dots, \bar{v}_n)$ be a positive eigenvector associated with $s(M(K))$. Choose $l > 0$ large enough such that $l\bar{v}_i > k,\ i = 1, 2, \dots, n$. If we rewrite (2.2) as $\frac{dS}{dt} = H(S)$, by (A4) we have

$$0 > s(M(K))l\bar{v} = M(K)l\bar{v} > H(l\bar{v}), \qquad \forall t \geq 0. \tag{2.4}$$

Let $S(t, l\bar{v})$ be the solution of (2.2) satisfying $S(0, l\bar{v}) = l\bar{v}$. Since (2.2) is cooperative, it follows from (2.4) that $S(t, l\bar{v})$ is nonincreasing in $t \geq 0$ and converges to an equilibrium as t approaches infinity (see, e.g., [12, Corollary 5.2.2]). It follows that every solution $S(t, x)$ of (2.2) in R_+^n exists globally on $[0,\infty)$. Let $\Phi_0(t)$ be the solution semiflow of (2.2) on R_+^n, that is, $\Phi_0(t)x = S(t, x)$ for all $t \geq 0$, $x \in R_+^n$. Then $\Phi_0(t)$ is strongly monotone on R_+^n in the sense that $x > y$ implies $\Phi_0(t)x \gg \Phi_0(t)y$ for all $t > 0$. Note that H is strongly sublinear on R_+^n. By [21, Proposition 2.2] as applied to $\Phi_0(t)\,(t > 0)$, system (2.2) has at most one positive (componentwise) equilibrium. In the case where $H(0) > 0$, $S(t, 0)$ is nondecreasing in $t \geq 0$ and converges to an equilibrium $\bar{x} > 0$ as t approaches infinity, and hence $\bar{x} = \Phi_0(t)\bar{x} \gg \Phi_0(t)0 = S(t, 0) \geq 0$ for any $t > 0$. Thus, the standard comparison method implies that the positive equilibrium $\bar{x}$ is globally attractive, and hence asymptotically stable, for (2.2) in R_+^n. In the case where $H(0) = 0$, [21, Corollary 3.2] implies that (2.2) has a positive equilibrium $S^* = (S_1^*, S_2^*, \dots, S_n^*)$, which is globally asymptotically stable for $S \in R_+^n \setminus \{0\}$. Consequently, (1.8) has a unique disease free equilibrium. ∎

Now we are ready to consider a basic reproduction number for the system (1.8). The basic reproduction number, denoted by $\mathcal{R}_0$, is "the expected number of secondary cases produced, in a completely susceptible population, by a typical infective individual" (see [7]). For the case of a single infected compartment, $\mathcal{R}_0$ is simply the product of the infection rate and the mean duration of the infection. For the model (1.8) with several infected compartments, the basic reproduction number can be defined as the number of new infections produced by a typical infective individual in the population at the disease free equilibrium (see [18]). Following [18], we let $x_i = I_i$ for $i = 1, \dots, n$ and $x_{n+i} = S_i$ for $i = 1, \dots, n$. Then model (1.8) can be rewritten as

$$\frac{dx_i}{dt} = f_i(x) = \mathcal{F}_i(x) - \mathcal{V}_i(x), \quad i = 1, \dots, 2n,$$

where

$$\mathcal{F}_i(x) = \beta_i x_i x_{n+i}, \quad \mathcal{V}_i(x) = (d_i + \gamma_i)x_i - \sum_{j=1}^{n} b_{ij}x_j \quad \text{for} \quad i = 1, \dots, n.$$

Here we do not write the expressions for $\mathcal{F}_i(x), \mathcal{V}_i(x)$ for $i = n+1, \ldots, 2n$, since they are not important at this moment. Furthermore, the disease free equilibrium E_0 now becomes $(0, \ldots, 0, S_1^*, S_2^*, \ldots, S_n^*)$.

For $i = 1, \ldots, n, \mathcal{F}_i(x)$ is the rate of appearance of new infections in compartment i, and $\mathcal{V}_i(x)$ is the net decreasing rate of infectives in compartment i due to infective flows inside the system of infected compartments.

If $\mathcal{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_{2n})$ and $\mathcal{V} = (\mathcal{V}_1, \ldots, \mathcal{V}_{2n})$, let us partition the derivatives $D\mathcal{F}(E_0)$ and $D\mathcal{V}(E_0)$ as

$$D\mathcal{F}(E_0) = \begin{bmatrix} F & 0 \\ 0 & 0 \end{bmatrix}, \quad D\mathcal{V}(E_0) = \begin{bmatrix} V & 0 \\ J_3 & J_4 \end{bmatrix},$$

where

$$F = \left[\frac{\partial \mathcal{F}_i}{\partial x_j}(E_0)\right]_{n\times n}, \quad V = \left[\frac{\partial \mathcal{V}_i}{\partial x_j}(E_0)\right]_{n\times n}.$$

Then it is easy to obtain

$$F = \begin{bmatrix} \beta_1 S_1^* & 0 & \cdots & 0 \\ 0 & \beta_2 S_2^* & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \beta_n S_n^* \end{bmatrix}$$

and

$$V = \begin{bmatrix} d_1 + \gamma_1 - b_{11} & -b_{12} & \cdots & -b_{1n} \\ -b_{21} & d_2 + \gamma_2 - b_{22} & \cdots & -b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ -b_{n1} & -b_{n2} & \cdots & d_n + \gamma_n - b_{nn} \end{bmatrix}.$$

According to [7, 18], the matrix FV^{-1} is called the next generation matrix, and its spectral radius is defined as the reproduction number for (1.8), that is,

$$\mathcal{R}_0 := \rho(FV^{-1}).$$

Let M_1 denote the matrix

$$\begin{bmatrix} \beta_1 S_1^* - d_1 - \gamma_1 + b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & \beta_2 S_2^* - d_2 - \gamma_2 + b_{22} & \cdots & b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ b_{n1} & b_{n2} & \cdots & \beta_n S_n^* - d_n - \gamma_n + b_{nn} \end{bmatrix}.$$

Clearly, M_1 is irreducible and has nonnegative off-diagonal elements. Then $s(M_1)$ is a simple eigenvalue of M_1 with a positive eigenvector. Furthermore, $M_1 = F - V$. Thus, the following observation is implied by the proof of [18, Theorem 2] with $J_1 = M_1$.

LEMMA 2.1. *There hold two equivalences:*

$$\mathcal{R}_0 > 1 \Leftrightarrow s(M_1) > 0, \qquad \mathcal{R}_0 < 1 \Leftrightarrow s(M_1) < 0. \tag{2.5}$$

By Lemma 2.1 and the proof of Theorem 2.1, it easily follows that $\mathcal{R}_0 < 1$ implies that E_0 is asymptotically stable, and that $\mathcal{R}_0 > 1$ implies that E_0 is unstable.

3. Global dynamics. The purpose of this section is to discuss the global extinction and persistence of the disease described by model (1.8).

Define $e_{ij} := \max\{a_{ij}, b_{ij}\}$, $i \neq j$, and $e_{ii} := \min\{a_{ii}, b_{ii}\}$. Let $\overline{M}(S)$ denote the matrix function

$$\begin{bmatrix} p_{11}B_1(S_1) - d_1 + e_{11} & e_{12} + p_{12}B_2(S_2) & \cdots & e_{1n} + p_{1n}B_n(S_n) \\ p_{21}B_1(S_1) + e_{21} & p_{22}B_2(S_2) - d_2 + e_{22} & \cdots & p_{2n}B_n(S_n) + e_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1}B_1(S_1) + e_{n1} & p_{n2}B_2(S_2) + e_{n2} & \cdots & p_{nn}B_n(S_n) - d_n + e_{nn} \end{bmatrix}.$$

Consider a time-delayed cooperative system

$$\frac{dA_i}{dt} = \overline{R}_i(A(t-r)) - d_iA_i + \sum_{j=1}^{n} e_{ij}A_j, \qquad i = 1, \ldots, n. \tag{3.1}$$

Assume that (A1)–(A5) and (A7) hold and that $s(\overline{M}(\infty)) < 0$. By [21, Theorem 3.2] and an argument similar to the proof of Theorem 2.1, it easily follows that system (3.1) admits a unique positive equilibrium $\bar{A} = (\bar{A}_1, \ldots, \bar{A}_n)$, and $\bar{A}$ is globally stable in $C_+ \setminus \{0\}$. Let $\overline{M}_1$ denote the matrix

$$\begin{bmatrix} \beta_1\bar{A}_1 - d_1 - \gamma_1 + b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & \beta_2\bar{A}_2 - d_2 - \gamma_2 + b_{22} & \cdots & b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ b_{n1} & b_{n2} & \cdots & \beta_n\bar{A}_n - d_n - \gamma_n + b_{nn} \end{bmatrix}.$$

We then have the following result.

THEOREM 3.1. *Let* (A1)–(A5) *and* (A7) *hold. Assume that* $s(\overline{M}(\infty)) < 0$. *If* $s(\overline{M}_1) < 0$, *then* E_0 *is globally attractive for* (1.8) *in* $(C_+ \setminus \{0\}) \times C_+$.

Proof. Note that $A_i = S_i + I_i$ for all $1 \leq i \leq n$. By (1.8), we have

$$\frac{dA_i}{dt} = \overline{R}_i(A(t-r)) - d_iA_i + \sum_{j=1}^{n} a_{ij}S_j + \sum_{j=1}^{n} b_{ij}I_j, \qquad i = 1, \ldots, n. \tag{3.2}$$

Then we have

$$\frac{dA_i}{dt} \leq \overline{R}_i(A(t-r)) - d_iA_i + \sum_{j=1}^{n} e_{ij}A_j, \quad i = 1, \ldots, n. \tag{3.3}$$

Given $(\phi, \psi) \in (C_+ \setminus \{0\}) \times C_+$, let $(S(t), I(t))$ be the solution of (1.8). It then follows from the comparison principle that

$$\limsup_{t\to\infty} A_i(t) \leq \bar{A}_i, \quad i = 1, \ldots, n.$$

As a consequence, each positive solution of (1.8) satisfies

$$\limsup_{t\to\infty} S_i(t) \leq \bar{A}_i, \quad i = 1, \ldots, n.$$

Choose a sufficiently small positive number ϵ such that $s(\overline{M}_1 + \epsilon\beta) < 0$, where $\beta = \operatorname{diag}(\beta_1, \ldots, \beta_n)$. Then there exists a $t_0 > 0$ such that $S_i(t) \leq \bar{A}_i + \epsilon$ for all $t \geq t_0$, $i = 1, \ldots, n$. It follows that

$$\frac{dI_i}{dt} \leq (\beta_i(\bar{A}_i + \epsilon) - d_i - \gamma_i)I_i + \sum_{j=1}^{n} b_{ij}I_j, \quad \forall t \geq t_0,\ i = 1, \ldots, n. \tag{3.4}$$

Let us consider an auxiliary system

$$\frac{dI_i}{dt} = (\beta_i(\bar{A}_i + \epsilon) - d_i - \gamma_i)I_i + \sum_{j=1}^{n} b_{ij}I_j, \quad i = 1, \dots, n. \tag{3.5}$$

Since $s(\bar{M}_1 + \epsilon\beta) < 0$, it is easy to see that all the solutions of (3.5) tend to the origin as t tends to infinity. Since (3.5) is cooperative, it follows from the comparison principle that $I_i(t) \to 0$ as $t \to \infty, i = 1, \dots, n$.

Let $\Phi(t) : C_+^2 \to C_+^2$ be the solution semiflow of (1.8), that is,

$$\Phi(t)(\phi, \psi) = (S_t(\phi, \psi), I_t(\phi, \psi)),$$

where $S_t(\phi, \psi)(\theta) = S(t + \theta, \phi, \psi)$, $I_t(\phi, \psi)(\theta) = I(t + \theta, \phi, \psi)$ for all $\theta \in [-r, 0]$, and $(S(t, \phi, \psi), I(t, \phi, \psi))$ is the solution of (1.8) through (ϕ, ψ). For any $(\phi, \psi) \in (C_+ \setminus \{0\}) \times C_+$, it is easy to see that $S(t, \phi, \psi) > 0$ for $t > r$. Let $\omega = \omega(\phi, \psi)$ be the omega limit set of $\Phi(t)(\phi, \psi)$. Since $I(t, \phi, \psi) \to 0$ as $t \to \infty$, there holds $\omega = \bar{\omega} \times \{0\}$. We claim that $\bar{\omega} \neq \{0\}$. Assume that, by contradiction, $\bar{\omega} = \{0\}$. Then $\lim_{t\to\infty} (S(t, \phi, \psi), I(t, \phi, \psi)) = (0, 0)$. Consequently, for any sufficiently small $\xi > 0$ there is a $T^* > 0$ such that

$$S_i(t) < \xi, \quad I_i(t) < \xi, \qquad \forall\, i = 1, \dots, n, \quad t \geq T^*, \tag{3.6}$$

where the dependence of S_i and I_i on (ϕ, ψ) is suppressed. It follows from (1.8) that

$$\frac{dS_i}{dt} \geq \sum_{j=1}^{n} p_{ij}G_j(S_j(t - r)) - (d_i + \beta_i\xi)S_i + \sum_{j=1}^{n} a_{ij}S_j, \quad i = 1, \dots, n, \tag{3.7}$$

for $t \geq T^*$. Consider the cooperative system of delay differential equations

$$\frac{dS_i}{dt} = \sum_{j=1}^{n} p_{ij}G_j(S_j(t - r)) - (d_i + \beta_i\xi)S_i + \sum_{j=1}^{n} a_{ij}S_j, \quad i = 1, \dots, n. \tag{3.8}$$

In the case where $H(0) = 0$, since $s(H'(0)) > 0$, we can further restrict $\xi > 0$ such that $s(H'(0) - \xi\beta) > 0$, where $\beta = \mathrm{diag}(\beta_1, \dots, \beta_n)$. Let $x(t, \phi)$ be the solution of (3.8) satisfying $x_0(\phi) = \phi \in C_+$. By [21, Theorem 3.1], it follows that either $x(t, \phi)$ is unbounded for every $\phi \in C_+ \setminus \{0\}$ or there exists an equilibrium $x^* \gg 0$ of (3.8) such that $\liminf_{t\to\infty} x(t, \phi) \geq x^*$ for every $\phi \in C_+ \setminus \{0\}$. In the case where $H(0) > 0$, [12, Corollary 5.2.2]) implies that $x(t, 0)$ is nondecreasing in t. Moreover, if $x(t, 0)$ is bounded, then $x(t, 0)$ converges to an equilibrium $x^* > 0$. Since the ordinary differential system (3.8) with $r = 0$ is cooperative and irreducible, it follows that $x^* \gg 0$. Consequently, the comparison principle (see [12, Theorem 5.1.1]) implies that either $S(t)$ is unbounded or $\liminf_{t\to\infty} S(t) \geq x^*$, which contradicts $\lim_{t\to\infty} S_i(t) = 0$ for all $i = 1, \dots, n$.

It is easy to see that

$$\Phi(t) \mid_{\omega} (\phi, 0) = (\Phi_1(t)\phi, 0),$$

where $\Phi_1(t)$ is the solution semiflow of the following system:

$$\frac{dS_i}{dt} = \overline{R}_i(S(t - r)) - d_iS_i + \sum_{j=1}^{n} a_{ij}S_j, \quad i = 1, \dots, n. \tag{3.9}$$

By [9, Lemma 2.1′], ω is an internal chain transitive set for $\Phi(t)$, and hence $\bar{\omega}$ is an internal chain transitive set for $\Phi_1(t)$. By the proof of Theorem 2.1 and [21, Theorem 3.2], it follows that S^* is globally asymptotically stable for (3.9) in $C_+ \setminus \{0\}$. Since $\bar{\omega} \neq \{0\}$, we have $\bar{\omega} \cap W^s(S^*) \neq \emptyset$. By [9, Theorem 3.1 and Remark 4.6], we then get $\bar{\omega} = S^*$. This proves $\omega = \{(S^*, 0)\}$. Consequently, $(S(t,\phi,\psi), I(t,\phi,\psi)) \to (S^*, 0)$ as $t \to \infty$. □

REMARK 3.1. *Let* (A1)–(A5) *and* (A7) *hold. Assume that* $s(\overline{M}(\infty)) < 0$. *Note that* $\bar{A}_i$ *is near to* S_i^* *for* $i = 1, \ldots, n$, *provided that the dispersal coefficients* a_{ij} *and* b_{ij} *are small for* $i, j = 1, \ldots, n$. *Hence, it is one implication of Theorem* 3.1 *that the disease, which dies out in the absence of population dispersal, still dies out when population dispersal is weak.*

COROLLARY 3.1. *Let* (A1)–(A5) *and* (A7) *hold. Assume that* $s(\overline{M}(\infty)) < 0$. *If* $a_{ij} = b_{ij}$ *for* $i = 1, \ldots, n, j = 1, \ldots, n$, *and* $\mathcal{R}_0 < 1$, *then* E_0 *is globally attractive for* (1.8) *in* $(C_+ \setminus \{0\}) \times C_+$.

Proof. In this special case, $e_{ij} = a_{ij} = b_{ij}$ for all $1 \le i, j \le n$, and hence $\bar{A} = S^*$, $\bar{M} = M$, and $\bar{M}_1 = M_1$. By Lemma 2.1, $s(M_1) < 0$. Thus, the conclusion follows from Theorem 3.1. □

The subsequent result shows that the disease is uniformly persistent in the case where $\mathcal{R}_0 > 1$.

THEOREM 3.2. *Let* (A1)–(A7) *hold and* $\mathcal{R}_0 > 1$. *If* $s(\overline{M}(\infty)) < 0$, *then there is a positive constant* ϵ *such that every solution* $(S(t,\phi,\psi), I(t,\phi,\psi))$ *of* (1.8) *with* $(\phi,\psi) \in C_+^2$ *and* $\psi(0) > 0$ *satisfies*

$$\liminf_{t\to\infty} I_i(t,\phi,\psi) \ge \epsilon \qquad \forall 1 \le i \le n.$$

Furthermore, (1.8) *admits at least one (componentwise) positive equilibrium.*

Proof. Define

$$\begin{aligned} X &:= \{(\phi,\psi) \in C_+ \times C_+\}, \\ X_0 &:= \{(\phi,\psi) \in X : \psi(0) > 0\}, \\ \partial X_0 &:= X \setminus X_0. \end{aligned}$$

By the form of (1.8), it is easy to see that both X and X_0 are positively invariant. Clearly, ∂X_0 is relatively closed in X, and $\partial X_0 = \{(\phi,\psi) \in X : \psi(0) = 0\}$. Furthermore, system (1.8) is point dissipative in R_+^n since nonnegative solutions of (1.8) are ultimately bounded (see the proof of Theorem 3.1).

Let $(S(t,\phi,\psi), I(t,\phi,\psi))$ be the solution of (1.8) through (ϕ,ψ), and define

$$M_\partial := \{(\phi,\psi) \in X : (S_t(\phi,\psi), I_t(\phi,\psi)) \in \partial X_0 \ \forall\ t \ge 0\}.$$

Then we have the following claim.

CLAIM 1. $M_\partial = \{(\phi,\psi) \in \partial X_0 : I(t,\phi,\psi) = 0 \ \forall\ t \ge 0\}$.

Indeed, it suffices to prove that for each $(\phi,\psi) \in M_\partial$, $I(t,\phi,\psi) = 0$ for all $t \ge 0$. Suppose that this does not hold. Then there exist some $i_0, 1 \le i_0 \le n$, and $t_0 \ge 0$ such that $I_{i_0}(t_0,\phi,\psi) > 0$. We partition $\{1, 2, \ldots, n\}$ into two sets Q_1 and Q_2 such that

$$\begin{aligned} I_i(t_0,\phi,\psi) &= 0 \quad \forall i \in Q_1, \\ I_i(t_0,\phi,\psi) &> 0 \quad \forall i \in Q_2. \end{aligned}$$

By the definition of M_∂ we see that Q_1 is nonempty. Q_2 is also nonempty since $I_{i_0}(t_0, \phi, \psi) > 0$. For any $j \in Q_1$, we have

$$\left.\frac{dI_j(t,\phi,\psi)}{dt}\right|_{t=t_0} \geq b_{ji_0} I_{i_0}(t_0, \phi, \psi) > 0.$$

It follows that there is an $\epsilon_0 > 0$ such that $I_j(t, \phi, \psi) > 0, j \in Q_1$, for $t_0 < t < t_0 + \epsilon_0$. Clearly, we can restrict $\epsilon_0 > 0$ to be small enough such that $I_i(t, \phi, \psi) > 0, i \in Q_2$, for $t_0 < t < t_0 + \epsilon_0$. This means that $(S_t(\phi, \psi), I_t(\phi, \psi)) \notin \partial X_0$ for $t_0 < t < t_0 + \epsilon_0$, which contradicts the assumption that $(\phi, \psi) \in M_\partial$.

Define $M_3 := \operatorname{diag}(\beta_1, \ldots, \beta_n)$. Then we can choose $\eta > 0$ small enough such that $s(M_1 - \eta M_3) > 0$. Consider the following system:

$$\frac{dS_i}{dt} = \overline{R}_i(S(t-r)) - (d_i + \beta_i\epsilon_1)S_i + \sum_{j=1}^{n} a_{ij}S_j, \qquad i = 1, \ldots, n. \tag{3.10}$$

First, as in our previous analysis of system (2.2), we can restrict $\epsilon_1 > 0$ to be small enough such that (3.10) admits a unique positive equilibrium $S^*(\epsilon_1)$, which is globally asymptotically stable for (3.10) (see the proof of Theorem 2.1 and [21, Theorem 3.2]). By the implicit function theorem, it follows that $S^*(\epsilon_1)$ is continuous in ϵ_1. Thus, we can restrict ϵ_1 to be small enough such that $S^*(\epsilon_1) > S^* - \eta$. Let $S(t, \phi, \psi) = (S_1(t, \phi, \psi), \ldots, S_n(t, \phi, \psi))$ and $I(t, \phi, \psi) = (I_1(t, \phi, \psi), \ldots, I_n(t, \phi, \psi))$. Then we further have the following claim.

CLAIM 2. $\limsup_{t\to\infty} \max_i\{I_i(t, \phi, \psi)\} > \epsilon_1$ *for all* $(\phi, \psi) \in X_0$.

Suppose, for the sake of contradiction, that there is a $T > 0$ such that $0 < I_i(t) \leq \epsilon_1$, $i = 1, 2, \ldots, n$, for all $t \geq T$. Then for $t \geq T$ we have

$$\frac{dS_i}{dt} \geq \overline{R}_i(S(t-r)) - (d_i + \beta_i\epsilon_1)S_i + \sum_{j=1}^{n} a_{ij}S_j, \quad i = 1, \ldots, n. \tag{3.11}$$

Since the equilibrium $S^*(\epsilon_1)$ of (3.10) is globally asymptotically stable and $S^*(\epsilon_1) > S^* - \eta$, there is a $T_1 > 0$ such that $S(t) \geq S^* - \eta$ for $t \geq T + T_1$. As a consequence, for $t > T + T_1$, we have

$$\frac{dI_i}{dt} \geq \beta_i(S_i^* - \eta)I_i - (d_i + \gamma_i)I_i + \sum_{j=1}^{n} b_{ij}I_j, \qquad i = 1, \ldots, n. \tag{3.12}$$

Since the matrix $M_1 - \eta M_3$ has a positive eigenvalue $s(M_1 - \eta M_2)$ with a positive eigenvector, it is easy to see that $I_i(t) \to \infty$ as $t \to \infty$, $i = 1, 2, \ldots, n$, which leads to a contradiction.

Note that the delay differential system

$$\frac{dS_i}{dt} = \overline{R}_i(S(t-r)) - d_iS_i + \sum_{j=1}^{n} a_{ij}S_j, \qquad i = 1, \ldots, n, \tag{3.13}$$

admits a global asymptotic stable equilibrium S^* in $C([-r, 0], R_+^n) \setminus \{0\}$. In the case where $H(0) = 0$, there are exactly two equilibria $(0, 0)$ and $E_0 = (S^*, 0)$ in M_∂. By Claim 2, we see that $(0, 0)$ and E_0 are isolated invariant sets in X, $W^s((0, 0)) \cap X_0 = \emptyset$, and $W^s(E_0) \cap X_0 = \emptyset$. Clearly, every forward orbit in M_∂ converges to either $(0, 0)$ or E_0, and $(0, 0)$ and E_0 are acyclic in M_∂. By [17, Theorem 4.6], it follows that

the solution semiflow associated with (1.8) is uniformly persistent with respect to $(X_0, \partial X_0)$. In the case where $H(0) > 0$, there is only one equilibrium $E_0 = (S^*, 0)$ in M_∂, and every forward orbit in M_∂ converges to E_0. By a similar argument, we see that the solution semiflow is uniformly persistent with respect to $(X_0, \partial X_0)$. Consequently, [15, Theorem A.2] with $Z = C([-r, 0], R_+^{2n})$ and $e = (1, \dots, 1) \in R^{2n}$ implies the required uniform persistence of solutions of (1.8).

By applying [20, Theorem 2.4] to the autonomous semiflow generated by the ordinary differential system (1.8) with $A(t-r)$ replaced by $A(t)$, we conclude that system (1.8) has an equilibrium $(\bar{S}, \bar{I}) \in X_0$. Then $\bar{S} \in R_+^n$ and $\bar{I} \in \text{int}(R_+^n)$. We further claim that $\bar{S} \in R_+^n \setminus \{0\}$. Suppose that $\bar{S} = 0$. By the second equation in (1.4), we then get $0 = -\sum_{i=1}^n (\mu_i + \gamma_i)\bar{I}_i$, and hence $\bar{I}_i = 0,\ i = 1, 2, \dots, n$, a contradiction. By the first equation in (1.8) and the irreducibility of the cooperative matrix (a_{ij}), it follows that $\bar{S} = S(t, \bar{S}, \bar{I}) \in \text{int}(R_+^n)$ for all $t > 0$. Then $(\bar{S}, \bar{I})$ is a componentwise positive equilibrium of (1.8). □

4. A case study. In order to illustrate the results of the last section, we suppose that the birth rates are in the form of (B3) and that there are only two patches. Then model (1.8) becomes

$$
(4.1)\quad \begin{cases} \dfrac{dS_i}{dt} = R_i(t) - (d_i + a_i)S_i - \beta_i S_i I_i + \gamma_i I_i + a_j S_j, \quad i, j = 1, 2, i \neq j, \\ \dfrac{dI_i}{dt} = \beta_i S_i I_i - (d_i + \gamma_i + b_i)I_i + b_j I_j, \quad i, j = 1, 2, i \neq j, \\ (R_1(t), R_2(t))^T = \exp(C_J r)(k_1 + l_1 A_1(t-r), k_2 + l_2 A_2(t-r))^T, \end{cases}
$$

where a_i, b_i, and c_i are the migration rates of adult susceptibles, adult infectives, and juveniles, respectively, of patch i. Set $\exp(C_J r) = (p_{ij})$. Then we have

$$
\begin{aligned} R_1(t) &= p_{11}(k_1 + l_1 A_1(t-r)) + p_{12}(k_2 + l_2 A_2(t-r)), \\ R_2(t) &= p_{21}(k_1 + l_1 A_1(t-r)) + p_{22}(k_2 + l_2 A_2(t-r)). \end{aligned}
$$

Set

$$
M_4 := \begin{bmatrix} p_{11}l_1 - a_1 - d_1 & p_{12}l_2 + a_2 \\ p_{21}l_1 + a_1 & p_{22}l_2 - a_2 - d_2 \end{bmatrix}.
$$

Then it is easy to see that (A1)–(A7) are satisfied if $s(M_4) < 0$.

Let $E_0 = (S_1^*, S_2^*, 0, 0)$ be the disease free equilibrium of (4.1). Then S_1^* and S_2^* satisfy the following linear system:

$$
\begin{aligned} &p_{11}(k_1 + l_1 S_1^*) + p_{12}(k_2 + l_2 S_2^*) - (d_1 + a_1)S_1^* + a_2 S_2^* = 0, \\ &p_{21}(k_1 + l_1 S_1^*) + p_{22}(k_2 + l_2 S_2^*) - (d_2 + a_2)S_2^* + a_1 S_1^* = 0. \end{aligned}
$$

Set

$$
\begin{aligned} h_1 &= \beta_1 S_1^* + \beta_2 S_2^* - b_1 - b_2 - \gamma_1 - \gamma_2 - d_1 - d_2, \\ h_2 &= \beta_1 S_1^* \beta_2 S_2^* - b_1 b_2 - \beta_1 S_1^*(d_2 + \gamma_2 + b_2) \\ &\quad - \beta_2 S_2^*(d_1 + \gamma_1 + b_1) + (d_1 + \gamma_1 + b_1)(d_2 + \gamma_2 + b_2). \end{aligned} \tag{4.2}
$$

It is easy to see that

$$
s(M_1) = \frac{h_1 + \sqrt{h_1^2 - 4h_2}}{2}. \tag{4.3}
$$

Then $\mathcal{R}_0 > 1$ if and only if $h_1 \geq 0$ or

$$\begin{cases} h_1 < 0, \\ 1 + \dfrac{\beta_1 S_1^* \beta_2 S_2^* - b_1 b_2}{(d_1 + \gamma_1 + b_1)(d_2 + \gamma_2 + b_2)} < \dfrac{\beta_1 S_1^*}{d_1 + \gamma_1 + b_1} + \dfrac{\beta_2 S_2^*}{d_2 + \gamma_2 + b_2}. \end{cases} \tag{4.4}$$

Now, we illustrate that the dispersal of juvenile individuals has a significant effect on the spread of the disease, although they are immune to the disease and cannot transmit the disease.

Example 4.1. We fix $k_1 = k_2 = k, l_1 = l_2 = l, d_1 = d_2 = d, \mu_1 = \mu_2 = \mu, a_1 = a_2 = a, b_1 = b_2 = b, \gamma_1 = \gamma_2 = \gamma$. That is, we suppose that the population in the first patch has the same birth rate, death rate, dispersal rate for adults, and recovery rate as the population in the second patch. We let the contact rates β_i and the dispersal rates for juveniles vary in two patches. Let the migration rate of juveniles from the first patch to the second be c_1 and the rate from the second to the first be c_2. Thus we have

$$C_J = \begin{bmatrix} -\mu - c_1 & c_2 \\ c_1 & -\mu - c_2 \end{bmatrix}.$$

It is easy to obtain

$$\exp(C_J r) = \begin{bmatrix} \dfrac{c_2\, p + c_1\, q}{c_1 + c_2} & \dfrac{c_2\,(-q + p)}{c_1 + c_2} \\ \dfrac{c_1\,(-q + p)}{c_1 + c_2} & \dfrac{p\, c_1 + c_2\, q}{c_1 + c_2} \end{bmatrix},$$

with $p = e^{-\mu r}$ and $q = \exp(-r(\mu + c_1 + c_2))$. Note that

$$R_1(t) = \frac{k(c_1\, q + 2\, c_2\, p - c_2\, q) + l\,(c_1\, q + c_2\, p)A_1(t - r) + l\,(c_2\, p - c_2\, q)A_2(t - r)}{c_1 + c_2},$$
$$R_2(t) = \frac{k(2\, c_1\, p - c_1\, q + c_2\, q) - l\,(c_1\, q - c_1\, p)A_1(t - r) + l\,(c_2\, q + c_1\, p)A_2(t - r)}{c_1 + c_2}.$$

By direct calculations, we obtain

$$S_1^* = -\frac{(2\, c_1\, a\, p + 2\, a\, c_2\, p + c_1\, d\, q - c_1\, l\, p\, q + 2\, d\, c_2\, p - c_2\, l\, p\, q - c_2\, d\, q)\, k}{(c_1 + c_2)\,(p\, l - d)\,(-q\, l + 2\, a + d)},$$
$$S_2^* = -\frac{k\,(-c_1\, l\, p\, q + 2\, c_1\, a\, p - c_1\, d\, q + 2\, c_1\, d\, p - c_2\, l\, p\, q + 2\, a\, c_2\, p + c_2\, d\, q)}{(c_1 + c_2)\,(p\, l - d)\,(-q\, l + 2\, a + d)}.$$

Then we can use Theorem 3.1, Corollary 3.1, and Theorem 3.2 to discuss the effect of population dispersal. In order to be concise and clear, we further fix $a = b = 0.2$, $\mu = 2$, $r = 1$, $d = 0.5$, $l = 1$, $k = 1$, and $\gamma = 0$. Then, by the method of estimation, it is not hard to see that $s(M_4) < 0$. Hence (A1)–(A7) are satisfied. In this case, it is clear that $s(\overline{M}(K)) < 0$ for all large k.

Now, let us vary c_1, c_2, β_1, and β_2 to see the effect of the dispersal of juvenile members. First, we fix $\beta_1 = 1.4$ and $\beta_2 = 1$. If the dispersal of juvenile individuals is turned off, then

$$S_1^* = -\frac{p\, k}{p\, l - d} = 0.3711, \quad S_2^* = -\frac{p\, k}{p\, l - d} = 0.3711.$$

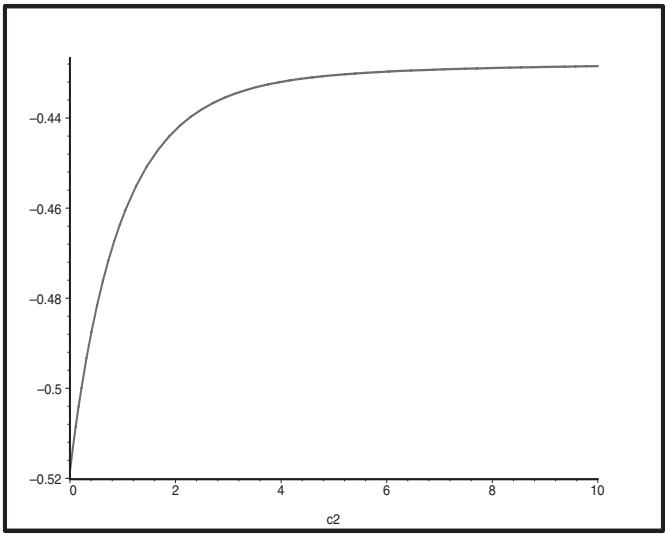


FIG. 1. *The graph of function* h_1 *when* $a = b = 0.2$, $\mu = 2$, $r = 1$, $d = 0.5$, $l = 1$, $k = 1$, $\beta_1 = 1.4$, $\beta_2 = 1$, $c_1 = 0.1$, *and* $\gamma = 0$.

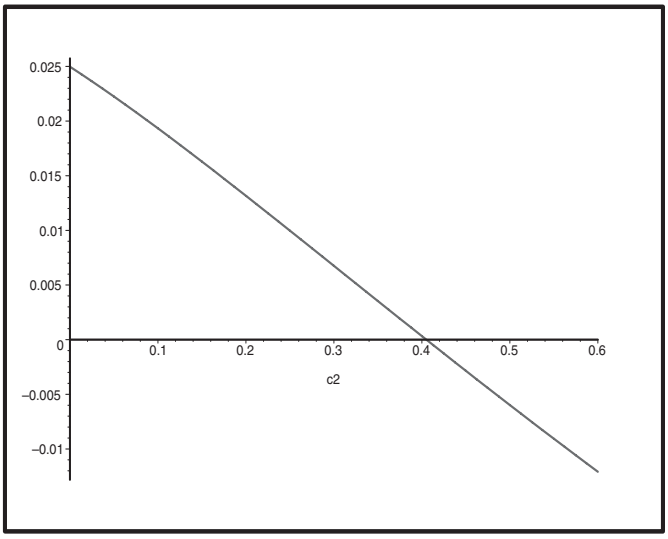


FIG. 2. *The graph of function* h_2 *when* $a = b = 0.2$, $\mu = 2$, $r = 1$, $d = 0.5$, $l = 1$, $k = 1$, $\beta_1 = 1.4$, $\beta_2 = 1$, $c_1 = 0.1$, *and* $\gamma = 0$.

As a consequence, $h_1 = -0.5093 < 0$, $h_2 = 0.0193 > 0$. Hence, the disease dies out in the patches if there is no dispersal for juvenile individuals. If we increase c_1 and c_2 from 0, by numerical calculations, we see that h_1 is always negative (see Figure 1). However, h_2 changes the sign as c_2 increases (see Figure 2). For the case in Figure 2, we see that the disease remains extinct in the two patches when the dispersal coefficient c_2 is weak, and breaks out in the two patches when c_2 is strong. In this case, the migration of juveniles from the second patch to the first patch could intensify the disease spread. Secondly, we fix $\beta_1 = 1.4, \beta_2 = 1.25$, and $c_1 = 0.6$. Then h_1 remains negative, and the graph of h_2 is given in Figure 3. From this figure and numerical calculations, we

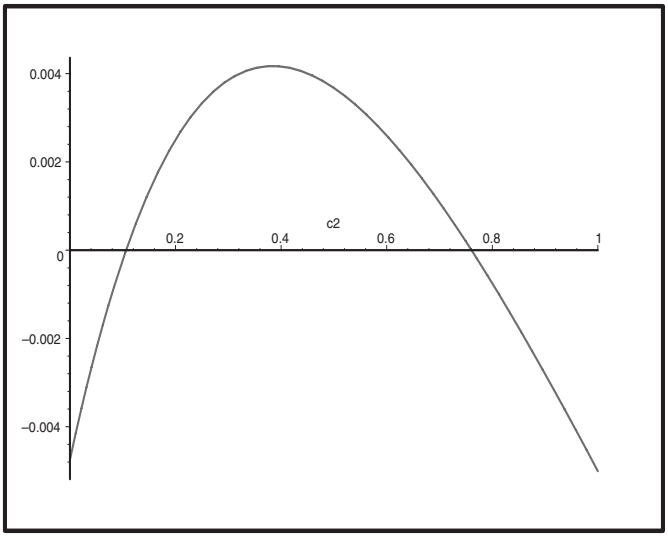


FIG. 3. *The graph of function* h_2 *when* $a = b = 0.2$, $\mu = 2$, $r = 1$, $d = 0.5$, $l = 1$, $k = 1$, $\beta_1 = 1.4$, $\beta_2 = 1.25$, $c_1 = 0.6$, *and* $\gamma = 0$.

see that the disease spreads in the two patches when $0 < c_2 < 0.1064$ or $0.7607 < c_2$, and dies out when $0.1064 < c_2 < 0.7607$. Thus, the disease undergoes a transition from uniform persistence to extinction and a second transition from extinction to uniform persistence; i.e., there exist persistence-extinction-persistence switches, as the dispersal rate c_2 increases from 0.

Let us now change the parameters to $a = b = 0.2$, $\mu = 2$, $r = 1$, $d = 0.5$, $l = 1$, $k = 1$, $\beta_1 = 2$, and $\beta_2 = 0.8$. In this case, if the dispersal of juvenile individuals is turned off, then $h_1 = -0.3609 < 0, h_2 = -0.0570 < 0$. Hence, Theorem 3.2 means that the disease spreads in two patches when there is no dispersal for juveniles. But if we take $c_2 = 0.01$, by means of Maple, we see that h_1 is always negative when c_1 varies. Further, the profile of h_2 is given in Figure 4. Numerical calculation shows that $h_2 = 0$ at $c_1 = 0.4484$. Hence, by Theorems 3.1 and 3.2, the disease spreads in two patches if $0 < c_1 < 0.4484$ and dies out in two patches if $c_1 > 0.4484$. This means that the dispersal of juveniles can also lower the risk of a disease outbreak.

Example 4.1 indicates the existence of the persistence-extinction-persistence switches where the migration rates of juveniles are independent of the migration rates of adults. However, the mobility of juveniles may be associated with that of their adults. To find functions that accurately describe patterns of juvenile movement and adult movement, a good way is to construct a submodel by considering factors such as behaviors of the population and resource differences among patches. For simplicity, we assume that juveniles and susceptible adults move together at the same migration rate. This is a very simple way to relate the movements of juveniles and adults. Our next example shows that persistence-extinction-persistence switches can also occur in this special case.

Example 4.2. We fix $\mu_1 = \mu_2 = 2$, $d_1 = d_2 = 0.5$, $l_1 = 1$, $l_2 = 0.5$, $r = 1$, $k_1 = 1$, $k_2 = 2$, $\beta_1 = 1.4$, $\beta_2 = 0.7$, $b_1 = 0.2$, $b_2 = 0.01$, $\gamma_1 = \gamma_2 = 0$, $a_1 = c_1 = 0.5$, and $a_2 = c_2$. Since $a_i = c_i$, $i = 1, 2$, we have assumed that juveniles and their susceptible adults in each patch move together. By arguments similar to those in Example 4.1, we see that $h_1 < 0$ if $0 \le a_2 \le 0.9$, $h_2 < 0$ if $0 \le a_2 < 0.1589$ or $0.4891 < a_2 \le 0.9$, and

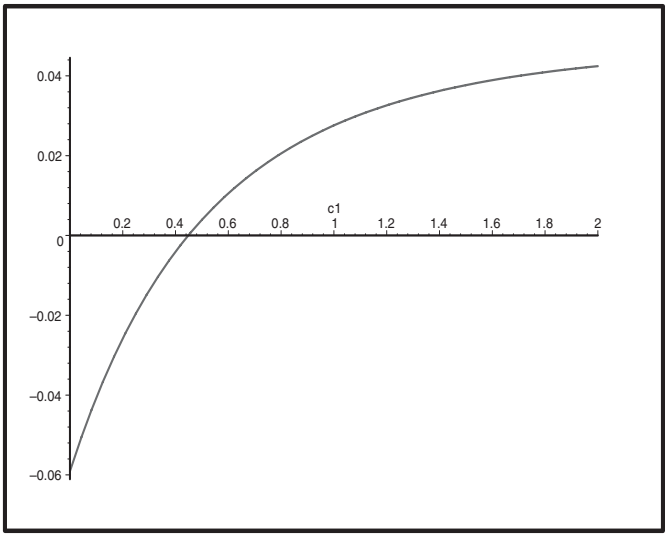


FIG. 4. *The graph of function* h_2 *when* $a = b = 0.2$, $\mu = 2$, $r = 1$, $d = 0.5$, $l = 1$, $k = 1$, $\beta_1 = 2$, $\beta_2 = 0.8$, $c_2 = 0.01$, *and* $\gamma = 0$.

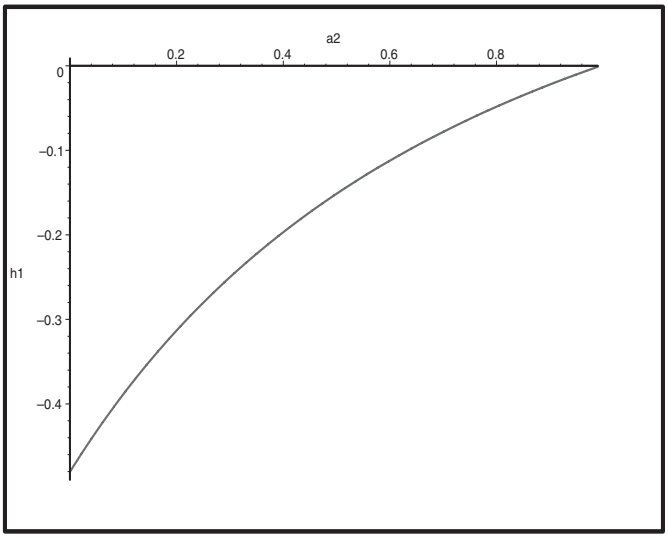


FIG. 5. h_1 *is negative as* a_2 *varies from* 0 *to* 0.9.

$h_2 > 0$ if $0.1589 < a_2 < 0.4891$ (see Figures 5 and 6). Hence, the disease is uniformly persistent in two patches when $0 \le a_2 < 0.1589$ or $0.4891 < a_2 \le 0.9$, and dies out when $0.1589 < a_2 < 0.4891$. This shows that the persistence-extinction-persistence switching phenomenon happens as the dispersal rate a_2 increases from 0.

5. Discussion. Spatial heterogeneity plays an important role in the persistence and dynamics of epidemics. Papers [4, 8, 10, 11] have discussed the effect of immigration of infectious individuals or age structure on the persistence of diseases from different point of views. In [19] we have shown that spatial heterogeneity can increase

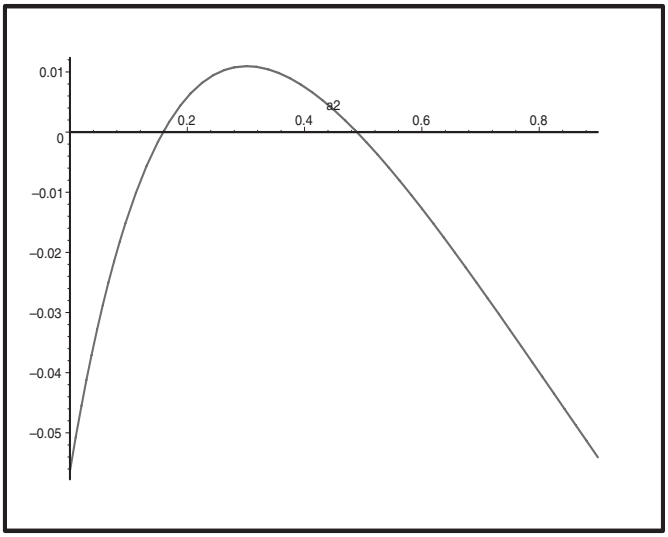


FIG. 6. *h_2 transits from negative value to positive value and then has a second transit from positive to negative as a_2 varies from* 0 *to* 0.9.

the persistence of a disease, and can also reduce the persistence of a disease in the presence of suitable parameters. In this paper, we have proposed an epidemic model with population dispersal among n patches, which may simulate n cities. This is clearly applicable to real situations because data for many diseases are available on a city-by-city scale. We have divided a population into two stages, juvenile stage and adult stage, and assumed that the disease spreads only in adult population. (Sexual diseases are typical examples for this.) The novelty of this paper is that we have incorporated the age structure into the model and studied the effect of the composition of spatial heterogeneity and age heterogeneity. In Theorem 2.1, we have obtained sufficient conditions under which the model admits a unique disease free equilibrium. Then we have defined the reproduction number of the model with many patches according to the ideas in [7, 18]. In Theorems 3.1 and 3.2, we have given sufficient conditions for global extinction and uniform persistence of the disease in terms of the reproduction number. Then we applied our general results to a case in which the birth rates are in the form of (B3) and there are only two patches. We have found that the dispersal of juveniles, although they are immune and cannot transmit the disease, has significant effect on the disease spread. Specifically, in Example 4.1 we have shown that the disease can spread to all patches by increasing the dispersal rate of juveniles in one patch, although the disease dies out in any patch in the absence of the dispersal of juveniles; we have also shown that the disease can die out in all patches by increasing the dispersal rate of juveniles in one patch, although the disease spreads in each patch in the absence of the dispersal of juveniles. Further, we have found the new phenomenon that the disease could undergo persistence-extinction-persistence switches in the sense that the disease admits a transition from uniform persistence to extinction and a second transition from extinction to uniform persistence as the dispersal rate c_2 increases. Example 4.2 shows that this switching phenomenon also happens when juveniles move together with susceptible adults.

The general behavior of the model with more than two patches or with birth rates of other forms is not clear at present. It should be more reasonable to consider

many stages or varying lengths of juvenile stages in different patches. We leave this as future work.

Acknowledgments. We are very grateful to two anonymous referees for careful reading and valuable comments which led to an improvement of our original manuscript.

REFERENCES

[1] J. Arino and P. van den Driessche, *A multi-city epidemic model*, Math. Popul. Stud., 10 (2003), pp. 175–193.

[2] J. Arino and P. van den Driessche, *The basic reproduction number in a multi-city compartmental epidemic model*, in Proceedings of the Positive Systems Conference (Rome, 2003), Lecture Notes in Control and Inform. Sci. 294, Springer, Berlin, 2003, pp. 135–142.

[3] R. M. Bolker and B. T. Grenfell, *Space, persistence, and dynamics of measles epidemics*, Phil. Trans. Roy. Soc. Lond. Ser. B, 348 (1995), pp. 309–320.

[4] F. Brauer and P. van den Driessche, *Models for transmission of disease with immigration of infectives*, Math. Biosci., 171 (2001), pp. 143–154.

[5] A. Cliff, P. Haggett, and M. Smallman-Raynor, *Measles: An Historical Geography of a Major Human Viral Disease*, Blackwell, Oxford, UK, 1993.

[6] K. Cooke, P. van den Driessche, and X. Zou, *Interaction of maturation delay and nonlinear birth in population and epidemic models*, J. Math. Biol., 39 (1999), pp. 332–352.

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

[8] Z. Feng and H. R. Thieme, *Endemic models with arbitrarily distributed periods of infection* II: *Fast disease dynamics and permanent recovery*, SIAM Appl. Math., 61 (2000), pp. 983–1012.

[9] W. M. Hirsch, H. L. Smith, and X.-Q. Zhao, *Chain transitivity, attractivity, and strong repellers for semidynamical systems*, J. Dynam. Differential Equations, 13 (2001), pp. 107–131.

[10] M. J. Keeling, *Metapopulation momnets: Coupling, stochasticity and persistence*, J. Animal Ecology, 69 (2000), pp. 725–736.

[11] A. L. Lloyd and R. M. May, *Spatial heterogeneity in epidemic models*, J. Theoret. Biol., 179 (1996), pp. 1–11.

[12] H. L. Smith, *Monotone Dynamical Systems. An Introduction to the Theory of Competitive and Cooperative Systems*, Math Surveys and Monographs 41, American Mathematical Society, Providence, RI, 1995.

[13] H. L. Smith and H. R. Thieme, *Strongly order preserving semiflows generated by functional differential equations*, J. Differential Equations, 93 (1991), pp. 332–363.

[14] H. L. Smith and P. Waltman, *The Theory of the Chemostat*, Cambridge University Press, Cambridge, UK, 1995.

[15] H. L. Smith and X.-Q. Zhao, *Microbial growth in a plug flow reactor with wall adherence and cell motility*, J. Math. Anal. Appl., 241 (2000), pp. 134–155.

[16] J. W.-H. So, J. Wu, and X. Zou, *Structured population on two patches: Modeling dispersal and delay*, J. Math. Biol., 43 (2001), pp. 37–51.

[17] H. R. Thieme, *Persistence under relaxed point-dissipativity (with application to an endemic model)*, SIAM J. Math. Anal., 24 (1993), pp. 407–435.

[18] P. van den Driessche and J. Watmough, *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci., 180 (2002), pp. 29–48.

[19] W. Wang and X.-Q. Zhao, *An epidemic model in a patchy environment*, Math. Biosci., 190 (2004), pp. 97–112.

[20] X.-Q. Zhao, *Uniform persistence and periodic coexistence states in infinite-dimensional periodic semiflows with applications*, Canadian Appl. Math. Quart., 3 (1995), pp. 473–495.

[21] X.-Q. Zhao and Z.-J. Jing, *Global asymptotic behavior in some cooperative systems of functional differential equations*, Canadian Appl. Math. Quart., 4 (1996), pp. 421–444.

[22] X.-Q. Zhao and X. Zou, *Threshold dynamics in a delayed SIS epidemic model*, J. Math. Anal. Appl., 257 (2001), pp. 282–291.

SIAM J. APPL. MATH.
Vol. 65, No. 5, pp. 1615–1632

A MAXIMUM PRINCIPLE FOR BELTRAMI COLOR FLOW*

LORINA DASCAL† AND NIR A. SOCHEN†

Abstract. We study, in this work, the maximum principle for the Beltrami color flow and the stability of the flow's numerical approximation by finite difference schemes. We discuss, in the continuous case, the theoretical properties of this system and prove the maximum principle in the strong and the weak formulations. In the discrete case, all the second order explicit schemes that are currently used violate, in general, the maximum principle. For these schemes we give a theoretical stability proof, accompanied by several numerical examples.

Key words. maximum principle, Beltrami framework, parabolic PDEs, finite difference schemes

AMS subject classifications. 35K40, 35B50

DOI. 10.1137/S0036139903430835

1. Introduction.

The maximum principle in its various forms is a powerful and instrumental tool for establishing results concerning existence, uniqueness, and other qualitative properties of linear and nonlinear partial differential equations (PDEs). A complete overview of this subject through 1967 can be found in [14].

We are interested in the property of the maximum principle for Beltrami color flow in the context of scale-space theory. This theory claims that significant information exists in all levels of resolution/scale of the image. It is important to create a simplification process, called a "scale-space," from which the information can be extracted.

The notion of causality in the context of image processing and especially in scale-space and denoising arenas was put forward in the work of Koenderink [10]. In the one-dimensional case it is desirable that the simplification process of the signal not create new maxima. This demand, together with homogeneity, leads to filtering with a Gaussian kernel. The convolution of this kernel with the initial image is the solution of the linear diffusion equation with the initial image as initial condition. For higher-dimensional signals the noncreation of new maxima cannot be achieved. It is usually replaced by a new principle—the noncreation of new level sets. This new principle is called in the scale-space literature "the causality principle." It is directly related, in the scalar case, to the extremum principle as observed by Hummel [7]. The extremum principle is taken as the natural generalization of the causality principle to the vectorial case.

Moreover, the relevance of investigating the maximum principle for the Beltrami flow and other PDE-based models can be seen through the work of Alvarez et al. [1]. In this paper, the authors propose a rigorous connection between scale-space analysis and PDEs. They start from a very natural set of filtering axioms and show that the resulting filtered image must necessarily be the solution of a second order fully nonlinear parabolic PDE. The maximum principle is one of their axioms, which is imposed so that smoothing of the original image is made with no enhancement.

*Received by the editors July 2, 2003; accepted for publication (in revised form) August 11, 2004; published electronically May 12, 2005. This research was supported in part by the Israel Academy of Science, the Tel-Aviv University fund, the Adams Super-Center for brain research, and the Israeli Ministry of Science.

http://www.siam.org/journals/siap/65-5/43083.html

†Department of Applied Mathematics, Tel Aviv University, Ramat-Aviv, Tel-Aviv 69978, Israel (lorina@math.tau.ac.il, sochen@math.tau.ac.il).

The problem of the discrete maximum principle was studied in many works. This issue is important as it ensures that intensity values in the evolving image are constrained by the initial image values and do not grow without bounds. Perona and Malik [12] proposed a numerical scheme which satisfies this property as proved in [22]. Also the PDE introduced by Catte et al. [3] was proved to satisfy the discrete maximum principle [22].

In this paper we treat the Beltrami flow for color images and study two aspects of the maximum principle (continuous and discrete). First we deal with the continuous formulation of the maximum principle and prove it for both the strong and the weak formulations. The motivation for considering these two formulations is threefold. First, the strong formulation is presented here in order to check the validity of the maximum principle for a smooth solution. Second, this property is generalized for a class of nonregular functions via a weak formulation of the maximum principle. In what concerns the weak formulation, we follow the duality approach of Florack [6] and later Mumford and Gidas [11]. In these works the image is conceived as a generalized function. The duality approach describes the sensor space (also called "device space") as a functional space. The data we usually process, which result from the interaction of the physical/optical data and the sensor, are modeled as an inner product of the sensor function and the "true image." In this context, the set of images is equivalent to the set of linear functionals on the sensor functional space. It is natural from this point of view to study the flow equations on the image space directly. We are able to do so by defining generalized (weak) solutions to our flow equations. Our third motivation lies in the fact that to achieve a proper analysis of images, we must consider functions with less regular structure than the smooth functions we dealt with in the strong formulation. This again leads to the study of weak solutions.

To the best of our knowledge, for highly nonlinear and strongly coupled systems like the one we describe here, no mathematical analysis has been performed even for smooth functions. In previous works [4, 2], well-posedness and the maximum principle were treated for scalar valued functions only and for initial data of Lipschitz type, which excludes discontinuous functions.

In the last part of the paper we study the discrete maximum principle for a certain explicit difference scheme by which the nonlinear differential equation is approximated. We show that to approximate the various derivatives to a given order is not enough to guarantee the maximum principle. This scheme can *violate* the maximum principle. We present, however, a proof of the stability of this scheme along with examples that clearly demonstrate stability while failing to obey the maximum principle.

The paper is organized as follows. In section 2 we review the Beltrami framework. In section 3 we deal with the continuous formulation of the maximum principle. We prove the extremum principle for the strong solution of the parabolic quasi-linear system that characterizes the Beltrami color flow. In section 4 we introduce a weak (generalized) solution for this system and prove the extremum principle in a weak formulation. In section 5 we discuss the properties of the second order central difference scheme, which in general violates the maximum principle. For this scheme we give a theoretical stability proof. In section 6 we present numerical results. We summarize and conclude in section 7.

2. The Beltrami framework. Let us briefly review the Beltrami framework for nonlinear diffusion in computer vision [8, 18, 19].

The space of interest in computer vision such as images, texture, disparity in

stereo-vision, optical flow, and distortion in registration is represented as a *fiber bundle*. The base manifold is the image domain. We consider in this work a flat and compact domain. A nonflat domain is treated, for example, in [17]. The feature space, be it gray-values, color, optical flow, texture, etc., is the fiber space. Any particular image or vector field is a *section* of this fiber bundle. We assume that the Riemannian structure can be defined for the base manifold and for the fiber bundle. Thus, we represent an image and other local features as an embedding map of a Riemannian manifold in a higher-dimensional space—the fiber bundle. The simplest example is a gray-level image which is represented as a two-dimensional surface embedded in $\mathbb{R}^3$. We denote the map by $X : \Sigma \to \mathbb{R}^3$, where Σ is a two-dimensional surface, and we denote the local coordinates on it by (σ^1, σ^2). The map U is given in general by $(U^1(\sigma^1, \sigma^2), U^2(\sigma^1, \sigma^2), U^3(\sigma^1, \sigma^2))$. In our example we represent the map U as follows: $(U^1 = \sigma^1, U^2 = \sigma^2, U^3 = I(\sigma^1, \sigma^2))$. We choose on this surface a Riemannian structure, namely a metric. The metric is a positive definite and symmetric 2-tensor that may be defined through the local distance measurements:

$$ds^2 = g_{11}(d\sigma^1)^2 + 2g_{12}d\sigma^1 d\sigma^2 + g_{22}(d\sigma^2)^2.$$

The canonical choice of coordinates in image processing is Cartesian. For such a choice, which we follow in the rest of the paper, we identify $\sigma^1 = x^1$ and $\sigma^2 = x^2$. We use below the Einstein summation convention in which a pair of upper and lower identical indices is summed over. With this convention, the above equation is written as $ds^2 = g_{ij}dx^i dx^j$. We denote the elements of the inverse of the metric by superscripts $g^{ij} = (g^{-1})_{ij}$, and the determinant by $g = \det(g_{ij})$.

Once the image is defined as an embedding mapping of Riemannian manifolds, it is natural to look for a measure on this space of embedding maps.

2.1. Polyakov action: A measure on the space of embedding maps. Denote by (Σ, g) the image manifold and its metric, and by (M, h) the space-feature manifold and its metric. Then the functional $S[U]$ attaches a real number to a map $U : \Sigma \to M$:

$$S[U^a, g_{ij}, h_{ab}] = \int dV ||dU||^2_{g,h},$$

where dV is a volume element that is expressed in a local coordinate system as $dV = \sqrt{g}dxdy$. The integrand $||dU||^2_{g,h}$ is the Riemannian Frobenius norm of the tangent map. It is expressed in a local coordinate system by $||dU||^2_{g,h} = (\partial_{x_i}U^a)g^{ij}(\partial_{x_j}U^b)h_{ab}$. This functional, for $m = 2$ and $h_{ab} = \delta_{ab}$, was first proposed by Polyakov [13] in the context of high energy physics, and the theory is known as *string theory*.

Let us formulate the Polyakov action in matrix form: (Σ, G) is the image manifold and its metric as before. Similarly, (M, H) is the spatial-feature manifold and its metric. Define

$$A^{ab} = (\vec{\nabla}U^a)^t G^{-1}\vec{\nabla}U^b.$$

The map $U : \Sigma \to M$ has a weight

$$S[U, G, H] = \int d^m\sigma\sqrt{g}\ \mathrm{Tr}(AH),$$

where m is the dimension of Σ and $g = \det(G)$.

Using standard methods in the calculus of variations, the Euler–Lagrange equations with respect to the embedding (assuming a Euclidean embedding space) are (see [18] for explicit derivation)

$$0 = -\frac{1}{2\sqrt{g}} h^{ab} \frac{\delta S}{\delta U^b} = \frac{1}{\sqrt{g}} \partial_{x_i} (\sqrt{g} g^{ij} \partial_{x_j} U^a), \tag{2.1}$$

or in matricial form,

$$0 = -\frac{1}{2\sqrt{g}} h^{ab} \frac{\delta S}{\delta U^b} = \underbrace{\frac{1}{\sqrt{g}} \operatorname{div} (D \nabla U^a)}_{\Delta_g U^a} \tag{2.2}$$

(where the matrix $D = (d^{ij})_{i,j=1,2} = \sqrt{g} G^{-1}$). The extension for non-Euclidean embedding space is treated in [9, 19, 20, 21]. The elements of the induced metric for color images with Cartesian color coordinates are

$$g_{ij} = \delta_{ij} + \beta^2 \sum_{a=1}^{3} U^a_{x_i} U^a_{x_j}, \tag{2.3}$$

where $\beta > 0$ is the ratio between the spatial and color distances, and the subscript of U denotes partial derivation. Note that this metric is different from the Di Zenzo matrix [24] (which is not a metric since it is not positive definite). A generalization of Di Zenzo's gradient for color images was investigated in [23] by constructing an anisotropic vector-valued diffusion model with a common tensor-valued structure descriptor.

The value of the parameter β, present in the elements of the metric g_{ij}, is very important and determines the nature of the flow. In the limit $\beta \to 0$, for example, the flow degenerates to the decoupled channel by channel linear diffusion flow. In the other limit $\beta \to \infty$ we get a new nonlinear flow. The gray-value analogue of this limit is the total variation (TV) flow of [15] (see details in [19]).

Since (g_{ij}) is positive definite, $g \equiv \det(g_{ij}) > 0$ for all σ^i. This factor is the simplest one that does not change the minimization solution while giving a reparameterization invariant expression. The operator that acts on U^a is the natural generalization of the Laplacian from flat spaces to manifolds and is called the Laplace–Beltrami operator, denoted by Δ_g.

The nonlinear diffusion or scale-space equation emerges via the gradient descent minimization:

$$U^a_t = \frac{\partial}{\partial t} U^a = -\frac{1}{2\sqrt{g}} h^{ab} \frac{\delta S}{\delta U^b} = \Delta_g U^a. \tag{2.4}$$

The mathematical properties of this system, together with the initial and boundary conditions which will be detailed below, are studied in the rest of the paper with an emphasis on the extremum principle.

3. The extremum principle in the strong formulation. Here we establish the maximum principle for the strong solution of the initial boundary-value problem which characterizes the Beltrami color flow. We refer to the term "strong solutions" when we talk about solutions which are functions with some smoothness criteria that we detail below. Let us first introduce some notation. We denote the image domain

by Ω. It is a bounded open domain in $\mathbb{R}^2$. We denote by $\partial\Omega$ the boundary of Ω. We define the space-time cylinder $Q_T = \Omega \times (0,T)$, and denote its lateral surface by $S_T = \{(x,t)|x \in \partial\Omega, t \in (0,T)\}$. We also define the parabolic boundary by the union of the bottom and the lateral boundaries of the cylinder $\Gamma_T = \Omega \bigcup S_T$.

The PDE is the gradient descent equation for the Polyakov action, as was described in the previous section. We rearrange the equation by explicitly carrying out the calculation of the derivation operator *div*. The result is the sum of two terms: The first term results from applying the *div* to $\sqrt{g}G^{-1}$, and the second from applying the *div* to the gradient's components $U^a_{x_i}$. Remember that the metric, and consequently its inverse and its determinant, depends on first order derivatives. Applying the *div* operator to it gives rise to second order derivatives of the different channels as well. Rearranging the right-hand side of (2.4) according to the second order derivatives, and the coefficients thereof, we arrive at the following coupled system of PDEs:

$$U^a_t = (F^a_{\ b})^{ij} U^b_{x_i x_j}, \quad (x,t) \in Q_T, \tag{3.1}$$

where $a,b = 1,2,3$ are indices in color space, $i,j = 1,2$ are spatial indices, and summation is applied to all repeated indices. Note that $(F^a_{\ b})$ are nine 2×2 matrices. Denote by $H^a = (U^a_{x_i x_j})_{ij=1}^2$ the Hessian of U^a. This system of PDEs can be written in terms of a trace in the spatial domain as

$$U^a_t = \text{Trace}\left(F^a_{\ b} H^b\right), \quad (x,t) \in Q_T, \tag{3.2}$$

where, as before, the repeated b index implies a summation over the color indices.

The system of PDEs for which we establish the extremum principle is

$$U^a_t = \Delta_g U^a = \frac{1}{\sqrt{g}} \text{div}(D \nabla U^a), \quad U^a = R,G,B, \tag{3.3}$$

where D is defined as before: $D = \sqrt{g}G^{-1}$.

The initial and boundary conditions are

$$U^a(x,0) = U^a_0(x), \quad x \in \Omega, \tag{3.4}$$

$$D\vec{\nabla}U^a \cdot \vec{n}\Big|_{S_T} = 0, \tag{3.5}$$

where $\vec{n}$ is the outer normal to $\partial\Omega$ and the dot product denotes, as usual, the Euclidean scalar product in $\mathbb{R}^2$.

LEMMA 3.1. *The nine* 2×2 *matrices* $(F^a_{\ b})$ *are symmetric, positive definite, and their elements* $(F^a_{\ b})^{ij}$ *are rational functions of the first derivatives of the different channels. These matrix elements are, moreover,* uniformly bounded functions *on* Q_T.

Proof. The proof is by direct calculation. One finds, for example,

$$\left(F^2_{\ 1}\right)^{11} = -R_x G_x \frac{g_{22}^2}{g^2} + (R_x G_y + R_y G_x)\frac{g_{12}g_{22}}{g^2} - \frac{R_y G_y}{g}\left(1 + \frac{g_{12}^2}{g}\right), \tag{3.6}$$

$$\left(F^2_{\ 1}\right)^{12} = \left(F^2_{\ 1}\right)^{21} = \frac{R_x G_y + R_y G_x}{g} - \frac{R_x G_y + R_y G_x}{g^2} g_{11} g_{22} \tag{3.7}$$

$$- \frac{R_x G_y + R_y G_x}{g^2} g_{12}^2 + 2\frac{R_x G_x g_{22} + R_y G_y g_{11}}{g^2} g_{12}^2, \tag{3.8}$$

$$\left(F^2_{\ 3}\right)^{22} = -R_y G_y \frac{g_{11}^2}{g^2} + (R_x G_y + R_y G_x)\frac{g_{11}g_{12}}{g^2} - \frac{R_x G_x}{g}\left(1 + \frac{g_{12}^2}{g}\right) \tag{3.9}$$

(here we have denoted by R, G, B the three components of the color vector $\overrightarrow{U}$).

These are rational functions of the first derivatives. The diagonal elements of (F^a_b) are strictly positive (by a direct check), and the negativity of the discriminant implies the positive definiteness of these matrices. One can verify directly that the coefficients are bounded functions of the first derivatives. These properties are verified along the same lines for all matrices. □

Next we state the maximum principle for the strong solutions of the coupled system of PDEs (3.3), with initial data (3.4) and boundary condition (3.5).

THEOREM 3.2. *Let $\overrightarrow{U_0} \in C^2(\Omega)$. Then a solution $\overrightarrow{U} \in C^{2,1}(\bar{Q}_T)$ satisfies the following maximum principle:*

$$(1) \quad \max_{\bar{Q}_T} \sum_{a=1}^{3} U^a = \max_{\Omega} \sum_{a=1}^{3} U_0^a, \tag{3.10}$$

$$(2) \quad \max_{\bar{Q}_T} U^a = \max_{\Omega} U_0^a. \tag{3.11}$$

Proof. Note that assertion (2) does not imply, in general, assertion (1). We start by proving assertion (1). Consider the following system of inequalities:

$$V_t^a < (F^a_b)^{ij} V^b_{x_i x_j}, \quad (x,t) \in Q_T, \tag{3.12}$$

where $F^a_b = F^a_b(\nabla \vec{V})$.

We now show that a smooth solution of this system of inequalities satisfies

$$\max_{\bar{Q}_T} \sum_{a=1}^{3} V^a = \max_{\Gamma_T} \sum_{a=1}^{3} V^a. \tag{3.13}$$

Let $\bar{V} = \sum_{a=1}^{3} V^a$, and suppose on the contrary that the maximum of $\bar{V}$ is attained at an interior point $(x_0, t_0) \in \bar{Q}_T \backslash \Gamma_T$. This assumption leads to a contradiction as follows: The maximality at the point (x_0, t_0) implies

$$\bar{V}_t|_{(x_0,t_0)} \geq 0 \quad (\bar{V}_t|_{(x_0,t_0)} = 0 \quad \text{if} \quad 0 \leq t_0 < T). \tag{3.14}$$

Based on (2.1), the system of inequalities (3.12) is equivalent to

$$V_t^a < \frac{1}{\sqrt{g}} \partial_{x_i} (g^{ij} \sqrt{g} \partial_{x_j} V^a), \quad (x,t) \in Q_T. \tag{3.15}$$

Since the Laplace–Beltrami operator that acts on the components depends only on the geometry, the sum of the components obeys the same inequality:

$$\bar{V}_t < \frac{1}{\sqrt{g}} \partial_{x_i} (g^{ij} \sqrt{g} \partial_{x_j} \bar{V}). \tag{3.16}$$

On the other hand, carrying out the *div* computation explicitly, we rewrite this inequality as

$$\frac{1}{\sqrt{g}} \partial_{x_i} (g^{ij} \sqrt{g} \partial_{x_j} \bar{V}) = g^{ij} \bar{V}_{x_i x_j} + \omega^j \bar{V}_{x_j}.$$

The functions ω^j depend on the first and second derivatives of each of the components of the color vector. They are bounded on Q_T by the smoothness of $\overrightarrow{V} \in C^{2,1}(\bar{Q}_T)$.

The positive definiteness of the matrix g^{ij} and the maximality at the point (x_0, t_0) imply

$$\frac{1}{\sqrt{g}}\partial_{x_i}(g^{ij}\sqrt{g}\partial_{x_j}\bar{V})|_{(x_0,t_0)} = g^{ij}\bar{V}_{x_ix_j}|_{(x_0,t_0)} < 0. \tag{3.17}$$

Clearly (3.14) and (3.17) contradict (3.16). We conclude that (3.13) holds for a solution of the system of inequalities (3.12).

Using the result for the solution of the system of inequalities, we prove the result concerning the solution $\vec{U}$ of our system (3.3). We define $W^a = U^a - \epsilon t$ and $\bar{W} = \sum_1^3 W^a$, $\bar{U} = \sum_1^3 U^a$. Then $\nabla W^a = \nabla U^a$, $g_{ij}(\vec{W}) = g_{ij}(\vec{U})$ and we obtain the following inequalities:

$$W_t^a - \frac{1}{\sqrt{g}}\partial_{x_i}(g^{ij}\sqrt{g}\partial_{x_j}W^a) = U_t^a - \frac{1}{\sqrt{g}}\partial_{x_i}(g^{ij}\sqrt{g}\partial_{x_j}U^a) - \epsilon < 0. \tag{3.18}$$

Since $\vec{W} = (W^a)_{a=1,2,3}$ satisfies (3.18), it follows that

$$\max_{\bar{Q}_T}\bar{W} = \max_{\Gamma_T}\bar{W}.$$

Letting $\epsilon \to 0$, we establish that

$$\max_{\bar{Q}_T}\bar{U} = \max_{\Gamma_T}\bar{U}.$$

Due to the boundary condition (3.5), the maximum cannot be attained on S_T (see [14, pp. 65–67]). Therefore assertion (1) is proved.

Next we prove assertion (2). Observe first that the off-diagonal matrices $F^a_{\ b}$ with $a \neq b$ can be written as $U^a_{x_i}$ times a bounded function. Taking, for example, $(a, b) = (1, 2)$, one finds by rearranging the terms in (3.1) that

$$(F_2^1)^{ij} = U^1_{x_1} \cdot f_1^{ij}(\nabla\overrightarrow{U}) + U^1_{x_2} \cdot f_2^{ij}(\nabla\overrightarrow{U}). \tag{3.19}$$

Thus if $i = 1, j = 1$, for example, then

$$f_1^{11} = -U^2_{x_1}\frac{g_{22}^2}{g^2} + U^2_{x_2}\frac{g_{12}g_{22}}{g^2}, \qquad f_2^{11} = U^2_{x_1}\frac{g_{12}g_{22}}{g^2} + U^2_{x_2}\left(\frac{g_{22}^2}{g^2} + \frac{1}{g}\right).$$

One can write the other off-diagonal matrices similarly. For the structure of the induced metric, we can easily see that

$$g \geq 1, \quad \frac{g_{ij}}{g} \leq 1 \quad \text{for all} \quad i, j = 1, 2. \tag{3.20}$$

Since the solution $\overrightarrow{U}$ is in $C^{2,1}(\bar{Q}_T)$, one can readily establish, using (3.20), that the functions f_1^{ij}, f_2^{ij} are bounded on Q_T. We can, therefore, write the first equation of the system (3.1), (3.4), (3.5) in the following form:

$$U_t^1 = (F_1^1)^{ij}U^1_{x_ix_j} + U^2_{x_ix_j}(U^1_{x_1}f_1^{ij} + U^1_{x_2}f_2^{ij}) + U^3_{x_ix_j}(U^1_{x_1}g_1^{ij} + U^1_{x_2}g_2^{ij}), \tag{3.21}$$

where g_1^{ij}, g_2^{ij} are, as above, bounded functions depending on the first derivatives of the vector solution $\overrightarrow{U}$. We rewrite this equation:

$$U_t^1 = (F_1^1)^{ij}U^1_{x_ix_j} + T^iU^1_{x_i},$$

where T^i are continuous functions on a compact domain and therefore bounded functions. Again using the maximality of the point (x_0, t_0), the positive definiteness of the matrix F_1^1, and reasoning similar to that in the proof of assertion (1), we can conclude that

$$\max_{\bar{Q}_T} U^1 = \max_{\Omega} U_0^1.$$

For the other two components the proof is the same, mutatis mutandis. Thus assertion (2) is proved. □

In the next section we define a weak solution for the Beltrami color flow and prove that it obeys the extremum principle if it exists.

4. Weak formulation of the extremum principle. In this section we define a weak solution of the system (3.3), (3.4), (3.5) under the smoothness assumptions that are detailed below. We further prove the extremum principle for this type of solution.

Let us introduce the following notations. Denote by $V(Q_T)$ the space of functions which belong to $L^2(Q_T)$ and have first weak derivatives satisfying $\nabla u \in L^\infty(Q_T)$, $u_t \in L^\infty(Q_T)$. The Sobolev space $W_r^{p,q}$ is the space of functions for which the L_r norm of their first generalized p spatial derivatives and q time derivatives is finite (below we omit the second superscript for functions on the spatial domain only).

First we define a weak solution as follows.

DEFINITION 4.1. *A weak solution of the system* (3.3)*, with initial and boundary conditions* (3.4), (3.5)*, is a vector function* $\vec{U} \in V(Q_T)$ *such that for any vector function* $\vec{\eta} \in V(Q_T)$ *(i.e., each of the components of the vector are in* $V(Q_T)$*) the following integral identities hold for almost all* $t \in [0, T]$:

$$\int_{Q_T} U_t^a \eta^a \sqrt{g}\, dxdt + \int_{Q_T} g^{ij} U_{x_i}^a \eta_{x_j}^a \sqrt{g}\, dxdt = 0. \tag{4.1}$$

Remark 4.1. The integral $\int_\Omega \sqrt{g}\, dx$ means the area of the two-dimensional manifold embedded in R^5.

Remark 4.2. Florack [6], in viewing an image as a tempered distribution (see [16]), adopted the space of the so-called slow growth functions (smooth functions of rapid decay) as the sensor space. In this paper we take $V(Q_T)$ as the sensor functional space, which we choose in accordance with the weak formulation of our problem.

Next we prove that if a weak solution exists and it satisfies $\nabla(\vec{U}_t) \in L^\infty(Q_T)$, the following weak extremum principle holds.

THEOREM 4.1. *Assume the initial data* $\vec{U_0} \in W_2^1(\Omega)$*. For a weak solution of the system* (3.3), (3.4), (3.5) *such that* $\nabla(\vec{U}_t) \in L^\infty(Q_T)$ *we have for almost all* $(x, t) \in Q_T$

$$\operatorname*{ess\,inf}_{\Omega} U_0^a \le U^a(x,t) \le \operatorname*{ess\,sup}_{\Omega} U_0^a. \tag{4.2}$$

Proof. We prove (4.2) for one of the components. We divide the cylinder Q_T into a finite number of cylinders of equal height $Q_{t_s} = \Omega \times (t_{s-1}, t_s)$, where $t_s = \frac{T}{N} s$ and $s = 1, 2, \ldots, N$. For the cylinder Q_{t_1} we define $k_a = \operatorname{ess\,sup}_\Omega U_0^a(x)$ and $(U^a)^{k_a} = \max\{0, U^a - k_a\}$ for $(x, t) \in \Omega \times (0, t_1)$.

Choose the test function $\eta^1 = R^{k_1}$. Note that by the hypothesis, $\vec{U} \in V(Q_T)$, and therefore the choice of such η is justified. Identity (4.1) for component R is now

$$\int_{Q_{t_1}} R_t R^{k_1} \sqrt{g} dxdt + \int_{Q_{t_1}} d^{ij} R_{x_i} R_{x_j}^{k_1}\, dx\, dt = 0. \tag{4.3}$$

Since the matrix D ($D = \sqrt{g}G^{-1}$) depends on the gradient $\nabla\vec{U}$ and $\nabla\vec{U} \in L^\infty(Q_{t_1})$, we have that D is a uniformly positive definite matrix, and thus there exists a constant $\nu > 0$ such that

$$d^{ij}(\nabla\bar{U})R_{x_i}R^{k_1}_{x_j} \geq \nu|\nabla R^{k_1}|^2 \quad \text{for almost every} \quad (x,t) \in Q_{t_1}.$$

Therefore, (4.3) for the component R becomes

$$\int_{Q_{t_1}} R_t R^{k_1}\sqrt{g}dxdt + \nu\int_{Q_{t_1}} |\nabla R|^2\,dx\,dt \leq 0. \tag{4.4}$$

Since for almost all $t \in (0,t_1)$

$$\int_\Omega R_t(x,t)R^{k_1}(x,t)\,dx = \frac{1}{2}\frac{d}{dt}\int_\Omega \left(R^{k_1}(x,t)\right)^2 dx,$$

we get

$$\int_{Q_{t_1}} R_t R^{k_1}\sqrt{g}\,dx\,dt = \frac{1}{2}\int_{Q_{t_1}} ((R^{k_1})^2)_t\sqrt{g}\,dx\,dt \tag{4.5}$$

$$= \frac{1}{2}\left(\int_\Omega (R^{k_1})^2\sqrt{g}\Big|_0^{t_1} dx - \int_{Q_{t_1}} (R^{k_1})^2(\sqrt{g})_t\,dx\,dt\right), \tag{4.6}$$

and using (4.4), we get

$$\frac{1}{2}\int_\Omega \left(R^{k_1}(x,t)\right)^2\Big|_{t=t_1} dx + \nu\int_{Q_{t_1}} |\nabla R^{k_1}|^2 dxdt \leq \frac{1}{2}\int_{Q_{t_1}} (R^{k_1})^2(\sqrt{g})_t\,dx\,dt + \frac{1}{2}\int_\Omega (R_0^{k_1})^2\sqrt{g_0}\,dx. \tag{4.7}$$

Since $\vec{U_0} \in W_2^1(\Omega)$ and $R^{k_1}(x,0) = R_0^{k_1} = 0$, then $\int_\Omega (R_0^{k_1})^2\sqrt{g_0}\,dx = 0$ and (4.7) becomes

$$\frac{1}{2}\int_\Omega \left(R^{k_1}(x,t)\right)^2 dx + \nu\int_{Q_{t_1}} |\nabla R^{k_1}|^2\,dx\,dt \leq \frac{1}{2}\int_{Q_{t_1}} (R^{k_1})^2(\sqrt{g})_t\,dx\,dt. \tag{4.8}$$

Denote by $||\cdot||_{V(Q_T)}$ the norm on the space $V(Q_T)$, where

$$||u||_{V(Q_T)} = \max_{0\leq t\leq T} ||u||_{L^2(\Omega)} + \sqrt{\int_{Q_T} |\nabla u|^2\,dx\,dt}.$$

By assumption, $\vec{U}_{tx} \in L^\infty(Q_T)$, and then there exists a positive constant C such that $C = \sup_{Q_T}(\sqrt{g})_t$.

The Cauchy–Schwarz inequality leads us to

$$\left|\int_{Q_{t_1}} (R^{k_1})^2(\sqrt{g})_t\,dx\,dt\right| \leq C\int_{Q_{t_1}} (R^{k_1})^2\,dx\,dt$$

$$\leq Ct_1\left(\max_{0\leq t\leq t_1} ||R^{k_1}||_{L^2(\Omega)}\right)^2 \leq Ct_1||R^{k_1}||^2_{V(Q_{t_1})}. \tag{4.9}$$

Therefore (4.8) becomes

$$\min\left\{\frac{1}{2},\nu\right\}||R^{k_1}||^2_{V(Q_{t_1})} \leq Ct_1||R^{k_1}||^2_{V(Q_{t_1})}.$$

Then for sufficiently small t_1 such that

$$Ct_1 < \min\left\{\frac{1}{2},\nu\right\}, \tag{4.10}$$

we obtain $||R^{k_1}||_{V(Q_{t_1})} = 0$, which implies that for almost every $(x,t) \in \Omega \times (0,t_1)$ we have

$$R(x,t) \leq \operatorname{ess\,sup}_{\Omega} R_0.$$

The same argument is valid for the cylinders $Q_{t_s} = \Omega \times (t_{s-1}, t_s)$, $2 \leq s \leq N$, as long as their height satisfies the requirement analogue of (4.10). Thus, after a finite number of steps we obtain for the component R the estimate (4.2), for almost every $(x,t) \in Q_T$. □

In a similar way we can proceed with the other components.

5. The discrete maximum principle and stability. In this section we show that the commonly used central difference second order explicit schemes in general violate the discrete maximum principle. Nevertheless, for these schemes, we give a theoretical proof of stability.

We work on a rectangular grid

$$x_i = i\Delta x, \quad y_j = j\Delta y, \quad t_m = m\Delta t,$$

$$i,j = 0,1,2,\ldots,M, \quad m = 0,1,2,\ldots,\left[\frac{T}{\Delta t}\right].$$

The spatial units are normalized such that $\Delta x = \Delta y = 1$. The approximate solution $(R^m_{ij}, G^m_{ij}, B^m_{ij})$ samples the functions

$$R^m_{ij} \equiv U^1(i\Delta x, j\Delta y, m\Delta t),$$

$$G^m_{ij} \equiv U^2(i\Delta x, j\Delta y, m\Delta t),$$

$$B^m_{ij} \equiv U^3(i\Delta x, j\Delta y, m\Delta t).$$

On the boundary we impose the Neumann boundary condition. This corresponds to a prolongation by reflection of the image across the boundary.

We replace the second spatial derivatives and the first time derivative by a central difference and forward difference, respectively. Based on (2.4), the first element R of the color vector satisfies the following equation:

$$R_t = \frac{1}{\sqrt{g}}\operatorname{div}(D\nabla R). \tag{5.1}$$

The diffusion matrix is written here as

$$D = \begin{pmatrix} a & b \\ b & c \end{pmatrix},$$

where the coefficients are given in terms of the image metric $a = g_{22}/\sqrt{g}$, $c = g_{11}/\sqrt{g}$, $b = -g_{12}/\sqrt{g}$. With this notation, (5.1) is written as

$$R_t = \frac{1}{\sqrt{g}}((aR_x + bR_y)_x + (bR_x + cR_y)_y). \tag{5.2}$$

We approximate (5.2) by the following central difference explicit scheme:

$$R_{ij}^{m+1} = R_{ij}^m + \beta \Delta t O_{ij}(R^m, G^m, B^m), \tag{5.3}$$

where $O_{ij}(R^m, G^m, B^m)$ is the discrete version of the right-hand side of (5.1) and is given explicitly, in the central difference framework, by

(5.4)

$$\begin{aligned} O_{ij} \;=\; & \frac{1}{\sqrt{g}_{i,j}^m}\Big[a_{i+\frac{1}{2},j}^m(R_{i+1,j}^m - R_{i,j}^m) - a_{i-\frac{1}{2},j}^m(R_{i,j}^m - R_{i-1,j}^m) \\ & + c_{i,j+\frac{1}{2}}^m(R_{i,j+1}^m - R_{i,j}^m) - c_{i,j-\frac{1}{2}}^m(R_{i,j}^m - R_{i,j-1}^m) \\ & + \frac{1}{4}b_{i,j+1}^m(R_{i+1,j+1}^m - R_{i-1,j+1}^m) - \frac{1}{4}b_{i,j-1}^m(R_{i+1,j-1}^m - R_{i-1,j-1}^m) \\ & + \frac{1}{4}b_{i+1,j}^m(R_{i+1,j+1}^m - R_{i+1,j-1}^m) - \frac{1}{4}b_{i-1,j}^m(R_{i-1,j+1}^m - R_{i-1,j-1}^m)\Big], \end{aligned}$$

where the half indices are obtained by linear interpolation. The equations for the two other color components are discretized in the same manner. This scheme is stable under CFL-like bound requirements of the time step. The stability, as well as the lack of extremum principle property, can be seen in the following theorem.

THEOREM 5.1. *If, for all* $m = 0, 1, 2, \ldots, [\frac{T}{\Delta t}]$, Δt *satisfies the condition*

$$\Delta t \le \frac{1}{8\beta \max\limits_{i,j}\left\{\frac{a_{i+\frac{1}{2},j}^m}{\sqrt{g_{i,j}^m}}, \frac{a_{i-\frac{1}{2},j}^m}{\sqrt{g_{i,j}^m}}, \frac{c_{i,j+\frac{1}{2}}^m}{\sqrt{g_{i,j}^m}}, \frac{c_{i,j-\frac{1}{2}}^m}{\sqrt{g_{i,j}^m}}\right\}}, \tag{5.5}$$

then the solution satisfies

$$\begin{aligned} |R_{i,j}^m| &\le e^{S\frac{\beta}{2}t_m} \max_{i,j}|R_{i,j}^0|, \\ |G_{i,j}^m| &\le e^{S\frac{\beta}{2}t_m} \max_{i,j}|G_{i,j}^0|, \\ |B_{i,j}^m| &\le e^{S\frac{\beta}{2}t_m} \max_{i,j}|B_{i,j}^0|, \end{aligned} \tag{5.6}$$

where

(5.7)

$$S = \max_{0\le p\le m} S_p \text{ and } S_p = \max_{i,j}\frac{|b_{i,j+1}^p|}{\sqrt{g_{ij}^p}} + \max_{i,j}\frac{|b_{i,j-1}^p|}{\sqrt{g_{ij}^p}} + \max_{i,j}\frac{|b_{i+1,j}^p|}{\sqrt{g_{ij}^p}} + \max_{i,j}\frac{|b_{i-1,j}^p|}{\sqrt{g_{ij}^p}}.$$

Proof. We give the proof for only one of the components, since the proof is the same for the other two. We introduce the following notation:

$$L_{ij}^m = a_{i+\frac{1}{2},j}^m (R_{i+1,j}^m - R_{i,j}^m) - a_{i-\frac{1}{2},j}^m (R_{i,j}^m - R_{i-1,j}^m), \tag{5.8}$$

$$M_{ij}^m = c_{i,j+\frac{1}{2}}^m (R_{i,j+1}^m - R_{i,j}^m) - c_{i,j-\frac{1}{2},j}^m (R_{i,j}^m - R_{i,j-1}^m), \tag{5.9}$$

$$N_{ij}^m = \frac{1}{4}\Big[b_{i,j+1}^m (R_{i+1,j+1}^m - R_{i-1,j+1}^m) - b_{i,j-1}^m (R_{i+1,j-1}^m - R_{i-1,j-1}^m)\Big], \tag{5.10}$$

$$P_{ij}^m = \frac{1}{4}\Big[b_{i+1,j}^m (R_{i+1,j+1}^m - R_{i+1,j-1}^m) - b_{i-1,j}^m (R_{i-1,j+1}^m - R_{i-1,j-1}^m)\Big]. \tag{5.11}$$

Therefore, using (5.3) and (5.5), we can write

$$\begin{aligned}\left|R_{ij}^{m+1}\right| \le \frac{1}{2}\left|R_{ij}^m\right| + \left|\frac{1}{4}R_{ij}^m + \beta\frac{\Delta t}{\sqrt{g_{ij}^m}}L_{ij}^m\right| + \left|\frac{1}{4}R_{ij}^m + \beta\frac{\Delta t}{\sqrt{g_{ij}^m}}M_{ij}^m\right| \\ + \left|\beta\frac{\Delta t}{\sqrt{g_{ij}^m}}N_{ij}^m\right| + \left|\beta\frac{\Delta t}{\sqrt{g_{ij}^m}}P_{ij}^m\right| .\end{aligned} \tag{5.12}$$

If

$$\Delta t \le \frac{1}{8\beta \max\limits_{i,j} \frac{a_{i+\frac{1}{2},j}^m}{\sqrt{g_{i,j}^m}}}$$

for all $m = 0, 1, 2, \ldots, [\frac{T}{\Delta t}]$, then

$$\begin{aligned}&\left|\frac{1}{8}R_{ij}^m + \beta\frac{\Delta t}{\sqrt{g_{ij}^m}}a_{i+\frac{1}{2},j}^m (R_{i+1,j}^m - R_{i,j}^m)\right| \\ &\qquad \le \max(|R_{i,j}^m|, |R_{i+1,j}^m|)\left(\frac{1}{8} - \frac{\beta\Delta t}{\sqrt{g_{ij}^m}}a_{i+\frac{1}{2},j}^m + \frac{\beta\Delta t}{\sqrt{g_{ij}^m}}a_{i+\frac{1}{2},j}^m\right),\end{aligned}$$

which implies

$$\left|\frac{1}{8}R_{ij}^m + \beta\frac{\Delta t}{\sqrt{g_{ij}^m}}a_{i+\frac{1}{2},j}^m (R_{i+1,j}^m - R_{i,j}^m)\right| \le \frac{1}{8}\max_{i,j}|R_{ij}^m| .$$

Thus we can get the estimate

$$\left|2\cdot\frac{1}{8}R_{ij}^m + \beta\frac{\Delta t}{\sqrt{g_{ij}^m}}L_{ij}^m\right| \le \frac{1}{4}\max_{ij}|R_{ij}^m| \quad \text{if} \quad \Delta t \le \frac{1}{8\beta\max\limits_{i,j}\left\{\frac{a_{i+\frac{1}{2},j}^m}{\sqrt{g_{i,j}^m}}, \frac{a_{i-\frac{1}{2},j}^m}{\sqrt{g_{i,j}^m}}\right\}} . \tag{5.13}$$

In the same way we obtain

$$\left|2\cdot\frac{1}{8}R_{ij}^m + \beta\frac{\Delta t}{\sqrt{g_{ij}^m}}M_{ij}^m\right| \le \frac{1}{4}\max_{ij}|R_{ij}^m| \quad \text{if} \quad \Delta t \le \frac{1}{8\beta\max\limits_{i,j}\left\{\frac{c_{i,j+\frac{1}{2}}^m}{\sqrt{g_{i,j}^m}}, \frac{c_{i,j-\frac{1}{2}}^m}{\sqrt{g_{i,j}^m}}\right\}} . \tag{5.14}$$

Let

$$N_{i,j}^{m,1} = \beta \frac{\Delta t}{4\sqrt{g_{ij}^m}} b_{i,j+1}^m (R_{i+1,j+1}^m - R_{i-1,j+1}^m)$$

and

$$N_{i,j}^{m,2} = \beta \frac{\Delta t}{4\sqrt{g_{ij}^m}} b_{i,j-1}^m (R_{i+1,j-1}^m - R_{i-1,j-1}^m).$$

Then we obtain for $N_{i,j}^{m,1}$

$$|N_{i,j}^{m,1}| \le 2 \cdot \beta \frac{\Delta t}{4\sqrt{g_{ij}^m}} |b_{i,j+1}^m| \max_{ij} |R_{i,j}^m|. \tag{5.15}$$

A similar inequality can be written for $|N_{i,j}^{m,2}|$ and for $|P_{i,j}^m|$. Thus we have

$$\begin{aligned} |N_{ij}^m| &\le \beta \frac{\Delta t}{2\sqrt{g_{ij}^m}} (|b_{i,j+1}^m| + |b_{i,j-1}^m|) \max_{ij} |R_{i,j}^m| \\ &\le \frac{\beta}{2} \Delta t \left(\max_{i,j} \frac{|b_{i,j+1}^m|}{\sqrt{g_{ij}^m}} + \max_{i,j} \frac{|b_{i,j-1}^m|}{\sqrt{g_{ij}^m}} \right) \max_{ij} |R_{i,j}^m| \end{aligned} \tag{5.16}$$

and

$$\begin{aligned} |P_{ij}^m| &\le \beta \frac{\Delta t}{2\sqrt{g_{ij}^m}} (|b_{i+1,j}^m| + |b_{i-1,j}^m|) \max_{ij} |R_{i,j}^m| \\ &\le \frac{\beta}{2} \Delta t \left(\max_{i,j} \frac{|b_{i+1,j}^m|}{\sqrt{g_{ij}^m}} + \max_{i,j} \frac{|b_{i-1,j}^m|}{\sqrt{g_{ij}^m}} \right) \max_{ij} |R_{i,j}^m|. \end{aligned} \tag{5.17}$$

From (5.13)–(5.17) it follows that

$$\begin{aligned} &|R_{ij}^{m+1}| \\ &\le \max_{i,j} |R_{ij}^m| \left(1 + \frac{\beta}{2} \Delta t \left(\max_{i,j} \frac{|b_{i,j+1}^m|}{\sqrt{g_{ij}^m}} + \max_{i,j} \frac{|b_{i,j-1}^m|}{\sqrt{g_{ij}^m}} + \max_{i,j} \frac{|b_{i+1,j}^m|}{\sqrt{g_{ij}^m}} + \max_{i,j} \frac{|b_{i-1,j}^m|}{\sqrt{g_{ij}^m}} \right) \right) \end{aligned} \tag{5.18}$$

if

$$\Delta t \le \frac{1}{8\beta \max\limits_{i,j} \left\{ \frac{a_{i+\frac{1}{2},j}^m}{\sqrt{g_{i,j}^m}}, \frac{a_{i-\frac{1}{2},j}^m}{\sqrt{g_{i,j}}}, \frac{c_{i,j+\frac{1}{2}}^m}{\sqrt{g_{i,j}^m}}, \frac{c_{i,j-\frac{1}{2}}^m}{\sqrt{g_{i,j}^m}} \right\}}. \tag{5.19}$$

Let

$$M_1^m = \max_{i,j} \frac{|b_{i,j+1}^m|}{\sqrt{g_{ij}^m}}, \qquad M_2^m = \max_{i,j} \frac{|b_{i,j-1}^m|}{\sqrt{g_{ij}^m}}$$

and

$$M_3^m = \max_{i,j} \frac{|b_{i+1,j}^m|}{\sqrt{g_{ij}^m}}, \qquad M_4^m = \max_{i,j} \frac{|b_{i-1,j}^m|}{\sqrt{g_{ij}^m}}.$$

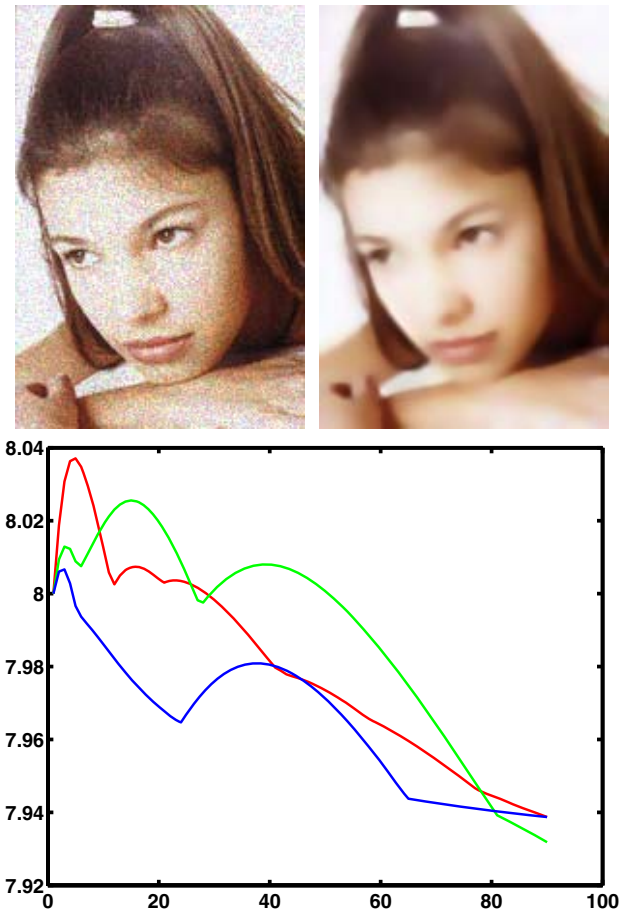


FIG. 5.1. *Top-left: noisy Camila image. Top-right: result of the Beltrami flow after* 90 *iteration. Bottom: plot of maximum of each of the channels versus number of iterations. Parameters:* $\beta^2 = 100$, $\Delta t = 0.0091$.

Next, applying (5.18) repeatedly, we find that if condition (5.19) is satisfied, then

$$\begin{aligned}|R_{ij}^m| &\le \left(1+\frac{\beta}{2}\Delta t S_{m-1}\right)\left(1+\frac{\beta}{2}\Delta t S_{m-2}\right)\cdots\left(1+\frac{\beta}{2}\Delta t S_0\right)\max_{i,j}|R_{i,j}^0| \\ &\le \left(1+\frac{\beta}{2}\Delta t S\right)^m \max_{i,j}|R_{i,j}^0| \le e^{\frac{\beta}{2}t_m S}\max_{i,j}|R_{i,j}^0|,\end{aligned}$$

where S is given in (5.7). □

The inequalities (5.6) clearly show that the maximum principle can be violated, but we still have stability. The inequalities in Theorem 5.1 show that the numerical solution is bounded in each iteration by the maximum value of the initial image multiplied by a factor. It guarantees that the flow does not blow up in finite time and ensures its stability. At the same time it is clear from the positivity of β that the maximum principle can be violated. One can actually see it in practice (see Figures 5.1, 5.2, and 5.3). We note that this does not indicate that the scheme is not accurate. This situation is not unprecedented. The Crank–Nicolson scheme for the

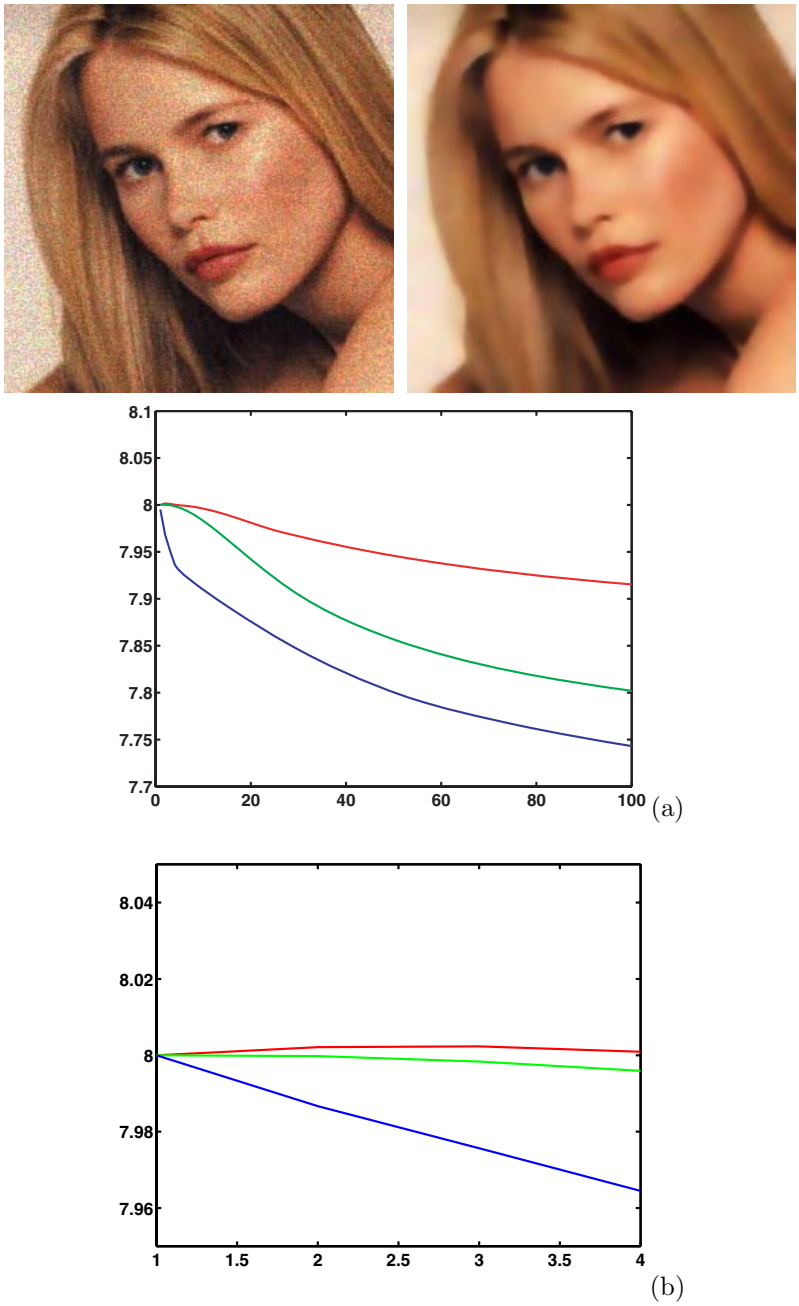


FIG. 5.2. *Top-left: noisy Claudia image. Top-right: result of the Beltrami flow after* 100 *iterations. Bottom*: (a) *Plot of the maximum of each of the channels versus number of iterations.* (b) *Detail of the previous graph (the first* 4 *iterations). Parameters:* $\beta^2 = 100$, $\Delta t = 0.01$.

one-dimensional heat equation, for example, is also known not to obey the maximum principle while being a useful and accurate scheme.

The reason for this discrepancy between the continuous and the discrete setting is that this second order approximation is not a nonnegative one. Indeed, the mixed

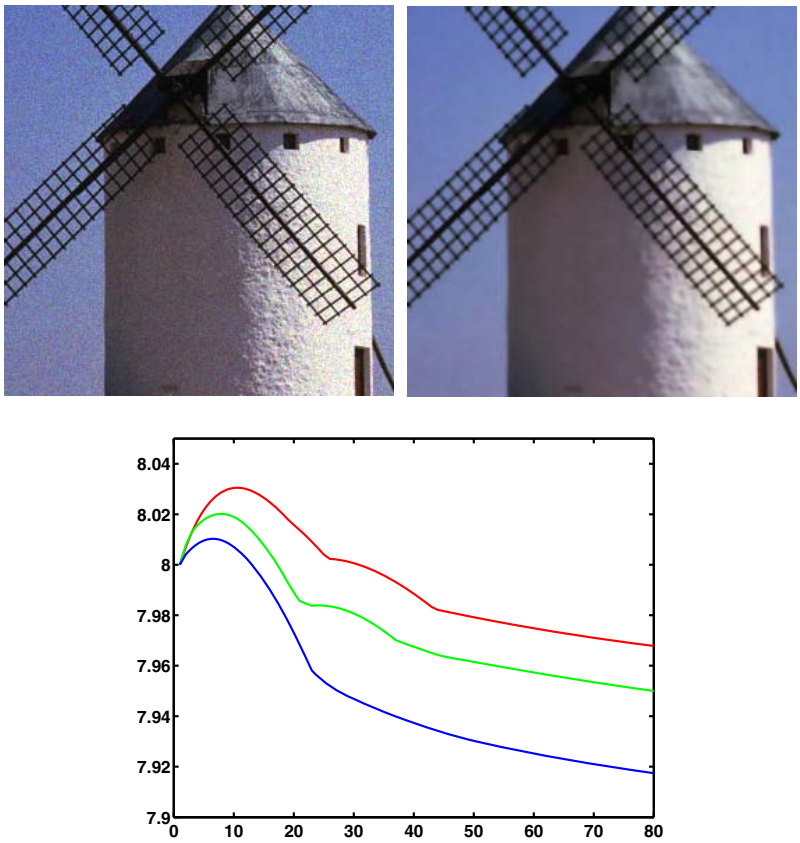


FIG. 5.3. *Top-left: noisy windmill. Top-right: denoised image by the Beltrami flow after* 80 *iterations. Bottom: plot of maximum of each of the channels versus number of iterations. Parameters:* β^2=3, $\Delta t = 0.0091$.

derivatives in (5.2) can create negative weights in certain pixels. One can easily show that a scheme which is based on a nonnegative discretization does satisfy the discrete maximum principle. Based on this result, the problem of proving the discrete maximum principle boils down to the problem of finding a nonnegative second order difference approximation. In [22], Weickert proposed a way to build a nonnegative scheme. The nonnegativity of his proposed scheme depends, however, on the condition number of the diffusion tensor D. Only in pixels where the condition number is smaller than $3 + 2\sqrt{2}$ are the weights nonnegative. This limits the application of the scheme, since in many images the condition number is higher than this limit in many pixels.

6. Details of the implementation and results. In this section we present results that represent the numerical behavior of the above described numerical scheme. The initial data are given in three channels r, g, and b in the range 0 to 255. We first transfer the images to the more perceptually adaptive coordinates $R = \log(1+r)$, $G = \log(1+g)$, $B = \log(1+b)$. The dynamic range of these variables is 0 to 8, and these adaptive coordinates do not limit the generality of our analysis. In the two examples presented below we corrupt the images with random noise and then denoise them using the scheme mentioned above. In the implementation, the parameters β and Δt were chosen to satisfy the stability condition (5.5).

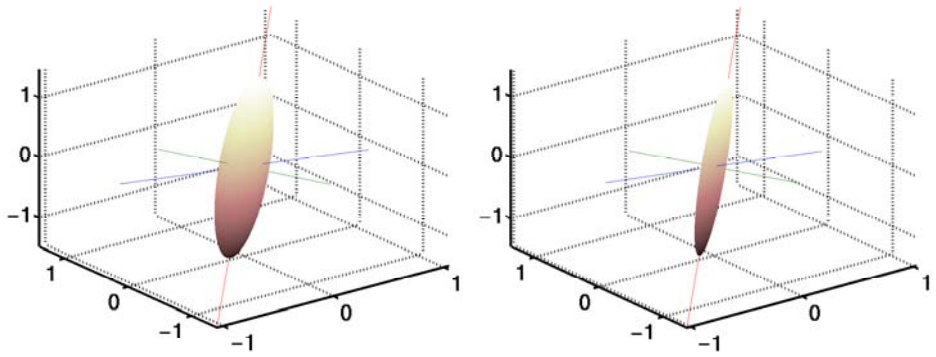


FIG. 6.1. *Camila image. Left: ellipsoid—initial noisy data. Right: ellipsoid—after applying Beltrami,* 90 *iterations,* $\beta = 10$.

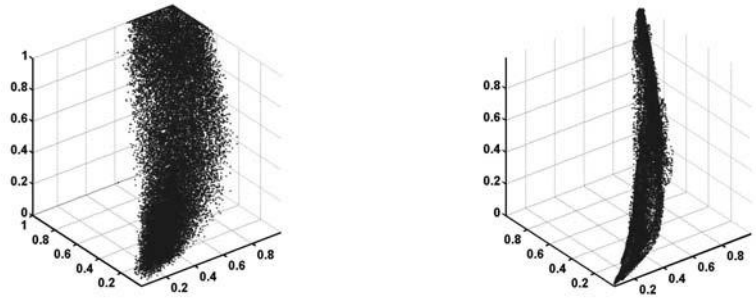


FIG. 6.2. *For Camila image* ($\beta = 10$). *Left: distribution of the initial noisy data. Right: distribution of the data after* 100 *iterations.*

Figures 5.1, 5.2, and 5.3 all demonstrate the stability of the process, on the one hand, and the violation of the maximum principle, on the other. In Figures 5.2 and 5.3, one notices that after a certain small number of iterations the maximum principle is satisfied. This is not the case in Figure 5.1, where the violation of the maximum principle is stable and is observable over the whole evolution. The stability can also be explained by the experiments presented in Figures 6.1 and 6.2.

Figure 6.2 depicts the distribution of colors in the Camila image before and after the Beltrami color flow. In Figure 6.1 the ellipsoids have as principle axes the eigenvectors of the covariance-matrix of the color image. The contracting form of the ellipsoid after applying the Beltrami flow indicates a stable denoising process.

7. Concluding remarks. In this paper, we studied the extremum principle property for the Beltrami color flow. We adapted the duality paradigm of Florack and considered "true images" as generalized functions. We therefore investigated, besides the strong solutions, also the generalized (weak) solutions. We proved the extremum principle in both the strong and the weak formulations.

We also addressed the problem of the discrete maximum principle and its close relationship with stability. In contrast to the continuous case, the discrete maximum principle cannot automatically be guaranteed. The central difference scheme does not necessarily satisfy the extremum principle. Though this property is violated, we

proved the stability of the scheme. Numerical examples show, nevertheless, that this scheme is a useful tool in denoising.

Questions of existence and uniqueness, as well as analysis of more elaborated numerical schemes, are currently being studied.

Acknowledgment. We thank Shoshana Kamin for interesting discussions.

REFERENCES

[1] L. Alvarez, F. Guichard, P. L. Lions, and J. M. Morel, *Axioms and fundamental equations of image processing*, Arch. Ration. Mech. Anal., 123 (1993), pp. 199–257.
[2] V. Caselles, R. Kimmel, and G. Sapiro, *Geodesic active contours*, Int. J. Computer Vision, 22 (1997), pp. 61–79.
[3] F. Catté, P.-L. Lions, J.-M. Morel, and T. Coll, *Image selective smoothing and edge detection by nonlinear diffusion*, SIAM J. Numer. Anal., 29 (1992), pp. 182–193.
[4] Y. Chen, B. C. Vemuri, and Wang Li, *Image denoising and segmentation via nonlinear diffusion*, Comput. Math. Appl., 39 (2000), pp. 131–149.
[5] L. Evans, *Partial Differential Equations*, Berkeley Mathematics, Berkeley, CA, 1994.
[6] L. Florack, *Duality principles in image processing and analysis*, in Image and Vision Computing, R. Klette, G. Gimel'farb, and R. Kakarala, eds., University of Auckland, Auckland, New Zealand, 1998, pp. 286–296.
[7] R. A. Hummel, *Representations based on zero-crossings in scalespace*, in Proceedings of the IEEE Computation Society Conference on Computer Vision and Pattern Recognition, IEEE Press, Piscataway, NJ, 1986, pp. 204–209.
[8] R. Kimmel, R. Malladi, and N. Sochen, *Images as embedding maps and minimal surfaces: Movies, color, texture, and volumetric medical images*, Int. J. Computer Vision, 39 (2000), pp. 111–129.
[9] R. Kimmel and N. Sochen, *Orientation diffusion or how to comb a porcupine*, J. Visual Communication and Image Representation, 13 (2001), pp. 238–248.
[10] J. J. Koenderinck, *The structure of images*, Biol. Cybernet., 50 (1993), pp. 363–370.
[11] D. Mumford and B. Gidas, *Stochastic models for generic images*, Quart. Appl. Math, 59 (2001), pp. 85–111.
[12] P. Perona and J. Malik, *Scale-space and edge detection using anisotropic diffusion*, IEEE Trans. Pattern Analysis and Machine Intelligence, 12 (1990), pp. 629–639.
[13] A. M. Polyakov, *Quantum geometry of bosonic strings*, Phys. Lett., 103 (1981), pp. 207–210.
[14] M. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1967.
[15] L. Rudin, S. Osher, and E. Fatemi, *Non-linear total variation-based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.
[16] L. Schwartz, *Functional Analysis*, Courant Institute of Mathematical Sciences, New York, 1964.
[17] N. Sochen, R. Deriche, and L. Perez-Lopez, *The Beltrami flow over implicit manifolds*, in Proceedings of the International Conference on Computer Vision, Nice, France, 2003, IEEE Computer Society Press, Los Alamitos, CA, 2003, pp. 832–839.
[18] N. Sochen, R. Kimmel, and R. Malladi, *From High Energy Physics to Low Level Vision*, Report 39243, Lawrence Berkeley National Laboratory, University of California Berkeley, 1996.
[19] N. Sochen, R. Kimmel, and R. Malladi, *A general framework for low level vision*, IEEE Trans. Image Processing, 7 (1998), pp. 310–318.
[20] N. Sochen and Y. Y. Zeevi, *Representation of colored images by manifolds embedded in higher dimensional non-Euclidean space*, in Proceedings of the IEEE International Conference on Image Processing (ICIP'98), Chicago, IEEE Press, Piscataway, NJ, 1998, pp. 166–170.
[21] N. Sochen and Y. Y. Zeevi, *Representation of images by surfaces and higher dimensional manifolds in non-Euclidean space*, in Mathematical Methods for Curves and Surfaces II, Vanderbilt University Press, Nashville, TN, 1998, pp. 469–476.
[22] J. Weickert, *Anisotropic Diffusion in Image Processing*, Teubner, Stuttgart, 1998.
[23] J. Weickert, *Coherence-enhancing diffusion of color images*, Image and Vision Computing, 17 (1999), pp. 201–212.
[24] S. Di Zenzo, *A note on the gradient of a multiimage*, Computer Vision, Graphics, and Image Processing, 33 (1986), pp. 116–125.

SIAM J. APPL. MATH.
Vol. 65, No. 5, pp. 1633–1656

PROBLEMS OF STATIONARY FLOW OF ELECTRORHEOLOGICAL FLUIDS IN A CYLINDRICAL COORDINATE SYSTEM*

R. H. W. HOPPE†‡, W. G. LITVINOV‡, AND T. RAHMAN§

Abstract. We consider the general problem on stationary flow of the electrorheological fluid with the constitutive equation developed in [R. H. W. Hoppe and W. G. Litvinov, *Comm. Pure. Appl. Anal.,* 3 (2004), pp. 809–848] in the cylindrical coordinate system. The problem is studied under mixed boundary conditions wherein velocities are specified on one part of the boundary and surface forces are given on the other part. The existence of a solution to this problem and the convergence of Galerkin approximations are established. Then, we consider the occasion where the flow is axially symmetric and study a problem on an electrorheological clutch. This problem is solved numerically, and the results of calculations of the electric field and velocities are presented.

Key words. electrorheological fluid, generalized solution, existence theorem, approximate solutions, electrorheological clutch

AMS subject classification. 35Q35

DOI. 10.1137/S0036139903432883

1. Introduction.

Electrorheological fluids are smart materials which are concentrated suspensions of polarizable particles in a nonconducting dielectric liquid. In moderately large electric fields, the particles form chains along the field lines, and these chains then aggregate to form columns [16]. These chainlike and columnar structures cause dramatic changes in the rheological properties of the suspensions. The fluids become anisotropic; the apparent viscosity (the resistance to flow) in the direction orthogonal to the direction of electric field abruptly increases, while the apparent viscosity in the direction of the electric field changes not so drastically.

The chainlike structures directed along the magnetic field lines are formed in magnetic suspensions whose behavior is similar to the behavior of electrorheological suspensions. It was shown experimentally that the apparent viscosity of the flow of magnetic suspensions in the direction orthogonal to the direction of the magnetic field is about three times greater than the apparent viscosity of the flow in the direction of the magnetic field; see [18, p. 85].

The chainlike and columnar structures are destroyed under the action of large stresses, and then the apparent viscosity of the fluid decreases and the fluid becomes less anisotropic.

The following constitutive equation of electrorheological fluids was developed in [8]:

(1.1)
$$\sigma_{ij}(p,u,E) = -p\delta_{ij} + 2\varphi(I(u),|E|,\mu(u,E))\varepsilon_{ij}(u), \quad i,j = 1,\dots,n, \quad n = 2 \text{ or } 3.$$

*Received by the editors August 7, 2003; accepted for publication (in revised form) August 16, 2004; published electronically June 14, 2005. This work has been supported by the German National Science Foundation (DFG) within the Collaborative Research Center SFB 438.
http://www.siam.org/journals/siap/65-5/43288.html
†Department of Mathematics, University of Houston, Houston, TX 77204-3008 (rohop@math.uh.edu).
‡Institute of Mathematics, University of Augsburg, Universitaetsstr. 14 D-86159 Augsburg, Germany (litvinov@math.uni-augsburg.de).
§The Bergen Center for Computational Science, Thormhlensgt. 55, N-5008 Bergen, Norway (talal@ii.uib.no).

Here, $\sigma_{ij}(p,u,E)$ are the components of the stress tensor which depend on the pressure p, the velocity vector $u = (u_1,\dots,u_n)$, and the electric field strength $E = (E_1,\dots,E_n)$; δ_{ij} are the components of the unit tensor (the Kronecker delta); and $\varepsilon_{ij}(u)$ are the components of the rate of strain tensor

$$\varepsilon_{ij}(u) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \tag{1.2}$$

where x_i are Cartesian coordinates of a point $x = (x_1,\dots,x_n)$.

Moreover, $I(u)$ is the second invariant of the rate-of-strain tensor

$$I(u) = \sum_{i,j=1}^{n} (\varepsilon_{ij}(u))^2, \tag{1.3}$$

and φ the viscosity function depending on $I(u)$, $|E|$, and $\mu(u,E)$.

The function μ is introduced into the constitutive equation (1.1) in order to take into account the anisotropy of the electrorheological fluid under which the viscosity of the fluid depends on the angle between the vector of the electric field and the vector of the velocity with respect to the charged electrode (the counter electrode is not charged usually). The electrode can move relative to the body of an electrorheological device, and hence we consider that the electrode can move relative to the reference frame under consideration.

Let $\breve{u}(x,t) = (\breve{u}_1(x,t), \breve{u}_2(x,t), \breve{u}_3(x,t))$ be a vector of transfer velocity; $\breve{u}(x,t)$ is the velocity of a point of the electrode which coincides with the point x of the frame at an instant t. It is assumed that $\breve{u}$ is a known function.

We define the function μ as the square of the cosine of the angle between the vector of the electric field and the vector of the velocity relative to the electrode, i.e.,

$$\mu(u,E) = \left(\frac{u-\breve{u}}{|u-\breve{u}|}, \frac{E}{|E|}\right)^2_{\mathbb{R}^3} = \frac{((u_i-\breve{u}_i)E_i)^2}{\left(\sum_{i=1}^3 (u_i-\breve{u}_i)^2\right)\left(\sum_{i=1}^3 E_i^2\right)}. \tag{1.4}$$

Here and below, the Einstein convention on summation over a repeated index is applied, and we denote by $(.,.)_{\mathbb{R}^3}$ the scalar product in $\mathbb{R}^3$.

If the electrode does not move relative to the reference frame, then $\breve{u} = 0$ and the function μ takes the form

$$\mu(u,E) = \left(\frac{u}{|u|}, \frac{E}{|E|}\right)^2_{\mathbb{R}^3}. \tag{1.5}$$

In the general case, the function $\breve{u}$ is defined as follows:

$$\breve{u}(x,t) = \mathring{u}(t) + w(x,t), \tag{1.6}$$

where $\mathring{u}(t) = (\mathring{u}_1(t), \mathring{u}_2(t), \mathring{u}_3(t))$ is a vector of the translation velocity and $w(x,t) = (w_1(x,t), w_2(x,t), w_3(x,t))$ is a vector of the rotational velocity.

The function $\mu(u,E)$ is an invariant which is independent of the choice of the reference frame and the motion of the frame with respect to the electrode.

The viscosity function φ is identified by approximation of flow curves (see [8]) and it was shown in [8] (see also the appendix) that it can be represented as follows:

$$\varphi(I(u),|E|,\mu(u,E)) = b(|E|,\mu(u,E))(\lambda + I(u))^{-\frac{1}{2}} + \psi(I(u),|E|,\mu(u,E)), \tag{1.7}$$

where λ is a small parameter, $\lambda \geq 0$.

The constitutive equation (1.1) with the viscosity function (1.7) allows one to describe the following main peculiarities of flow of electrorheological fluids:

(a) singular or almost singular viscosity function at zero value of the rate-of-strain tensor,
(b) an arbitrary nonlinear relationship between the shear rates and the shear stresses,
(c) an arbitrary dependence of the viscosity on the module of the vector of the electric field and on the angle between the vectors of the velocity and electric field (the anisotropy).

With some assumptions natural from a physical point of view, the constitutive equation (1.1) with the viscosity function (1.7) leads to well-posed mathematical problems (see sections 4 and 5 below and [8]).

The functions b and ψ in (1.7) can be identified so that a set of flow curves obtained for different electric fields E is approximated in an arbitrary range of the shear rates with an arbitrarily high degree of accuracy (for example, by splines).

The Bingham constitutive equation of electrorheological fluids, which is of considerable current use (see, e.g., [4], [16], [22]), gives no way to closely approximate a set of flow curves, especially at small shear rates (see Figure A-1 in the appendix). In addition, the Bingham constitutive equation takes no account of the anisotropy of electrorheological fluids.

We consider Maxwell's equations in the following form (see, e.g., [10]):

$$\begin{aligned} &\operatorname{curl} E + \frac{1}{c}\frac{\partial B}{\partial t} = 0, \qquad \operatorname{div} B = 0, \\ &\operatorname{curl} H - \frac{1}{c}\frac{\partial D}{\partial t} = 0, \qquad \operatorname{div} D = 0. \end{aligned} \tag{1.8}$$

Here E is the electric field, B the magnetic induction, D the electric displacement, H the magnetic field, and c the speed of light. One can assume that

$$D = \epsilon E, \qquad B = \mu H, \tag{1.9}$$

where ϵ is the dielectric permittivity and μ the magnetic permeability.

Since electrorheological fluids are dielectrics, the magnetic field H can be neglected. Then (1.8), (1.9) give the following relations:

$$\operatorname{curl} E = 0, \tag{1.10}$$

$$\operatorname{div}(\epsilon E) = 0. \tag{1.11}$$

It follows from (1.10) that there exists a function of potential θ such that

$$E = -\operatorname{grad}\theta, \tag{1.12}$$

and (1.11) implies

$$\operatorname{div}(\epsilon \operatorname{grad}\theta) = 0 \quad \text{in} \quad \Omega_1. \tag{1.13}$$

Here Ω_1 is the domain of the fluid flow in the Cartesian coordinate system.

The boundary conditions are the following:

$$\theta = U_i(t) \quad \text{on} \quad \Gamma_i, \quad i = 1, \dots, k, \tag{1.14}$$

$$\theta = 0 \quad \text{on} \quad \Gamma_{i0}, \tag{1.15}$$

$$\nu \cdot \epsilon \operatorname{grad}\theta = 0 \quad \text{on} \quad \Gamma \setminus \left(\bigcup_{i=1}^{k} \overline{(\Gamma_i \cup \Gamma_{i0})} \right). \tag{1.16}$$

Here Γ_i and Γ_{i0} are the surfaces of the ith control and null electrodes, respectively, and it is supposed that Γ_i, Γ_{i0} are open subsets of the boundary Γ of Ω_1.

Therefore, the equations for the functions E and (p, v) are separated. Because of this, we assume hereafter that the function of electric field E is known.

In the special case that the direction of the velocity relative to the electrode $u(x,t) - \check{u}(x,t)$ at each point (x,t) at which $E(x,t) \neq 0$ is known, the function $(x,t) \to \mu(u,E)(x,t)$ becomes well known, and the viscosity functions (1.7) takes the form

$$\varphi(I(u), |E|, x, t) = e(|E|, x, t)(\lambda + I(u))^{-\frac{1}{2}} + \psi_1(I(u), |E|, x, t), \tag{1.17}$$

where

$$\begin{aligned} e(|E|, x, t) &= b(|E|, \mu(u,E)(x,t)), \\ \psi_1(I(u), |E|, x, t) &= \psi(I(u), |E|, \mu(u,E)(x,t)). \end{aligned} \tag{1.18}$$

In many electrorheological devices the fluid flows in domains of which the boundaries are the surfaces of revolution. Problems on flow of electrorheological fluids in such domains are convenient to consider in cylindrical coordinates.

In section 2, we present governing equations. In section 3, we formulate a general boundary value problem on stationary flow of the electrorheological fluid in the cylindrical coordinate system and adduce some auxiliary results. Section 4 contains approximate solutions and existence theorems for the general boundary value problem. In section 5, we consider a problem on stationary axially symmetric flow. A problem on an electrorheological clutch is formulated and solved numerically in section 6.

2. Governing equations and assumptions. We consider the system of cylindrical coordinates r, α, z. An element of the length dl is defined in cylindrical coordinates as $dl = (dr^2 + r^2\, d\alpha^2 + dz^2)^{\frac{1}{2}}$. Denote the components of a vector v in the mobile orthonormal basis e_r, e_α, e_z by v_1, v_2, v_3; i.e., $v = (v_1, v_2, v_3)$.

Let $u = (u_1, u_2, u_3)$ be a velocity vector. The components of the rate-of-strain tensor have the following form in cylindrical coordinates,

$$\begin{aligned} &\epsilon_{11}(u) = \frac{\partial u_1}{\partial r}, \quad \epsilon_{22}(u) = \frac{1}{r}\frac{\partial u_2}{\partial \alpha} + \frac{u_1}{r}, \qquad \epsilon_{33}(u) = \frac{\partial u_3}{\partial z}, \\ &\epsilon_{12}(u) = \epsilon_{21}(u) = \frac{1}{2}\left(\frac{1}{r}\frac{\partial u_1}{\partial \alpha} + \frac{\partial u_2}{\partial r} - \frac{u_2}{r}\right), \\ &\epsilon_{23}(u) = \epsilon_{32}(u) = \frac{1}{2}\left(\frac{\partial u_2}{\partial z} + \frac{1}{r}\frac{\partial u_3}{\partial \alpha}\right), \\ &\epsilon_{13}(u) = \epsilon_{31}(u) = \frac{1}{2}\left(\frac{\partial u_1}{\partial z} + \frac{\partial u_3}{\partial r}\right), \end{aligned} \tag{2.1}$$

and the second invariant of the rate-of-strain tensor is defined by

$$I(u) = \sum_{i,j=1}^{3} (\epsilon_{ij}(u))^2. \tag{2.2}$$

We assume the following.

(A0) Ω_1 is a bounded domain in $\mathbb{R}^3$ with a Lipschitz continuous boundary Γ.

Let $\mathcal{P}$ be the operator of translation from cylindrical coordinates to Cartesian ones,

$$\mathcal{P} : (r, \alpha, z) \to \mathcal{P}(r, \alpha, z) = (x_1, x_2, x_3),$$
$$x_1 = r\cos\alpha, \quad x_2 = r\sin\alpha, \quad x_3 = z, \quad r \in \mathbb{R}_+, \quad \alpha \in [0, 2\pi), \quad z \in \mathbb{R}, \tag{2.3}$$

where $\mathbb{R}_+ = \{y \in \mathbb{R},\, y \geq 0\}$ and we identify the points $(0, \alpha, z)$ with the point $(0, 0, z)$, $\alpha \in [0, 2\pi)$. The inverse operator $\mathcal{P}^{-1}$ is defined by

$$\mathcal{P}^{-1} : \mathcal{P}^{-1}(x_1, x_2, x_3) = (r, \alpha, z),$$
$$r = (x_1^2 + x_2^2)^{\frac{1}{2}}, \qquad \alpha = \arctan\frac{x_2}{x_1}, \qquad z = x_3. \tag{2.4}$$

Here, we consider that the mapping $(x_1, x_2) \to \arctan\frac{x_2}{x_1}$ is a multifunction at the point $x_1 = x_2 = 0$, namely, $\arctan\frac{0}{0} = [0, 2\pi)$.

Let

$$\Omega = \mathcal{P}^{-1}(\Omega_1), \qquad S = \mathcal{P}^{-1}(\Gamma). \tag{2.5}$$

We consider a stationary flow problem under the Stokes approximation; i.e., we ignore inertial forces, which are assumed to be small as compared with the internal forces caused by the viscous stresses. Then the motion equations take the following form:

$$\frac{\partial p}{\partial r} - 2\frac{\partial}{\partial r}(\varphi\epsilon_{11}(u)) - \frac{2}{r}\frac{\partial}{\partial\alpha}(\varphi\epsilon_{12}(u)) - 2\frac{\partial}{\partial z}(\varphi\epsilon_{13}(u)) - \frac{2\varphi}{r}(\epsilon_{11}(u) - \epsilon_{22}(u)) = K_1 \quad \text{in } \Omega, \tag{2.6}$$

$$\frac{1}{r}\frac{\partial p}{\partial\alpha} - 2\frac{\partial}{\partial r}(\varphi\epsilon_{21}(u)) - \frac{2}{r}\frac{\partial}{\partial\alpha}(\varphi\epsilon_{22}(u)) - 2\frac{\partial}{\partial z}(\varphi\epsilon_{23}(u)) - \frac{4}{r}\varphi\epsilon_{12}(u) = K_2 \text{ in } \Omega, \tag{2.7}$$

$$\frac{\partial p}{\partial z} - 2\frac{\partial}{\partial r}(\varphi\epsilon_{31}(u)) - \frac{2}{r}\frac{\partial}{\partial\alpha}(\varphi\epsilon_{32}(u)) - 2\frac{\partial}{\partial z}(\varphi\epsilon_{33}(u)) - \frac{2}{r}\varphi\epsilon_{13}(u) = K_3 \text{ in } \Omega. \tag{2.8}$$

Here the viscosity function φ is defined either by (1.7) or by (1.17), and K_1, K_2, K_3 are the components of the volume force vector K.

The velocity function u meets the incompressibility condition

$$\operatorname{div}_c u = \frac{\partial u_1}{\partial r} + \frac{1}{r}\frac{\partial u_2}{\partial\alpha} + \frac{\partial u_3}{\partial z} + \frac{u_1}{r} = 0 \quad \text{in } \Omega. \tag{2.9}$$

Here and below, we denote by div_c the operator of divergence in cylindrical coordinates.

Suppose that S_1 and S_2 are open subsets of S such that S_1 is nonempty, $S_1 \cap S_2 = \emptyset$, and $\overline{S}_1 \cup \overline{S}_2 = S$. We consider mixed boundary conditions, wherein velocities are specified on S_1 and surface forces are given on S_2, i.e.,

$$u = \hat{u} \quad \text{on } S_1, \tag{2.10}$$

$$[-p\delta_{ij} + 2\varphi\epsilon_{ij}(u)]\nu_j = F_i \quad \text{on } S_2, \qquad i, j = 1, 2, 3. \tag{2.11}$$

Here, by ν_j and F_i we denote the components of the unit outward normal to S_2 and the components of the vector of surface force with respect to the basis vectors e_r, e_α, e_z.

Let

$$\tilde{S} = \{(r, \alpha, z) | r = 0, \quad \alpha \in [0, 2\pi), \quad z \in \mathbb{R}_+\}, \qquad S_0 = \Omega \cap \tilde{S}. \tag{2.12}$$

In particular, S_0 can be an empty set. It follows from (2.9) that

$$u_1 = 0 \quad \text{on} \quad S_0, \tag{2.13}$$

and therefore $\frac{\partial u_1}{\partial z} = 0$ on S_0, and since $\epsilon_{13}(u) = 0$ on S_0 (see (2.8)), we obtain

$$\frac{\partial u_3}{\partial r} = 0 \quad \text{on} \quad S_0. \tag{2.14}$$

It follows also from (2.6), (2.7), and (2.1) that

$$\begin{aligned} &\lim_{r \to 0} (\epsilon_{11}(u) - \epsilon_{22}(u))(r, \alpha, z) = 0, \\ &\lim_{r \to 0} (\epsilon_{12}(u))(r, \alpha, z) = 0, \qquad u_2 = 0 \quad \text{on } S_0. \end{aligned} \tag{2.15}$$

In the case when the viscosity function is defined by (1.7), we assume the following.

(A1) $b : y_1, y_2 \to b(y_1, y_2)$ is a function continuous in $\mathbb{R}_+ \times [0, 1]$, and, in addition,

$$0 \le b(y_1, y_2) \le a_0, \qquad (y_1, y_2) \in \mathbb{R}_+ \times [0, 1], \tag{2.16}$$

where a_0 is a positive constant.

(A2) $\psi : (y_1, y_2, y_3) \to \psi(y_1, y_2, y_3)$ is a function continuous in $\mathbb{R}^2_+ \times [0, 1]$, and for an arbitrarily fixed $(y_2, y_3) \in \mathbb{R}_+ \times [0, 1]$ the partial function $\psi(., y_2, y_3) : y_1 \to \psi(y_1, y_2, y_3)$ is continuously differentiable in $\mathbb{R}_+$, and the following inequalities hold:

$$a_2 \ge \psi(y_1, y_2, y_3) \ge a_1, \tag{2.17}$$

$$\psi(y_1, y_2, y_3) + 2 \frac{\partial \psi}{\partial y_1}(y_1, y_2, y_3) y_1 \ge a_3, \tag{2.18}$$

$$\left| \frac{\partial \psi}{\partial y_1}(y_1, y_2, y_3) \right| y_1 \le a_4, \tag{2.19}$$

where a_1–a_4 are positive constants.

In the case that the viscosity function is defined by (1.17), we suppose the following.

(A3) for an arbitrary fixed $(y_2, x, t) \in \mathbb{R}_+ \times \Omega_1 \times \mathbb{R}_+$, the partial function $\psi_1(., y_2, x, t) : y_1 \to \psi_1(y_1, y_2, x, t)$ is continuously differentiable in $\mathbb{R}_+$, and the following inequalities hold:

$$a_2 \ge \psi_1(y_1, y_2, x, t) \ge a_1, \tag{2.20}$$

$$\psi_1(y_1, y_2, x, t) + 2 \frac{\partial \psi_1}{\partial y_1}(y_1, y_2, x, t) y_1 \ge a_3, \tag{2.21}$$

$$\left| \frac{\partial \psi_1}{\partial y_1}(y_1, y_2, x, t) \right| y_1 \le a_4. \tag{2.22}$$

As for the function e, we assume

$$e \in L_\infty(\mathbb{R}_+ \times \Omega_1 \times \mathbb{R}_+), \quad 0 \le e(y, x, t) \le a_0, \qquad y \in \mathbb{R}_+, \quad x \in \Omega_1, \quad t \in \mathbb{R}_+. \tag{2.23}$$

At $\lambda = 0$ the viscosity function φ defined by (1.7) is singular at $I(u) = 0$, $\varphi(0, |E|, \mu(u, E)) = \infty$, and flow problems for such viscosity function reduce to the solution of variational inequalities.

The equation (1.7) with a small positive λ defines a fluid with a finite but possibly large viscosity at $I(u) = 0$. From a physical point of view a fluid with bounded viscosity is more reasonable than the fluid with singular unbounded viscosity (all is bounded in actuality). It is shown in [8] that the solutions of the problems with bounded viscosities converge to the solution of the problem with the singular viscosity as λ tends to zero. Because of this, we assume that

$$\lambda > 0 \quad \text{in (1.7) and (1.17).} \tag{2.24}$$

Let us dwell on the physical sense of the inequalities (2.16)–(2.23). The inequalities (2.16) and (2.17) indicate that the viscosity is bounded from below and from above by positive constants. The inequality (2.18) implies that for fixed values of $|E|$ and $\mu(u, E)$ the derivative of the function $I(u) \to G(u)$ is positive, where $G(u)$ is the second invariant of the stress deviator

$$G(u) = \sum_{i,j=1}^{n} (\sigma_{ij}(p, u, E) + p\delta_{ij})^2 = 4[\varphi(I(u), |E|, \ \mu(u, E))]^2 I(u).$$

This means that in case of simple shear flow the shear stress increases with increasing shear rate. (2.19) is a restriction on $\frac{\partial \varphi}{\partial y_1}$ for large values of y_1.

The inequalities (2.20)–(2.23) are analogous to the inequalities (2.16)–(2.19).

All inequalities (2.16)–(2.23) are natural from a physical point of view.

The viscosity function is identified by approximation of a set of flow curves which are obtained experimentally by viscometric testing for different electric fields. The inequalities (2.16)–(2.23) are consistent with the shapes of the flow curves and enable one to approximate a set of flow curves over a wide range of shear rates with a high degree of accuracy (see the appendix below and [3], [8], [19]).

3. Generalized solution of the problem. We define the following sets:

$$J_0 = \Bigg\{ v | v = (v_1, v_2, v_3) \in C^\infty(\overline{\Omega})^3, \ v_1\big|_{S_0} = 0, \ v_2\big|_{S_0} = 0,$$

$$\frac{\partial v^k}{\partial \alpha^k}\Big|_{\alpha=0} = \frac{\partial v^k}{\partial \alpha^k}\Big|_{\alpha=2\pi}, \ k = 0, 1, 2, \dots \Bigg\}, \tag{3.1}$$

$$J = \{v | v \in J_0, \ v\big|_{S_1} = 0\}, \tag{3.2}$$

$$J_1 = \{v | v \in J, \ \mathrm{div}_c \, v = 0\}. \tag{3.3}$$

Let H and H_1 be the closures of J and J_1 with respect to the norm

$$||v||_H = \left(\int_\Omega I(v) r \, dr \, d\alpha \, dz \right)^{\frac{1}{2}}, \tag{3.4}$$

and let H_0 be the closures of J_0 relative to the norm

$$||v||_{H_0} = \left(||v||_H^2 + \int_{S_1} |v|^2 \, ds \right)^{\frac{1}{2}}. \tag{3.5}$$

Let also Y be the space of scalar functions which are square integrable in Ω with respect to the measure $r\,dr\,d\alpha\,dz$. The norm in Y is defined by

$$||h||_Y = \left(\int_\Omega h^2\, r\, dr\, d\alpha\, dz \right)^{\frac{1}{2}}. \tag{3.6}$$

We define the operator G that maps H_0 into a set of vector valued functions determined in Ω_1 as follows:

$$\begin{aligned} v = (v_1, v_2, v_3) \in H_0, \qquad & G(v) = \{G(v)_i\}_{i=1}^3, \\ G(v)_1 = (v_1 \cos\alpha - v_2 \sin\alpha) \circ \mathcal{P}^{-1}, & \\ G(v)_2 = (v_1 \sin\alpha + v_2 \cos\alpha) \circ \mathcal{P}^{-1}, \quad & G(v)_3 = v_3 \circ \mathcal{P}^{-1}. \end{aligned} \tag{3.7}$$

We assign also the following norm in H_0:

$$||v||_1 = ||G(v)||_{H^1(\Omega_1)^3}, \tag{3.8}$$

where $||\,\cdot\,||_{H^1(\Omega_1)^3}$ is the norm of the product of three Sobolev spaces $H^1(\Omega_1)$.

LEMMA 3.1. *Suppose that the condition* (A0) *is satisfied. Then the expressions* (3.4) *and* (3.8) *define equivalent norms in* H*, and the expressions* (3.5) *and* (3.8) *are equivalent norms in* H_0*. The operator* G *is an isomorphism of* H_0 *onto* $H^1(\Omega_1)^3$*, and the following equality holds:*

$$||h||_Y = ||h \circ \mathcal{P}^{-1}||_{L_2(\Omega_1)}. \tag{3.9}$$

Proof. The equivalence of the norms (3.4) and (3.8) in H, and the norms (3.5) and (3.8) in H_0, follows from the fact that $I(v)$ is the invariant, i.e.,

$$\sum_{i,j=1}^3 [(\epsilon_{ij}(v))(r,\alpha,z)]^2 = \sum_{i,j=1}^3 [(\varepsilon_{ij}(G(v)))(\mathcal{P}(r,\alpha,z))]^2, \tag{3.10}$$

and from the Korn inequality.

Therefore, $G(H_0) \subset H^1(\Omega_1)^3$. Let $g = (g_1, g_2, g_3) \in H^1(\Omega_1)^3$. We have

$$\begin{aligned} \left(\frac{\partial}{\partial r}(g_i \circ \mathcal{P}) \right)(r,\alpha,z) &= \frac{\partial g_i}{\partial x_1}(\mathcal{P}(r,\alpha,z)) \cos\alpha + \frac{\partial g_i}{\partial x_2}(\mathcal{P}(r,\alpha,z)) \sin\alpha, \\ \left(\frac{\partial}{\partial \alpha}(g_i \circ \mathcal{P}) \right)(r,\alpha,z) &= -\frac{\partial g_i}{\partial x_1}(\mathcal{P}(r,\alpha,z)) r \sin\alpha + \frac{\partial g_i}{\partial x_2}(\mathcal{P}(r,\alpha,z)) r \cos\alpha, \\ \left(\frac{\partial}{\partial z}(g_i \circ \mathcal{P}) \right)(r,\alpha,z) &= \frac{\partial g_i}{\partial x_3}(\mathcal{P}(r,\alpha,z)), \qquad i = 1,2,3. \end{aligned} \tag{3.11}$$

We define a vector-function $v = (v_1, v_2, v_3)$ as follows:

$$\begin{aligned} v_1 &= (g_1 \circ \mathcal{P}) \cos\alpha + (g_2 \circ \mathcal{P}) \sin\alpha, \\ v_2 &= (g_2 \circ \mathcal{P}) \cos\alpha - (g_1 \circ \mathcal{P}) \sin\alpha, \quad v_3 = g \circ \mathcal{P}. \end{aligned} \tag{3.12}$$

It follows from (3.7), (3.11), and (3.12) that $v = G^{-1} g \in H_0$, where G^{-1} is the inverse of G. Therefore, $G(H_0) = H^1(\Omega_1)^3$.

The equalities (3.11) imply $G^{-1} \in \mathcal{L}(H^1(\Omega_1)^3, H_0)$, and by the Banach theorem on closed range the operator G is an isomorphism of H_0 onto $H^1(\Omega_1)^3$. ∎

Everywhere below, we use the following notations: if $\mathcal{H}$ is a normed space, we denote by $\mathcal{H}^*$ the dual of $\mathcal{H}$ and by (f,h) the duality between $\mathcal{H}^*$ and $\mathcal{H}$, where $f \in \mathcal{H}^*$, $h \in \mathcal{H}$. In particular, if $f \in Y$ or $f \in Y^n$, $n = 2$ or 3, then (f,h) is the scalar product in Y or in Y^n, respectively. That is, we identify the spaces Y and Y^n with their dual spaces Y^* and $(Y^n)^*$, respectively.

The sign $\rightharpoonup$ denotes weak convergence in a Banach space.

We suppose that $\hat{u}$ belongs to the space of traces on S_1 of the functions from H_0, i.e., $\hat{u} \circ \mathcal{P}^{-1} \in H^{\frac{1}{2}}(\mathcal{P}(S_1))$. Then, there exists a function $\tilde{u}$ such that

$$\tilde{u} \in H_0, \qquad \tilde{u}\big|_{S_1} = \hat{u}, \qquad \operatorname{div}_c \tilde{u} = 0. \tag{3.13}$$

We assume also that

$$K = (K_1, K_2, K_3) \in Y^3, \qquad F \circ \mathcal{P}^{-1} = (F_1, F_2, F_3) \circ \mathcal{P}^{-1} \in L_2(\mathcal{P}(S_2))^3. \tag{3.14}$$

We define operators $L : H \to H^*$ and $B \in \mathcal{L}(H, Y^*)$ as follows:

$$(L(v), h) = 2 \int_\Omega \varphi \epsilon_{ij}(\tilde{u} + v) \epsilon_{ij}(h) r \, dr \, d\alpha \, dz, \qquad v, h \in H, \tag{3.15}$$

$$(Bv, w) = \int_\Omega (\operatorname{div}_c v) w r \, dr \, d\alpha \, dz, \qquad v \in H, \ w \in Y. \tag{3.16}$$

In (3.15) the function φ is defined either by (1.7) or by (1.17).

We consider the problem: Find a pair of functions (v,p) satisfying

$$v \in H, \qquad p \in Y, \tag{3.17}$$

$$(L(v), h) - (B^* p, h) = (K + F, h), \qquad h \in H, \tag{3.18}$$

$$(Bv, w) = 0, \qquad w \in Y. \tag{3.19}$$

Here, B^* is the operator adjoint of B and

$$(K + F, h) = \int_\Omega K_i h_i \, r \, dr \, d\alpha \, dz + \int_{S_2} F_i h_i \, ds. \tag{3.20}$$

The pair $(u = v + \tilde{u}, p)$, where (v,p) is a solution of the problem (3.17)–(3.19), will be called the generalized solution of the problem (2.6)–(2.11), (2.13)–(2.15).

Indeed, by use of Green's formula, it can be seen that, if (v,p) is a solution of the problem (3.17)–(3.19), then the pair $(u = v + \tilde{u}, p)$ is a solution of the problem (2.6)–(2.11), (2.13)–(2.15) in the distributional sense. On the contrary, if (u,p) is a smooth solution of the problem (2.6)–(2.11), (2.13)–(2.15), then the pair $(v = u - \tilde{u}, p)$ is a solution of the problem (3.17)–(3.19).

LEMMA 3.2. *Suppose that the condition* (A0) *is satisfied. Then, the following inf-sup condition,*

$$\inf_{g \in Y} \sup_{w \in H} \frac{(Bw, g)}{||w||_H ||g||_Y} \geq \beta_1 > 0, \tag{3.21}$$

holds true. The operator B is an isomorphism from $H_1^\perp$ onto Y, where $H_1^\perp$ is the orthogonal complement of H_1 in H, and the operator B^* is an isomorphism from Y onto the polar set

$$H_1^\circ = \{q \in H^*, \quad (q, v) = 0, \quad v \in H_1\}. \tag{3.22}$$

Moreover

$$||B^{-1}||_{\mathcal{L}(Y,H_1^\perp)} \le \frac{1}{\beta_1}, \qquad ||(B^*)^{-1}||_{\mathcal{L}(H_1^\circ,Y)} \le \frac{1}{\beta_1}. \tag{3.23}$$

This lemma follows from the corresponding result in Cartesian coordinates (see [2], [8], [13]), since G is an isomorphism of H_0 onto $H^1(\Omega_1)^3$ and $I(v)$ is invariant (see (3.10)) and $\operatorname{div}_c v$ is an invariant also; i.e.,

$$(\operatorname{div}_c v)(r,\alpha,z) = (\operatorname{div} G(v))(\mathcal{P}(r,\alpha,z)). \tag{3.24}$$

4. Approximate solutions and existence theorems. Let $\{X_m\}_{m=1}^\infty$ and $\{N_m\}_{m=1}^\infty$ be sequences of finite-dimensional subspaces in H and Y, respectively, such that

$$\lim_{m\to\infty} \inf_{h\in X_m} ||w-h||_H = 0, \qquad w \in H, \tag{4.1}$$

$$\lim_{m\to\infty} \inf_{g\in N_m} ||f-g||_Y = 0, \qquad f \in Y, \tag{4.2}$$

$$\inf_{g\in N_m} \sup_{h\in X_m} \frac{(Bh,g)}{||h||_H||g||_Y} \ge \beta > 0, \tag{4.3}$$

$$X_m \subset X_{m+1}, \qquad N_m \subset N_{m+1}, \qquad m \in \mathbb{N}. \tag{4.4}$$

We seek an approximate solution of the problem (3.17)–(3.19) of the form

$$v_m \in X_m, \qquad p_m \in N_m, \tag{4.5}$$

$$(L(v_m),h) - (B^* p_m, h) = (K+F,h), \qquad h \in X_m, \tag{4.6}$$

$$(Bv_m, g) = 0, \qquad g \in N_m. \tag{4.7}$$

THEOREM 4.1. *Suppose that the function φ defining the operator L (see* (3.15)*) is given by* (1.7) *and that the conditions* (A0), (A1), (A2), (2.24) *are satisfied. Let also* (3.13), (3.14), (4.1)–(4.4) *hold. Then there exists a solution (v,p) of the problem* (3.17)–(3.19)*, and for an arbitrary $m \in \mathbb{N}$ there exists a solution of the problem* (4.5)–(4.7)*, and a subsequence $\{v_k, p_k\}$ can be extracted from the sequence $\{v_m, p_m\}$ such that*

$$v_k \to v \quad \text{in } H, \qquad p_k \to p \quad \text{in } Y. \tag{4.8}$$

Indeed, we replace cylindrical coordinates r, α, z by Cartesian coordinates x_1, x_2, x_3 in the problems (3.17)–(3.19) and (4.5)–(4.7). Then, we use Lemma 3.1 and Theorem 5.1 from [8] for these problems in Cartesian coordinates and pass back to cylindrical coordinates. As a result, we obtain that there exists a solution of the problem (3.17)–(3.19), and there exists a solution of the problem (4.5)–(4.7) for any $m \in \mathbb{N}$, and a subsequence $\{v_k, p_k\}$ can be extracted from the sequence $\{v_m, p_m\}$ such that

$$v_k \rightharpoonup v \quad \text{in } H, \qquad p_k \rightharpoonup p \quad \text{in } Y. \tag{4.9}$$

(4.8) is proved by using (4.9) and the arguments of Theorem 2.1 from [2].

The next theorem follows from the results of [2].

THEOREM 4.2. *Suppose that the function φ defining the operator L (see* (3.15)*) is given by* (1.17) *and that the conditions* (A0), (A3), (2.23), (2.24) *are satisfied. Let also* (3.13), (3.14), (4.1)–(4.4) *hold. Then there exists a unique solution (u,p) of the problem* (3.17)–(3.19)*, and for an arbitrary $m \in \mathbb{N}$ there exists a unique solution (v_m, p_m) of the problem* (4.5)–(4.7)*; moreover,*

$$v_m \to v \quad in \quad H, \qquad p_m \to p \quad in \quad Y.$$

5. Problem of axially symmetric flow.

5.1. Formulation of the problem. In the case of axially symmetric flow the components of a velocity vector u in the mobile orthonormal basis (e_r, e_α, e_z) of the cylindrical coordinate system depend on r, z only, i.e., $u(r,z) = (u_1(r,z), u_2(r,z), u_3(r,z))$; the components of the rate-of-strain tensor have the form

$$\epsilon_{11}(u) = \frac{\partial u_1}{\partial r}, \quad \epsilon_{22}(u) = \frac{u_1}{r}, \quad \epsilon_{33}(u) = \frac{\partial u_3}{\partial z},$$

$$\epsilon_{12}(u) = \epsilon_{21}(u) = \frac{1}{2}\left(\frac{\partial u_2}{\partial r} - \frac{u_2}{r}\right), \quad \epsilon_{23}(u) = \epsilon_{32}(u) = \frac{1}{2}\frac{\partial u_2}{\partial z},$$

$$\epsilon_{13}(u) = \epsilon_{31}(u) = \frac{1}{2}\left(\frac{\partial u_1}{\partial z} + \frac{\partial u_3}{\partial r}\right); \tag{5.1}$$

and the second invariant of the rate-of-strain tensor is defined by

$$I(u) = \left(\frac{\partial u_1}{\partial r}\right)^2 + \left(\frac{u_1}{r}\right)^2 + \left(\frac{\partial u_3}{\partial z}\right)^2 + \frac{1}{2}\left(\frac{\partial u_2}{\partial r} - \frac{u_2}{r}\right)^2$$

$$+ \frac{1}{2}\left(\frac{\partial u_2}{\partial z}\right)^2 + \frac{1}{2}\left(\frac{\partial u_1}{\partial z} + \frac{\partial u_3}{\partial r}\right)^2. \tag{5.2}$$

We assume that the domain of flow of the electrorheological fluid Ω_1 satisfies the condition (A0) and has the following form:

$$\Omega_1 = \{x | x = (x_1, x_2, x_3),\ x_3 \in (0,l),\ (x_1^2 + x_2^2)^{\frac{1}{2}} < R_2(x_3),$$

$$(x_1^2 + x_2^2)^{\frac{1}{2}} > R_1(x_3) \text{ if } R_1(x_3) > 0,\ (x_1^2 + x_2^2)^{\frac{1}{2}} \geq R_1(x_3) \text{ if } R_1(x_3) = 0\}, \tag{5.3}$$

where R_1 and R_2 are functions given in $(0,l)$. The function R_1 takes nonnegative values, R_2 takes positive values, and $R_2(x_3) > R_1(x_3)$ for all $x_3 \in (0,l)$.

The condition (A0) imposes restrictions on the functions R_1 and R_2. The functions R_1 and R_2 can be Lipschitz continuous as well as discontinuous with a finite number of points of discontinuity. But in the second case the functions R_1 and R_2 must be Lipschitz continuous in between the points of discontinuity.

Let $\Omega_2 = \mathcal{P}^{-1}(\Omega_1)$; the mapping $\mathcal{P}^{-1}$ is defined by (2.4). Since the flow of the fluid is assumed to be axially symmetric—i.e., the functions of velocity, pressure, and electric field are independent of α in cylindrical coordinate system—we consider our problem in the domain Ω_3, which consists of points (r,z) such that $(r,\alpha,z) \in \Omega_2$, $\alpha \in [0, 2\pi)$.

According to (5.3), the domain Ω_3 is defined by

$$\Omega_3 = \{(r,z) | 0 < z < l,\ R_1(z) < r \text{ if } R_1(z) > 0,$$

$$R_1(z) \leq r \text{ if } R_1(z) = 0, \quad r < R_2(z)\}. \tag{5.4}$$

We consider the stationary flow problem under the neglect of the inertial forces. Taking into account (2.6)–(2.8) and (5.1), we obtain the following motion equations:

(5.5)
$$\frac{\partial p}{\partial r} - 2\frac{\partial}{\partial r}\left(\varphi\frac{\partial u_1}{\partial r}\right) - \frac{\partial}{\partial z}\left(\varphi\left(\frac{\partial u_1}{\partial z} + \frac{\partial u_3}{\partial r}\right)\right) - \frac{2}{r}\varphi\left(\frac{\partial u_1}{\partial r} - \frac{u_1}{r}\right) = K_1 \quad \text{in } \Omega_3,$$

(5.6)
$$-\frac{\partial}{\partial r}\left(\varphi\left(\frac{\partial u_2}{\partial r} - \frac{u_2}{r}\right)\right) - \frac{\partial}{\partial z}\left(\varphi\frac{\partial u_2}{\partial z}\right) - \frac{2}{r}\varphi\left(\frac{\partial u_2}{\partial r} - \frac{u_2}{r}\right) = K_2 \quad \text{in } \Omega_3,$$

(5.7)
$$\frac{\partial p}{\partial z} - \frac{\partial}{\partial r}\left(\varphi\left(\frac{\partial u_1}{\partial z} + \frac{\partial u_3}{\partial r}\right)\right) - 2\frac{\partial}{\partial z}\left(\varphi\frac{\partial u_3}{\partial z}\right) - \frac{\varphi}{r}\left(\frac{\partial u_1}{\partial z} + \frac{\partial u_3}{\partial r}\right) = K_3 \quad \text{in } \Omega_3,$$

where the function φ is defined either by (1.7) or by (1.17).

The equation of incompressibility takes the form

$$\operatorname{div}_c u = \frac{\partial u_1}{\partial r} + \frac{\partial u_3}{\partial z} + \frac{u_1}{r} = 0 \quad \text{in } \Omega_3. \tag{5.8}$$

Let S be the boundary of Ω_3 and

$$\begin{aligned} &\mathcal{T} = \{z | z \in (0,l),\ R_1(z) = 0\}, \\ &S_0 = \{(r,z) | r = 0,\ z \in \mathcal{T}\}. \end{aligned} \tag{5.9}$$

In particular, S_0 can be an empty set.

Let also

$$S' = \{(r,\alpha,z) | (r,z) \in S\backslash S_0, \quad \alpha \in [0,2\pi)\}. \tag{5.10}$$

Then $\mathcal{P}(S') = \Gamma$, where Γ is the boundary of the domain Ω_1 defined by (5.3).

Suppose that S_1 and S_2 are open subsets of $S\backslash S_0$ such that S_1 is not empty, $S_1 \cap S_2 = \emptyset$, and $\overline{S}_1 \cup \overline{S}_2 = S\backslash S_0$. We consider mixed boundary conditions, wherein velocities are specified on S_1 and surface forces are given on S_2, i.e.,

$$u|_{S_1} = \check{u}, \tag{5.11}$$

$$[(-p + 2\varphi\epsilon_{11}(u))\nu_1 + 2\,\varphi\epsilon_{13}(u)\nu_3]|_{S_2} = F_1, \tag{5.12}$$

$$[2\varphi\epsilon_{21}(u)\nu_1 + 2\varphi\epsilon_{23}(u)\nu_3]|_{S_2} = F_2, \tag{5.13}$$

$$[(-p + 2\varphi\epsilon_{33}(u))\nu_3 + 2\varphi\epsilon_{31}(u)\nu_1]|_{S_2} = F_3, \tag{5.14}$$

where ν_1 and ν_3 are the components of the unit outward normal $\nu = (\nu_1, 0, \nu_3)$ to the boundary S'. By analogy with the above (see (2.13)–(2.15)), we obtain the following boundary conditions on S_0:

$$\begin{aligned} &u_1\big|_{S_0} = 0, \qquad u_2\big|_{S_0} = 0, \qquad \frac{\partial u_3}{\partial r}\Big|_{S_0} = 0, \\ &\lim_{r\to 0}\left(\frac{\partial u_1}{\partial r} - \frac{u_1}{r}\right)(r,z) = 0, \qquad z \in \mathcal{T}, \\ &\lim_{r\to 0}\left(\frac{\partial u_2}{\partial r} - \frac{u_2}{r}\right)(r,z) = 0, \qquad z \in \mathcal{T}. \end{aligned} \tag{5.15}$$

5.2. Functional spaces and two lemmas. We introduce the following sets:

$$\begin{aligned} &\mathcal{J}_0 = \{v | v = (v_1, v_2, v_3) \in C^\infty(\overline{\Omega}_3)^3,\ v_1\big|_{S_0} = 0,\ v_2\big|_{S_0} = 0\}, \\ &\mathcal{J} = \{v | v \in \mathcal{J}_0,\ v = 0 \quad \text{on } S_1\}, \\ &\mathcal{J}_1 = \{v | v \in \mathcal{J},\ \operatorname{div}_c v = 0\}, \end{aligned} \tag{5.16}$$

where the operator div_c is defined by (5.8).

We denote by $\mathcal{H}$ and $\mathcal{H}_1$ the closures of $\mathcal{J}$ and $\mathcal{J}_1$ with respect to the norm

$$||v||_{\mathcal{H}} = \left(\int_{\Omega_3} I(v) r \, dr \, dz \right)^{\frac{1}{2}}, \tag{5.17}$$

where $I(v)$ is given by (5.2).

Let $\mathcal{H}_0$ be the closure of $\mathcal{J}_0$ relative to the norm

$$||v||_{\mathcal{H}_0} = \left(||v||^2_{\mathcal{H}} + \int_{S_1} |v|^2 \, ds \right)^{\frac{1}{2}}, \tag{5.18}$$

where $ds = (dz^2 + dr^2)^{\frac{1}{2}}$.

Let also $\mathcal{Y}$ be the space of scalar functions which are square integrable in Ω_3 with respect to the measure $r \, dr \, dz$. The norm in $\mathcal{Y}$ is defined by

$$||h||_{\mathcal{Y}} = \left(\int_{\Omega_3} h^2 r \, dr \, dz \right)^{\frac{1}{2}}. \tag{5.19}$$

By analogy with the Lemma 3.1, we obtain the following result.

LEMMA 5.1. *Suppose that the domain Ω_1 defined by* (5.3) *satisfies the condition* (A0). *Then, the expressions* (5.17) *and* (3.8) *define equivalent norms $\mathcal{H}$, and the expressions* (5.18) *and* (3.8) *are equivalent norms in $\mathcal{H}_0$; moreover, the following equality holds:*

$$(2\pi)^{\frac{1}{2}} ||h||_{\mathcal{Y}} = ||\tilde{h} \circ \mathcal{P}^{-1}||_{L_2(\Omega_1)}, \tag{5.20}$$

with $\tilde{h}(r, \alpha, z) = h(r, z)$, $\alpha \in [0, 2\pi)$.

LEMMA 5.2. *Suppose that the domain Ω_1 defined by* (5.3) *satisfies the condition* (A0). *Denote by $\mathcal{B}$ the operator* div_c *acting in the space $\mathcal{H}$, i.e.,*

$$\mathcal{B}v = \frac{\partial v_1}{\partial r} + \frac{\partial v_3}{\partial z} + \frac{v_1}{r}. \tag{5.21}$$

Then, the following inf-sup condition,

$$\inf_{g \in \mathcal{Y}} \sup_{v \in \mathcal{H}} \frac{(\mathcal{B}v, g)}{||v||_{\mathcal{H}} ||g||_{\mathcal{Y}}} \geq \beta_2 > 0, \tag{5.22}$$

holds true.

The operator $\mathcal{B}$ is an isomorphism from $\mathcal{H}_1^{\perp}$ onto $\mathcal{Y}$, where $\mathcal{H}_1^{\perp}$ is the orthogonal complement of $\mathcal{H}_1$ in $\mathcal{H}$, and the operator $\mathcal{B}^*$ that is the adjoint of $\mathcal{B}$ is an isomorphism from $\mathcal{Y}$ onto the polar set

$$\mathcal{H}_1^{\circ} = \{ q \in \mathcal{H}^*, \quad (q, v) = 0, \quad v \in \mathcal{H}_1 \}. \tag{5.23}$$

Moreover,

$$||\mathcal{B}^{-1}||_{\mathcal{L}(\mathcal{Y}, \mathcal{H}_1^{\perp})} \leq \frac{1}{\beta_2}, \qquad ||(\mathcal{B}^*)^{-1}||_{\mathcal{L}(\mathcal{H}_1^{\circ}, \mathcal{Y})} \leq \frac{1}{\beta_2}. \tag{5.24}$$

Lemma 5.2 does not follow from Lemma 3.2. For the proof of Lemma 5.2, see [14].

5.3. Generalized solution. We suppose that the function $\breve{u}$ (see (5.11)) belongs to the traces on S_1 of the functions from $\mathcal{H}_0$. Then there exists a function $\overset{*}{u}$ such that

$$\overset{*}{u} \in \mathcal{H}_0, \qquad \overset{*}{u}\big|_{S_1} = \breve{u}, \qquad \mathrm{div}_c \overset{*}{u} = 0. \tag{5.25}$$

We assume also that

$$K = (K_1, K_2, K_3) \in \mathcal{Y}^3, \qquad F = (F_1, F_2, F_3) \in L_2(S_2)^3. \tag{5.26}$$

Define an operator $\mathcal{M} : \mathcal{H} \to \mathcal{H}^*$ as follows:

$$(\mathcal{M}(v), h) = 2 \int_{\Omega_3} \varphi \epsilon_{ij}(\overset{*}{u} + v)\epsilon_{ij}(h) r \, dr \, dz, \qquad v, h \in \mathcal{H}, \tag{5.27}$$

where the function φ is given either by (1.7) or by (1.17), and $\epsilon_{ij}(v)$ are defined by (5.1). We consider the problem: Find a pair of functions (v, p) satisfying

$$v \in \mathcal{H}, \qquad p \in \mathcal{Y}, \tag{5.28}$$

$$(\mathcal{M}(v), h) - (\mathcal{B}^* p, h) = (K + F, h), \qquad h \in \mathcal{H}, \tag{5.29}$$

$$(\mathcal{B}v, w) = 0, \qquad w \in \mathcal{Y}. \tag{5.30}$$

Here $\mathcal{B}^*$ is the operator adjoint of $\mathcal{B}$ and

$$(K + F, h) = \int_{\Omega_3} K_i h_i r \, dr \, dz + \int_{S_2} F_i h_i \, ds. \tag{5.31}$$

The pair $(u = \overset{*}{u} + v, p)$, where (v, p) is a solution of problem (5.28)–(5.30), will be called the generalized solution of the problem (5.5)–(5.8), (5.11)–(5.15).

Let $\{\mathcal{X}_m\}_{m=1}^{\infty}$ and $\{\mathcal{N}_m\}_{m=1}^{\infty}$ be sequences of finite-dimensional subspaces in $\mathcal{H}$ and $\mathcal{Y}$, respectively, such that

$$\lim_{m\to\infty} \inf_{h \in \mathcal{X}_m} ||w - h||_{\mathcal{H}} = 0, \qquad w \in \mathcal{H}, \tag{5.32}$$

$$\lim_{m\to\infty} \inf_{g \in \mathcal{N}_m} ||f - g||_{\mathcal{Y}} = 0, \qquad f \in \mathcal{Y}, \tag{5.33}$$

$$\inf_{g \in \mathcal{N}_m} \sup_{h \in \mathcal{X}_m} \frac{(\mathcal{B}h, g)}{||h||_{\mathcal{H}} ||g||_{\mathcal{Y}}} \geq \beta > 0, \tag{5.34}$$

$$\mathcal{X}_m \subset \mathcal{X}_{m+1}, \qquad \mathcal{N}_m \subset \mathcal{N}_{m+1}, \qquad m \in \mathbb{N}. \tag{5.35}$$

We seek an approximate solution of the problem (5.28)–(5.30) of the form

$$v_m \in \mathcal{X}_m, \qquad p_m \in \mathcal{N}_m, \tag{5.36}$$

$$(\mathcal{M}(v_m), h) - (\mathcal{B}^* p_m, h) = (K + F, h), \qquad h \in \mathcal{X}_m, \tag{5.37}$$

$$(\mathcal{B}\, v_m, g) = 0, \qquad g \in \mathcal{N}_m. \tag{5.38}$$

THEOREM 5.1. *Suppose that the function φ is given by* (1.7) *and that the conditions* (A1), (A2), (2.24) *are satisfied. Let the conditions* (A0), (5.3) *hold and Ω_3 be defined by* (5.4). *Assume also that* (5.25), (5.26), (5.32)–(5.35) *are fulfilled. Then there exists a solution v, p of the problem* (5.28)–(5.30), *and for an arbitrary $m \in \mathbb{N}$ there exists a solution of the problem* (5.36)–(5.38), *and a subsequence $\{v_k, p_k\}$ can be extracted from the sequence $\{v_m, p_m\}$ such that*

$$v_k \to v \quad \textit{in } \mathcal{H}, \qquad p_k \to p \quad \textit{in } \mathcal{Y}. \tag{5.39}$$

Indeed, by using the arguments of Theorem 5.1 from [8], we prove that there exists a solution of the problem (5.36)–(5.38) for any $m \in \mathbb{N}$, and a subsequence $\{v_k, p_k\}$ can be extracted from the sequence $\{v_m, p_m\}$ such that

$$v_k \rightharpoonup v \quad \text{in } \mathcal{H}, \qquad p_k \rightharpoonup p \quad \text{in } \mathcal{Y}. \tag{5.40}$$

(5.39) is proved by using (5.40) and the reasonings of the Theorem 2.1 from [2].

THEOREM 5.2. *Suppose that the function φ is given by* (1.17) *and that the conditions* (A0), (A3), (2.23), (2.24) *are satisfied. Let also* (5.25), (5.26), (5.32)–(5.35) *hold. Then, there exists a unique solution of the problem* (5.28)–(5.30), *and there exists a unique solution of the problem* (5.36)–(5.38) *for any $m \in \mathbb{N}$; in addition, $v_m \to v$ in $\mathcal{H}$, $p_m \to p$ in $\mathcal{Y}$.*

The proof of this theorem is analogous to the proof of Theorem 2.1 from [2].

6. Electrorheological clutch.

6.1. Problem on an electric field. Figure 1 (left) displays a scheme of an electrorheological clutch consisting of two coaxial cylinders. The gap between the cylinders is filled with an electrorheological fluid. The inner cylinder hosts a high voltage lead supplying the lateral surface, which serves as the electrode, whereas the lateral surface of the outer cylinder acts as the counter electrode.

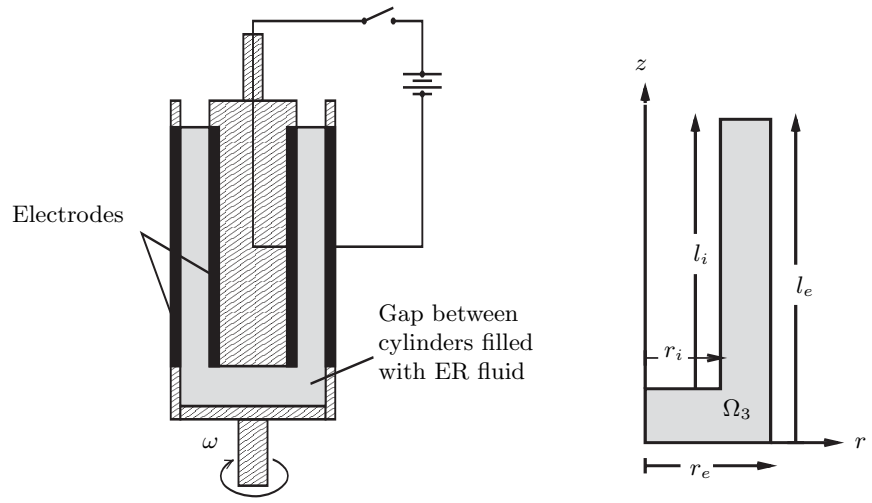


FIG. 1. *Simple model for an electrorheological fluid clutch (left) and the computational domain (right).*

By applying a voltage, one enhances the viscosity of the fluid. Under sufficiently large voltage the inner and external cylinders are almost rigidly bound and rotate practically at the same angular velocity. By varying the voltage, one obtains various slippage of the cylinders, i.e., various transmission ratio of the clutch.

The flow in the clutch is axially symmetric. According to Figure 1 (left) and (5.3), (5.4), the domain Ω_3 corresponding to the flow in the clutch has the form shown in Figure 1 (right).

The vector function of electric field E is defined as $E = -\operatorname{grad}\theta$, where θ is the function of the electric potential that meets the following equation:

$$\operatorname{div}(\chi \operatorname{grad}\theta) = 0, \tag{6.1}$$

where χ is the dielectric permittivity. In our case $\operatorname{grad}\theta = (\frac{\partial\theta}{\partial r}, \frac{\partial\theta}{\partial z})$, and (6.1) takes the form

$$\frac{\partial}{\partial r}\left(\chi\frac{\partial\theta}{\partial r}\right) + \frac{\chi}{r}\frac{\partial\theta}{\partial r} + \frac{\partial}{\partial z}\left(\chi\frac{\partial\theta}{\partial z}\right) = 0 \quad \text{in} \quad \Omega_3, \tag{6.2}$$

and θ satisfies the following boundary conditions:

$$\theta = U \quad \text{on } D_1, \qquad \theta = 0 \quad \text{on } D_2, \qquad \frac{\partial\theta}{\partial r} = 0 \quad \text{on } S_0,$$
$$\nu_1\chi\frac{\partial\theta}{\partial r} + \nu_3\chi\frac{\partial\theta}{\partial z} = 0 \quad \text{on } S\backslash(\overline{D}_1 \cup \overline{D}_2 \cup \overline{S}_0). \tag{6.3}$$

Here, $U = \text{constant} > 0$, S is the boundary of Ω_3, and

$$D_1 = \{(r,z)|\, r = r_i,\ z \in (l_e - l_i, l_e)\},$$
$$D_2 = \{(r,z)|\, r = r_e,\ z \in (l_e - l_i, l_e)\},$$
$$S_0 = \{(r,z)|\, r = 0,\ z \in (0, l_e - l_i)\}. \tag{6.4}$$

Let

$$\mathcal{Z} = \left\{w | w \in C^\infty(\overline{\Omega}_3),\ \frac{\partial w}{\partial r} = 0 \quad \text{on } S_0\right\}, \tag{6.5}$$

and let Z_0 be the closure of $\mathcal{Z}$ with respect to the norm

$$||w||_{Z_0} = \left(\int_{\Omega_3}\left[w^2 + \left(\frac{\partial w}{\partial r}\right)^2 + \left(\frac{\partial w}{\partial z}\right)^2\right] r\,dr\,dz\right)^{\frac{1}{2}}. \tag{6.6}$$

Again, we consider the following space:

$$Z = \{w | w \in Z_0,\ w = 0 \quad \text{on} \quad D_1 \cup D_2\}. \tag{6.7}$$

The expression

$$||w||_Z = \left(\int_{\Omega_3}\left[\left(\frac{\partial w}{\partial r}\right)^2 + \left(\frac{\partial w}{\partial z}\right)^2\right] r\,dr\,dz\right)^{\frac{1}{2}} \tag{6.8}$$

defines a norm in Z being equivalent to the norm of Z_0 determined by (6.6). Let θ_0 be a function such that

$$\theta_0 \in Z_0, \quad \theta_0 = U \quad \text{on } D_1, \qquad \theta_0 = 0 \quad \text{on } D_2. \tag{6.9}$$

We assume that χ is a function that is integrable in Ω_3 with respect to the measure $r\,dr\,dz$, and in addition,

$$b_1 \geq \chi \geq b_0 > 0 \quad \text{a.e. in} \quad \Omega, \tag{6.10}$$

where b_0 and b_1 are positive constants.

Define a bilinear form $a : Z_0 \times Z \to \mathbb{R}$ as follows:

$$a(q,h) = \int_{\Omega_3}\chi\left(\frac{\partial q}{\partial r}\frac{\partial h}{\partial r} + \frac{\partial q}{\partial z}\frac{\partial h}{\partial z}\right) r\,dr\,dz, \quad q \in Z_0, \quad h \in Z. \tag{6.11}$$

Consider the following problem: Find θ_1 satisfying

$$\theta_1 \in Z, \qquad a(\theta_1, h) = -a(\theta_0, h), \qquad h \in Z. \tag{6.12}$$

The function $\theta = \theta_0 + \theta_1$ is a generalized solution of the problem (6.2), (6.3).

The Riesz theorem implies the following result.

THEOREM 6.1. *Suppose that* (6.9) *and* (6.10) *are satisfied. Then there exists a unique solution of the problem* (6.12)*, and there exists a unique generalized solution* θ *of the problem* (6.2), (6.3)*. The function* θ *is represented in the form* $\theta = \theta_0 + \theta_1$*, where* θ_0 *is a function satisfying* (6.9) *and* θ_1 *is the solution of the problem* (6.12)*.*

6.2. Problem on the fluid flow. We assume that the velocity vector u and the pressure p depend only on r, z in the mobile orthonormal basis e_r, e_α, e_z of cylindrical coordinate system r, α, z and, in addition, $u(r, z) = (0, u_2(r, z), 0)$. We denote the function u_2 by u.

According to (2.1), we have

$$\epsilon_{12}(u) = \epsilon_{21}(u) = \frac{1}{2}\left(\frac{\partial u}{\partial r} - \frac{u}{r}\right), \qquad \epsilon_{23}(u) = \epsilon_{32}(u) = \frac{1}{2}\,\frac{\partial u}{\partial z},$$

$$\epsilon_{11}(u) = \epsilon_{22}(u) = \epsilon_{33}(u) = \epsilon_{13}(u) = \epsilon_{31}(u) = 0, \tag{6.13}$$

and

$$I(u) = \frac{1}{2}\left(\frac{\partial u}{\partial r} - \frac{u}{r}\right)^2 + \frac{1}{2}\left(\frac{\partial u}{\partial z}\right)^2. \tag{6.14}$$

In line with (5.5)–(5.7), the motion equations take the form

$$\frac{\partial p}{\partial r} = \frac{\partial p}{\partial z} = 0, \tag{6.15}$$

$$\frac{\partial}{\partial r}\left(\varphi\left(\frac{\partial u}{\partial r} - \frac{u}{r}\right)\right) + \frac{\partial}{\partial z}\left(\varphi\frac{\partial u}{\partial z}\right) + \frac{2}{r}\varphi\left(\frac{\partial u}{\partial r} - \frac{u}{r}\right) = 0, \tag{6.16}$$

where volume force vector is ignored.

In the case under consideration, the condition of incompressibility (2.9) is satisfied.

We prescribe velocities on the surfaces of the internal and external cylinders S_1 and specify surface forces on the top boundary of the electrorheological fluid S_2. In this case, we have (see Figure 1 (right))

$$S_1 = \bigcup_{i=1}^{4} S_{1i}, \tag{6.17}$$

where

$$S_{11} = \{(r, z) | z = 0,\ r \in (0, r_e)\}, \qquad S_{12} = \{(r, z) | r = r_e,\ z \in (0, l_e)\},$$

$$S_{13} = \{(r, z) | z = l_e - l_i,\ r \in (0, r_i)\}, \qquad S_{14} = \{(r, z) | r = r_i,\ z \in ((l_e - l_i), l_e)\}, \tag{6.18}$$

and

$$S_2 = \{(r, z) | z = l_e,\ r \in (r_i, r_e)\}. \tag{6.19}$$

In the case that the inner cylinder is leading, we deal with the following boundary conditions:

$$u(r,z) = \begin{cases} 0 & \text{on } S_{11} \cup S_{12} \cup S_0, \\ \omega r & \text{on } S_{13}, \\ \omega r_i & \text{on } S_{14}, \end{cases} \tag{6.20}$$

$$\varphi \frac{\partial u}{\partial z} = 0 \quad \text{on } S_2, \qquad p = \tilde{c} \quad \text{on } S_2. \tag{6.21}$$

Here ω is the angular velocity of the internal cylinder, and we assume that $F_1 = F_2 = 0$, $F_3 = -\tilde{c}$, $\tilde{c} = \text{constant} > 0$; see (5.12)–(5.14). S_0 is given in (6.4), and according to (5.15), we have

$$u\big|_{S_0} = 0, \qquad \lim_{r \to 0} \left(\frac{\partial u}{\partial r} - \frac{u}{r} \right)(r,z) = 0, \qquad z \in (0, l_e - l_i). \tag{6.22}$$

In the case that the external cylinder is leading, we consider the boundary conditions of this type:

$$u = \begin{cases} \omega r & \text{on } S_{11}, \\ \omega r_e & \text{on } S_{12}, \\ 0 & \text{on } S_{13} \cup S_{14} \cup S_0, \end{cases} \tag{6.23}$$

where ω is the angular velocity of the external cylinder, and, in addition, (6.21) and (6.22) hold.

In the case under consideration the set $\mathcal{J}_0$ has the form

$$\mathcal{J}_0 = \{v | v \in C^\infty(\overline{\Omega}_3), \quad v = 0 \quad \text{on } S_0\}, \tag{6.24}$$

and $\mathcal{H}_0$ is the closure of $\mathcal{J}_0$ relative to the norm (compare with (5.16)–(5.18), (6.14))

$$||v||_{\mathcal{H}_0} = \left(\int_{\Omega_3} \left[v^2 + \left(\frac{\partial v}{\partial r} - \frac{v}{r} \right)^2 + \left(\frac{\partial v}{\partial z} \right)^2 \right] r \, dr \, dz \right)^{\frac{1}{2}}. \tag{6.25}$$

The space $\mathcal{H}$ appears as

$$\mathcal{H} = \{v | v \in \mathcal{H}_0, \, v = 0 \quad \text{on } S_1\}, \tag{6.26}$$

and the norm in $\mathcal{H}$ is given by

$$||v||_{\mathcal{H}} = \left(\int_{\Omega_3} \left[\left(\frac{\partial v}{\partial r} - \frac{v}{r} \right)^2 + \left(\frac{\partial v}{\partial z} \right)^2 \right] r \, dr \, dz \right)^{\frac{1}{2}}. \tag{6.27}$$

It follows from Lemma 5.1 that the expressions (6.25) and (3.8) with $v_1 = 0$, $v_2 = v$, $v_3 = 0$ define equivalent norms in $\mathcal{H}_0$, whereas (6.27) and (3.8) are equivalent norms in $\mathcal{H}$.

Equation (6.15) implies $p = c = \text{constant}$, and (6.21) yields $c = \tilde{c}$.

Let $\overset{*}{u}$ be a function from $\mathcal{H}_0$ that satisfies either (6.20) or (6.23) according to which cylinder, inner or external, is leading.

The operator $\mathcal{M} : \mathcal{H} \to \mathcal{H}^*$ is defined as follows:

(6.28)
$$(\mathcal{M}(v), h) = \frac{1}{2} \int_{\Omega_3} \varphi \left[\left(\frac{\partial(\overset{*}{u} + v)}{\partial r} - \frac{\overset{*}{u} + v}{r} \right) \left(\frac{\partial h}{\partial r} - \frac{h}{r} \right) + \frac{\partial(\overset{*}{u} + v)}{\partial z} \frac{\partial h}{\partial z} \right] r \, dr \, dz.$$

In the case under consideration the velocity vector is orthogonal to the vector of the electric field at each point $(r, z) \in \overline{\Omega}_3$. Therefore, in (6.28) the function φ is defined by (see (1.7))

$$\varphi = b(|E|, 0)(\lambda + I(\overset{*}{u} + v))^{-\frac{1}{2}} + \psi(I(\overset{*}{u} + v), |E|, 0), \tag{6.29}$$

where the function I is given by (6.14).

We consider the following problem: Find a function v such that

$$v \in \mathcal{H}, \qquad (\mathcal{M}(v), h) = 0, \qquad h \in \mathcal{H}. \tag{6.30}$$

The pair $(u = \overset{*}{u} + v, p)$, where v is a solution of the problem (6.30) and $p = \tilde{c}$, is a generalized solution of the problem (6.15), (6.16), (6.21), (6.22), and (6.20) or (6.23).

Let $\{\mathcal{V}_m\}_{m=1}^{\infty}$ be a sequence of finite-dimensional subspaces in $\mathcal{H}$ such that

$$\lim_{m \to \infty} \inf_{h \in \mathcal{V}_m} ||w - h||_{\mathcal{H}} = 0, \qquad w \in \mathcal{H}, \tag{6.31}$$

$$\mathcal{V}_m \subset \mathcal{V}_{m+1}, \qquad m \in \mathbb{N}. \tag{6.32}$$

We define an approximate solution of the problem (6.30) of the form

$$v_m \in \mathcal{V}_m, \qquad (\mathcal{M}(v_m), h) = 0, \qquad h \in \mathcal{V}_m. \tag{6.33}$$

It follows from Theorem 5.2 that for the function φ defined by (6.29) there exists a unique solution of the problems (6.30) and (6.33); in addition, $v_m \to v$ in $\mathcal{H}$.

6.3. Simulation results. The nonlinear problem (6.30) is solved through solving a sequence of linear problems. Given $v^0 \in \mathcal{H}$, find $v^k \in \mathcal{H}$, $k = 1, 2, \ldots$, such that

$$d \in \mathcal{H}, \quad (\hat{\mathcal{M}}(v^{k-1})d, h) = -(\mathcal{M}(v^{k-1}), h) \quad \forall h \in \mathcal{H}, \tag{6.34}$$

$$v^k = v^{k-1} + \alpha \, d. \tag{6.35}$$

Here α is a relaxation parameter, and $\hat{\mathcal{M}}$ is the linearized version of the operator $\mathcal{M}$ (cf. [2]), defined as

$$\begin{aligned}(\hat{\mathcal{M}}(w)v, h) = \frac{1}{2} \int_{\Omega_3} & (b(|E|, 0)(\lambda + I(\overset{*}{u} + w))^{-\frac{1}{2}} + \psi(I(\overset{*}{u} + w), |E|, 0)) \\ & \times \left[\left(\frac{\partial(\overset{*}{u} + v)}{\partial r} - \frac{\overset{*}{u} + v}{r} \right) \left(\frac{\partial h}{\partial r} - \frac{h}{r} \right) + \frac{\partial(\overset{*}{u} + v)}{\partial z} \frac{\partial h}{\partial z} \right] r \, dr \, dz.\end{aligned} \tag{6.36}$$

Note that $(\hat{\mathcal{M}}(w)v, h) = (\mathcal{M}(v), h)$ whenever $w = v$. The algorithm can be termed the Birger–Kachanov method with relaxation; see [5] for the analysis of the original Birger–Kachanov method.

We consider the electrorheological fluid called the Rheobay TP AI 3656, a product of Bayer [1]. The experimentally obtained flow curves (relating the shear stress to the shear rate) of this product, corresponding to different electric field strengths

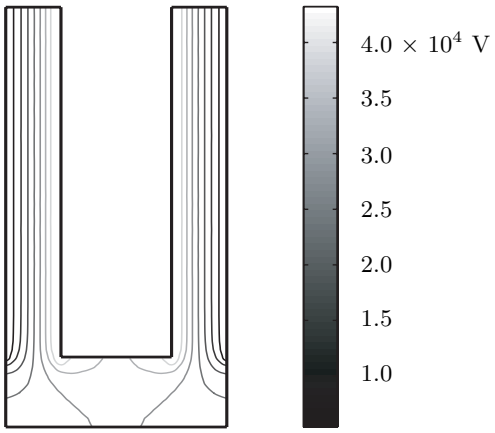


FIG. 2. *Contour plot of the electric potential: wide gap configuration.*

orthogonal to the velocities, have been approximated by cubic splines. The viscosity function φ is then calculated from these splines; see the appendix for details.

In order to understand the behavior of our electrorheological fluid in our model clutch, we study the flow in two different geometrical configurations of the clutch, the *wide*- and the *narrow* gap configurations. In the *wide* gap configuration, we take $r_i = 35$ mm, $r_e = 70$ mm, and $l_i = 250$ mm, $l_e = 300$ mm. During this test the cylinder (outer or inner, whichever is leading) rotates with an angular velocity of 125 rad sec^{-1}. For the *narrow* gap configuration, we take a much narrower gap between the cylinders by setting $r_i = 24$ mm and $r_e = 25$ mm. In this case $l_i = 25$ mm, $l_e = 30$ mm, and the angular velocity of the leading cylinder (outer or inner) is 5 rad sec^{-1}.

The function of the electric field potential θ was calculated approximately by using the Galerkin method with continuous and piecewise linear finite elements for the problem (6.12).

Figure 2 shows a contour plot of the electric potential calculated on the *wide* gap configuration for an applied voltage of 10 kV on the inner electrode. The distribution of this electric potential is linear along any cross-section inside the gap between the electrodes.

Angular velocity profiles for different applied voltages, calculated at one cross section of the gap, are shown in Figures 3 and 4.

From the calculations performed we arrive at the following conclusions:

1. The electric field $E = (E_r, E_z)$ in the gap between the cylinders is close to a constant vector $(U/(r_i - r_e), 0)$ for the narrow and wide gap configurations. At each point between the electrodes, with the exception of points in a very small zone by the ends of the electrodes, the electric field (E_r, E_z) tends to $(U/(r_i - r_e), 0)$ as $(r_e - r_i)/r_i$ tends to zero. The electric field decays sharply as the distance to the electrodes increases (see Figure 2).
2. In the case when the gap between the cylinders is wide and the outer cylinder is leading, a zone with a constant angular velocity is formed near the outer cylinder, and this zone increases with the increase of voltage (see Figure 3 left).
3. In the case when the gap between the cylinders is wide and the inner cylinder is leading, a zone with a constant angular velocity is formed near the outer

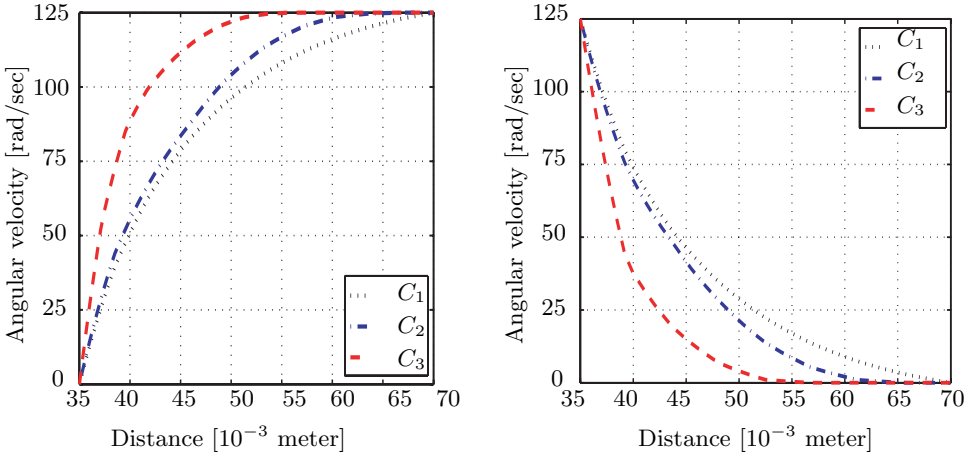


FIG. 3. *Angular velocity profile. Wide gap configuration with leading outer cylinder (left) and leading inner cylinder (right). The curves C_1, C_2, and C_3 correspond to $U = 0$ V, $U = 50$ kV, and $U = 100$ kV, respectively.*

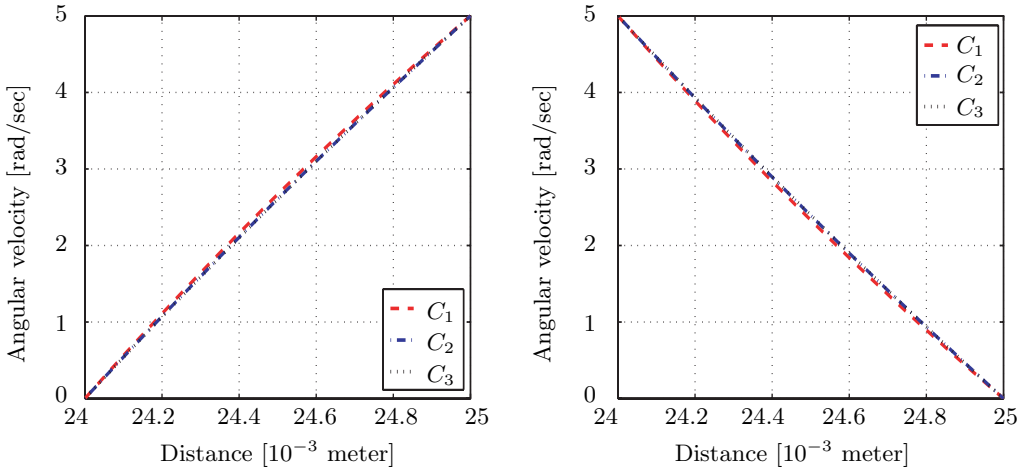


FIG. 4. *Angular velocity profile. Narrow gap configuration with leading outer cylinder (left) and leading inner cylinder (right). The curves C_1, C_2, and C_3 correspond to $U = 0$ V, $U = 2$ kV, and $U = 3$ kV, respectively.*

cylinder, as in the case when the outer cylinder is leading. This zone increases under the increase of voltage (see Figure 3 right).

4. In the case of a narrow gap between the cylinders, the zone with a constant angular velocity is not formed. The velocity profiles are almost linear at various voltages. No matter what cylinder is leading and what voltage is applied, the velocity profile tends to linear as $(r_e - r_i)/r_i$ tends to zero (see Figure 4). In this case essentially the velocity profile does not depend on the shape of a flow curve, and the shear rate is a constant.

We note that in the case of a wide gap between the cylinders, the zone with a constant angular velocity is also formed under the flow of the Bingham fluid. The proximity of the Bingham velocity profiles to the profiles presented in Figure 3

depends on the proximity of approximations of the flow curves by the affine functions $\tau_0+b_0\gamma$, where τ_0 and b_0 are the yield stress and the viscosity of the Bingham fluid and γ is the shear rate. (About the proximity of solutions for close flow curves, see [13, section 6.2].)

As may be seen from the appendix Figure A-1, one cannot obtain good approximations of the flow curves by affine functions, especially for small shear rates.

Appendix. Identification of the viscosity function. In the following, we present a set of cubic splines (flow splines) approximating a set of experimentally obtained flow curves and show how the viscosity function φ is calculated from these splines. These flow curves (splines) are for the electrorheological fluid called Rheobay TP AI 3656, a product of Bayer, based on a water-free dispersion of polymer particles in silicone oil (Baysilone Oil M); see [1] for specifications. The application of such a product can be found in various devices, such as shock absorbers, vibration dampers, clutches, and so on.

A.1. The flow splines. The set of cubic splines approximating experimentally obtained flow curves corresponding to a set of different electric field strengths, which are orthogonal to the velocity, are shown in Figure A-1. Complete information for the reconstruction of these splines, i.e., the sample points representing the shear rates γ, the data representing the shear stress τ, and the end slopes (derivatives), are provided in Table A-1.

Each flow curve has been approximated within the interval $[\gamma_0, \gamma_1]$ (in our case $\gamma_0 = 100\ \mathrm{sec}^{-2}$, $\gamma_1 = 2000\ \mathrm{sec}^{-2}$) by a cubic spline with the given end slopes. Outside of the interval $[\gamma_0, \gamma_1]$ the splines have been extended on $\mathbb{R}_+$ by straight lines (see the dotted lines in Figure A-1), so that the obtained function $\gamma \to \tau(\gamma)$ becomes continuously differentiable in $\mathbb{R}_+$.

A linear interpolation is used to calculate the function $\tau(\gamma)$ for values of $|E|$ intermediate between the values given in the Table A-1.

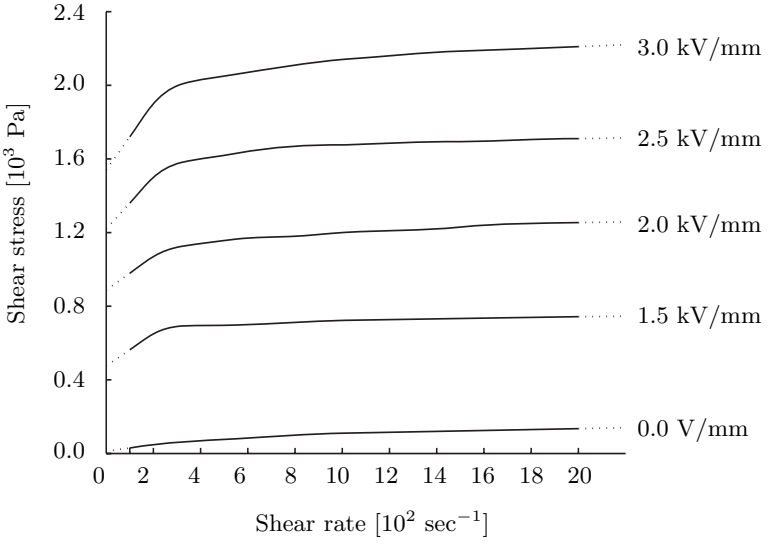


FIG. A-1. *Flow splines showing the effect of field strength (50Hz, AC) and shear rate γ on shear stress τ at 40°C. The cubic splines are constructed using the data provided in Table* A-1.

TABLE A-1

The table contains complete information for the reconstruction of the five cubic splines displayed in Figure A-1. *Each of the last five columns in the table corresponds to a spline approximating a flow curve, containing two end slopes and eleven data values corresponding to the eleven sample points in the first column of the table.*

	Shear stress (Pa)				
Shear rate γ [per sec]	0.0 V/mm	1.5 kV/mm	2.0 kV/mm	2.5 kV/mm	3.0 kV/mm
1.0×10^2	30.2	563.0	979.0	1360.0	1720.0
2.0×10^2	48.0	650.0	1070.0	1500.0	1900.0
4.0×10^2	69.3	695.0	1140.0	1600.0	2030.0
6.0×10^2	83.5	700.0	1170.0	1640.0	2070.0
8.0×10^2	100.0	712.0	1180.0	1670.0	2110.0
1.0×10^3	110.0	723.0	1200.0	1676.0	2140.0
1.2×10^3	115.0	727.0	1210.0	1686.0	2160.0
1.4×10^3	120.0	731.0	1220.0	1693.0	2180.0
1.6×10^3	125.0	735.0	1240.0	1696.0	2190.0
1.8×10^3	130.0	740.0	1250.0	1706.0	2200.0
2.0×10^3	135.0	743.0	1254.0	1710.0	2210.0
Slope at left end	0.180	0.870	0.910	1.400	1.800
Slope at right end	0.025	0.015	0.020	0.020	0.050

A.2. The viscosity function. We now show how the viscosity function φ is calculated from the function $\tau(\gamma)$. Let τ_0 be the point of intersection of a left dotted line with the shear stress axis. This dotted line is the continuation of the flow curve in the interval $[0, \gamma_0)$, and τ_0 is the yield stress.

The viscosity function, in the case of a simple shear flow, is determined as follows:

$$\varphi = \frac{1}{2}\frac{\tau}{\gamma}, \quad \text{where} \quad \gamma = \left(\frac{1}{2}I(u)\right)^{\frac{1}{2}}.$$

Generalizing it to an arbitrary flow, we get

$$b = \frac{\tau_0}{\sqrt{2}} \quad \text{and} \quad \psi = \frac{\tau - \tau_0}{2\gamma} = \frac{\tau - \tau_0}{(2I(u))^{\frac{1}{2}}}. \tag{A-1}$$

For a fixed value of $|E|$ and $I(u)$, we can thus find the value of φ using (A-1) and (6.29), i.e.,

$$\varphi = \frac{b}{(\lambda + I(u))^{\frac{1}{2}}} + \psi. \tag{A-2}$$

The parameter λ was chosen equal to $1.011e^{-11}$ sec^{-2}.

In the general case, when there are given flow curves for different values of $|E|$ and $\mu(u, E)$, one obtains expressions (A-2) for different values of $|E|$ and $\mu(u, E)$; see (1.7).

Acknowledgment. The authors are very thankful to the referees for useful comments.

REFERENCES

[1] Bayer AG, *Provisional Product Information of Rheobay TP AI* 3565 *and Rheobay TP AI* 3566, Technical Report AI 12601e, Bayer AG, Silicones Business Unit, Leverkusen, Germany, 1997.

[2] M. S. Belonosov and W. G. Litvinov, *Finite element method for nonlinearly viscous fluids*, Z. Angew. Math. Mech., 76 (1996), pp. 307–320.

[3] H. Block and J. P. Kelly, *Electro-rheology*, J. Phys. D: Appl. Phys., 21 (1988), pp. 1661–1677.

[4] F. Filisko, *Overview of ER technology*, in Progress in ER Technology, K. Havelka, ed., Plenum Press, New York, 1995.

[5] S. Fučik, A. Kratochvil, and J. Nečas, *Kachanov–Galerkin method*, Comment. Math. Univ. Carolinae, 14 (1973), pp. 651–659.

[6] H. Gajewski and K. Gröger, *Zacharias, Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*, Akademie-Verlag, Berlin, 1974.

[7] V. Girault and P. Raviart, *Finite Element Approximation of the Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.

[8] R. H. W. Hoppe and W. G. Litvinov, *Problems on electrorheological fluid flows*, Comm. Pure. Appl. Anal., 3 (2004), pp. 809–848.

[9] O. Ladyzhenskaya and V. Solonnikov, *Some problems of vector analysis and generalized formulation of boundary value problems for the Navier–Stokes equations*, Zap. Nauchn. Sem. Leningrad. Otdel Math. Inst. Steklov (LOMI), 59 (1976), pp. 81–116 (in Russian).

[10] L. D. Landau and E. M. Lifshitz, *Electrodynamics of Continuous Media*, Pergamon, Oxford, UK, 1984.

[11] J.-L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites non Linéaires*, Dunod Gauthier-Villars, Paris, 1969.

[12] W. G. Litvinov, *Motion of Nonlinearly Viscous Fluid*, Nauka, Moscow, 1982 (in Russian).

[13] W. G. Litvinov, *Optimization in Elliptic Problems with Applications to Mechanics of Deformable Bodies and Fluid Mechanics*, Birkhäuser Boston, Cambridge, MA, 2000.

[14] W. G. Litvinov, *A problem on non steady flow of a nonlinear viscous fluid in a deformable pipe*, Methods Funct. Anal. Topol., 2 (1996), pp. 85–113.

[15] J. Nečas and I. Hlavaček, *Mathematical Theory of Elastic and Elastico-Plastic Bodies: An Introduction*, Elsevier Scientific, Amsterdam, 1981.

[16] M. Parthasarathy and D. J. Kleingenberg, *Electrorheology: Mechanisms and models*, Material Sci. Engrg. 17 (1996), pp. 57–103.

[17] K. Rajagopal and A. Wineman, *Flow of electrorheological materials*, Acta Mechanica, 91 (1992), pp. 57–75.

[18] Z. P. Shulman and V. I. Kordonskii, *Magneto-rheological Effect*, Nauka i Technika, Minsk, 1982 (in Russian).

[19] Z. P. Shulman and B. M. Nosov, *Rotation of Nonconducting Bodies in Electrorheological Suspensions*, Nauka i Technika, Minsk, 1985 (in Russian).

[20] L. Schwartz, *Analyse Mathématique* 1, Hermann, Paris, 1967.

[21] M. M. Vainberg, *Variational Methods for the Study of Nonlinear Operators*, Holden Day, San Francisco, 1964.

[22] M. Whittle, R. J. Atkin, and W. A. Bullough, *Fluid dynamic limitations on the performance of an electrorheological clutch*, J. Non-Newtonian Fluid Mech., 57 (1995), pp. 61–81.

SIAM J. APPL. MATH.
Vol. 65, No. 5, pp. 1657–1676

DIFFUSION APPROXIMATION FOR LINEAR TRANSPORT WITH MULTIPLYING BOUNDARY CONDITIONS*

V. PROTOPOPESCU† AND L. THEVENOT‡

Abstract. We consider the diffusion limit of a suitably rescaled model transport equation in a slab with *multiplying boundary conditions*, as the scaling parameter ε tends to zero. We show that, for sufficiently smooth data, the solution converges in the L^2-norm for each $t > 0$ to the solution of a diffusion equation with *Robin boundary conditions* corresponding to an *incoming* flux. The derivation of the diffusive limit is based on an asymptotic expansion, which is rigorously justified.

Key words. diffusion approximation, transport equation, multiplying boundary condition, boundary layer, Milne problem, asymptotic expansion, spectral theory

AMS subject classifications. 35B25, 35B40, 35F15, 35C20, 35P05

DOI. 10.1137/S0036139903424242

1. Statement of the problem and main result. Traditionally, transport equations have been studied in media with either dissipative or conservative boundary conditions (BCs), while the possible multiplication of particles is confined to volume scattering effects (fission). Recently, multiplying BCs have attracted an interest on their own, first in connection with a model for cell population dynamics [R], and afterwards with the extension of general transport theory results [B1], [BE]. The cell population dynamics is described by an equation that shares strong formal similarities with the linear transport equation; in particular, multiplying BCs ensure the survivability of the cell population. For transport phenomena proper, multiplying BCs may occur in rarefied gas dynamics whereby, upon collision, molecules previously adsorbed at the boundaries may be liberated and reenter the medium. In heterogeneous reactors, with periodic structures of alternating slabs/rods/prisms of fissionable and moderating materials, one could model the effect of the fissionable material upon the moderating one as a multiplying boundary.

In general multidimensional geometries, multiplying BCs lead to difficulties, since the combined effect of geometry and net production of particles at the boundaries may result in the absence of overall control of the flux. In slab (one-dimensional) geometry, if the velocities are bounded, the growth effect can be controlled. This is also the case in general multidimensional geometries if the minimum travel time between two successive boundary collisions is bounded away from zero. Under these conditions, one can prove that the corresponding transport operator generates an exponentially bounded, positivity preserving evolution semigroup [BT], [MaT], [LaM], [B3].

In this paper, we consider the *diffusion limit* for the linear transport equation with *multiplying* BCs in slab geometry and provide the first proof of the existence

*Received by the editors March 5, 2003; accepted for publication (in revised form) December 6, 2004; published electronically July 13, 2005.
http://www.siam.org/journals/siap/65-5/42424.html

†Computer Science & Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6364 (protopopesva@ornl.gov). The research of this author was supported in part by the Division of Materials Science of the U.S. Department of Energy under contract DE-AC05-00OR22725 with UT-Battelle, LLC.

‡Department of Mathematics, Chalmers University of Technology, 41296 Göteborg, Sweden (thevenot@descartes.univ-fcomte.fr). The research of this author was supported in part by the HYKE Research Training Network financed by the European Union (HPRN-CT-2002-00282).

and precise meaning of such a limit. The derivation of diffusion (hydrodynamics) from transport with either dissipative or conservative BCs based on scaling arguments and/or asymptotic analysis has a rather long history. A more or less formal mathematical treatment can be found in [HM], [KL], [BLP], [Se], [Sa], [BSS], [DL2, Chap. XXI-5], [BM], [M1], [MPT], [MoT], and the references therein. However, when applied to *multiplying* BCs, the scaling method presents additional difficulties.

First, while conservative (dissipative) BCs for the transport equation result in Neumann (Dirichlet) BCs for the diffusion equation, without their own additional scaling, multiplying BCs have to be properly scaled: otherwise, in the limit of infinite time, they would lead to uncontrollable growth. To this end we introduce the scaled restitution coefficient at the surface $\alpha = 1 + \beta\varepsilon$ ($\beta > 0$), which means that we consider multiplying BCs which are almost conservative, and we prove that the corresponding transport semigroup is exponentially bounded in time, uniformly with respect to the scaling parameter ε. The time behavior is then controlled by adding an extra absorption term. We also study the spectral properties of the corresponding transport operator, which are related to the time asymptotic behavior of the transport semigroup. A direct estimate of the "leading eigenvalues," which could give a better understanding of the asymptotic behavior, is still missing.

The second difficulty is related to the boundary layer problem. Indeed, to carry out the analysis, we need to bound, uniformly in ε, the solution of the Milne problem with specularly reflecting BCs that arises from the boundary layer. The solvability condition of this Milne problem yields the BCs for the diffusion problem.

Finally, the proof of the convergence of the transport solution to the solution of the diffusive equation is based on rigorous estimates of the various terms in the asymptotic expansion. The interest of the asymptotic expansions is that the method permits one to solve the boundary layer problem. However, representing the solution of the transport equation as an expansion in the scaling parameter ε requires certain regularity of the solution of the diffusion equation and thereby of the data of the transport equation.

We state now the problem and the *main result*. Consider the linear transport equation

$$\text{Pb(1.1)} \quad \begin{cases} \dfrac{\partial f_\varepsilon}{\partial t}(x,\mu,t) = -\varepsilon^{-1}\mu\dfrac{\partial f_\varepsilon}{\partial x} - \varepsilon^{-2}Cf_\varepsilon - \gamma f_\varepsilon + S \quad \text{on } (-a,a)\times(-1,1)\times\mathbb{R}^+, \\ f_\varepsilon(-a,\mu,t) = (1+\beta\varepsilon)\ f_\varepsilon(-a,-\mu,t) + \varepsilon S^-(\mu), \quad \mu > 0, \\ f_\varepsilon(a,-\mu,t) = (1+\beta\varepsilon)\ f_\varepsilon(a,\mu,t) + \varepsilon S^+(\mu), \qquad \mu > 0, \\ f_\varepsilon(x,\mu,0) = f_0(x,\mu) \end{cases}$$

as an initial-boundary value problem in the space $L^2\big((-a,a)\times(-1,1)\big)$. Here γ and β are two positive constants, f_0 and S are in $L^2\big((-a,a)\times(-1,1)\big)$, and $S^+ \in L^2(0,1)$ and $S^- \in L^2(0,1)$. The collision operator C and the projection operator $\overline{P}$ are defined as

$$C\varphi(\mu) = \varphi(\mu) - P\varphi(\mu) = \varphi(\mu) - \frac{1}{2}\int_{-1}^{1}\varphi(\mu')\ d\mu', \quad \varphi \in L^2(-1,1),$$

$$\overline{P}\varphi = 2\int_0^1 \mu\varphi(\mu)d\mu, \quad \varphi \in L^2(0,1).$$

THEOREM 1. *Let γ be an absorption coefficient such that $\gamma > \frac{\beta}{2a} + \frac{\beta^2}{2}$. Let us assume that the data Pf_0 and PS are sufficiently smooth (for the precise conditions,*

see Lemma 3 *in section* 4*). Then the solution* f_ε *of the transport equation* Pb(1.1) *satisfies*

(1)
$$\|f_\varepsilon - f\|_{L^2((-a,a)\times(-1,1))} \le \varepsilon M e^{\delta T}, \quad \textit{uniformly on compact subsets of } \ t \in (0,T],$$

where $T > 0$*,* M *and* δ *are two positive constants independent of* T *and* ε*, and* f *is the solution of the related diffusion equation, considered as an initial-boundary value problem in* $H^1(-a,a)$*,*

$$\text{Pb(1.2)} \qquad \begin{cases} \dfrac{\partial f}{\partial t}(x,t) = \dfrac{1}{3}\dfrac{\partial^2 f}{\partial x^2} - \gamma f + PS \quad \textit{on } (-a,a)\times\mathbb{R}^+, \\ \dfrac{4}{3}\dfrac{\partial f}{\partial n} = \beta\ f + \overline{P}S^- \quad \textit{in } x = -a, \\ \dfrac{4}{3}\dfrac{\partial f}{\partial n} = \beta\ f + \overline{P}S^+ \quad \textit{in } x = a, \\ f(x,0) = Pf_0(x). \end{cases}$$

The additional absorption γ that appears in both the transport and diffusion problems is introduced in order to make the corresponding evolutions dissipative. Indeed, in the absence of net absorption, unscaled multiplying BCs lead to an unbounded semigroup, and even scaled multiplying BCs lead only to an exponentially bounded semigroup, with strictly positive type. On the other hand, introducing a suitable absorption, as described by the parameter γ, makes the semigroup a contraction, and as a result, the analysis is simplified. In this sense, the presence of γ may be viewed as a mathematical artifact. At the same time though, we note that γ estimates quite sharply the multiplicative effect of the boundaries. Besides and beyond the rigorous proof of the asymptotic equivalence between the transport and diffusive regimes as expressed by (1), this relationship between γ and β is an interesting result in itself, since it allows one to compare, within a specified geometry, the relative effects of the boundary multiplication and the volume absorption needed to counteract it.

The precise meaning and complete analysis of all the terms and operators appearing in the two problems above will be clarified in the next sections, as follows. In section 2 we study the properties of the transport operator with multiplying BCs in order to derive some exponential bound of the transport semigroup. In section 3 we investigate the Milne problem with specularly reflecting BCs that arises from the boundary layer. In section 4 we apply a scaling argument to prove the main result. Section 5 contains a short discussion.

2. Spectral properties of the transport operator. We define the space

$$\begin{aligned} W^2 &= W^2\big((-a,a)\times(-1,1)\big) \\ &= \left\{\varphi \in L^2\big((-a,a)\times(-1,1)\big);\ \mu\frac{\partial\varphi}{\partial x} \in L^2\big((-a,a)\times(-1,1)\big)\right\}, \end{aligned}$$

endowed with the norm $\|\varphi\|_{W^2} := (\|\varphi\|_{L^2}^2 + \|\mu\frac{\partial\varphi}{\partial x}\|_{L^2}^2)^{\frac{1}{2}}$, where L^2 denotes the usual $L^2\big((-a,a)\times(-1,1)\big)$-norm. It is known (see [DL2, Chap. XXI]) that functions of W^2 have traces on $\{a\}$ and on $\{-a\}$ in $L^2(-1,1)$.

We consider the transport operator

$$\text{(2)}\qquad \begin{cases} T_\varepsilon\varphi = -\dfrac{1}{\varepsilon}\ \mu\dfrac{\partial\varphi}{\partial x} - \dfrac{1}{\varepsilon^2}\ C\varphi, \\ D(T_\varepsilon) = \Big\{\varphi \in W^2,\ \varphi(-a,\mu) = \alpha\ \varphi(-a,-\mu) \text{ and } \varphi(a,-\mu) = \alpha\ \varphi(a,\mu) \text{ for } \mu > 0\Big\} \end{cases}$$

in a slab of width $2a$. The restitution coefficient at the surface, α, is strictly greater than 1, accounting for multiplying BCs. The collision operator C is a nonnegative self-adjoint operator on $L^2(-1,1)$. The spectrum of C is composed of two eigenvalues, namely, $\lambda = 0$ associated with the constant eigenvector and $\lambda = 1$ associated with the infinite-dimensional subspace of functions with zero mean. These two eigenspaces reduce the operator C. For convenience, we define the streaming operator [GMP]

$$A_\varepsilon\varphi = -\frac{1}{\varepsilon}\ \mu\frac{\partial\varphi}{\partial x} - \frac{1}{\varepsilon^2}\ \varphi, \qquad D(A_\varepsilon) = D(T_\varepsilon),$$

which generates a group (see [BT], [MaT], [LaM], and [B2]).

PROPOSITION 1. *The spectrum of A_ε, $\sigma(A_\varepsilon)$, is located in $\{\lambda \in \mathbb{C},\ -\frac{1}{\varepsilon^2} \le \operatorname{Re}\lambda \le -\frac{1}{\varepsilon^2} + \frac{\ln\alpha}{2a\varepsilon}\}$ and is composed of the line $\{\lambda \in \mathbb{C}, \operatorname{Re}\lambda = -\frac{1}{\varepsilon^2}\}$ and the segments*

$$\bigcup_{k\in\mathbb{Z}} \left\{ -\frac{1}{\varepsilon^2} + v\left(\frac{\ln\alpha}{2a\varepsilon} + i\ \frac{k\pi}{2a\varepsilon}\right),\ v \in [0,1]\right\}.$$

The spectrum of T_ε, $\sigma(T_\varepsilon)$, consists of the spectrum $\sigma(A_\varepsilon)$, plus possibly a denumerable set of isolated eigenvalues with finite algebraic multiplicities. Moreover, the semigroup generated by T_ε satisfies, for all $\varphi \in L^2\big((-a,a)\times(-1,1)\big)$, the estimate

$$\text{(3)}\qquad \left\| e^{tT_\varepsilon}\varphi\right\|_{L^2} \le \alpha\ e^{\omega(\alpha,\varepsilon)t}\ \|\varphi\|_{L^2}$$

with the exponential bound $\omega(\alpha,\varepsilon) = \frac{\ln\alpha}{2a\varepsilon} + \frac{1}{2\varepsilon^2}(\alpha + \frac{1}{\alpha} - 2)$.

Remark 1. With the scaled restitution coefficient $\alpha = 1 + \beta\varepsilon$ $(\beta > 0)$, the transport semigroup becomes exponentially bounded in time, uniformly in ε, since

$$w(1+\beta\varepsilon,\varepsilon) = \frac{\ln(1+\beta\varepsilon)}{2a\varepsilon} + \frac{1}{2\varepsilon^2}\left(1 + \beta\varepsilon + \frac{1}{1+\beta\varepsilon} - 2\right) \longrightarrow \frac{\beta}{2a} + \frac{\beta^2}{2}\ (\varepsilon \to 0).$$

Proof. First we study the spectrum of A_ε. Let us consider $\lambda = -\frac{1}{\varepsilon^2} + ib$, where $b \in \mathbb{R}$, and the sequence of functions $\varphi_n(x,\mu) = \sqrt{n}\ \mathrm{e}^{-i\frac{bx\varepsilon}{\mu} + \frac{1}{x^2-a^2}}\ f_n(\mu)$ $(n \in \mathbb{N})$, where $f_n(\mu) = n\mu\mathbb{1}_{[0,\frac{1}{n}]}(\mu) + (2-n\mu)\mathbb{1}_{[\frac{1}{n},\frac{2}{n}]}(\mu)$ and $\mathbb{1}_{[a,b]}(.)$ denotes the characteristic function of the interval $[a,b]$. We notice that φ_n is in the domain of A_ε and

$$(\lambda - A_\varepsilon)\ \varphi_n(x,\mu) = -\mu\ \frac{2x}{x^2-a^2}\ \varphi_n(x,\mu).$$

One verifies easily that there exists a constant $M > 0$ such that $\|\varphi_n\|_{L^2} \ge M$ and $\|(\lambda - A_\varepsilon)\varphi_n\|_{L^2}$ converges to zero as n tends to ∞. Therefore the line $\{\lambda \in \mathbb{C}, \operatorname{Re}\lambda = -\frac{1}{\varepsilon^2}\}$ belongs to the spectrum of A_ε.

Let us consider now $\lambda = -\frac{1}{\varepsilon^2} + v(\frac{\ln\alpha}{2a\varepsilon} + i\ \frac{k\pi}{2a\varepsilon})$, where $v \in (0,1)$ and $k \in \mathbb{Z}$, and the sequence of functions $\varphi_n(x,\mu) = \sqrt{n}\ \mathrm{e}^{-x\ \operatorname{sgn}(\mu)\ \left(\frac{\ln\alpha}{2a} + i\frac{k\pi}{2a}\right)}\ f_n(\mu)$ $(n \in \mathbb{N})$, where

$$\begin{aligned} f_n(\mu) = {} & (n(\mu+v)+1)\mathbb{1}_{[-v-\frac{1}{n},-v]}(\mu) + (-n(\mu+v)+1)\mathbb{1}_{[-v,-v+\frac{1}{n}]}(\mu) \\ & - (n(\mu-v)+1)\mathbb{1}_{[v-\frac{1}{n},v]}(\mu) - (-n(\mu-v)+1)\mathbb{1}_{[v,v+\frac{1}{n}]}(\mu). \end{aligned}$$

We notice that the functions φ_n are in the domain of A_ε and

$$(\lambda - A_\varepsilon)\ \varphi_n = \frac{\ln \alpha}{2a\varepsilon}\ (v - |\mu|)\ \varphi_n + \frac{k\pi i}{2a\varepsilon}\ (v - |\mu|)\ \varphi_n.$$

As before, one verifies that there exists a constant $M > 0$ such that $\|\varphi_n\|_{L^2} \geq M$ and $\|(\lambda - A_\varepsilon)\varphi_n\|_{L^2}$ converges to zero as n tends to ∞. Therefore the segment $\{-\frac{1}{\varepsilon^2} + v(\frac{\ln \alpha}{2a\varepsilon} + i\ \frac{k\pi}{2a\varepsilon}),\ v \in [0,1]\}$ belongs to the spectrum of A_ε.

The remainder of the complex plane is the resolvent set of the operator A_ε. This can be checked by writing down explicitly the expression of the evolution generated by A_ε, calculating the resolvent of A_ε via Laplace transform, and estimating its norm for complex numbers belonging to the complement of $\sigma(A_\varepsilon)$. A well-known compactness result (see [M2, Chap. 3]) ensures that the operators A_ε and T_ε have the same essential spectrum. Thus the spectrum of T_ε is composed of the spectrum of A_ε together with, possibly, isolated eigenvalues with finite algebraic multiplicities.

In order to prove the estimate (3), we consider the solution f_ε of the equation

$$\text{(4)}\qquad \begin{cases} \dfrac{\partial f_\varepsilon}{\partial t}(x,\mu,t) = -\dfrac{1}{\varepsilon}\mu\dfrac{\partial f_\varepsilon}{\partial x} - \dfrac{1}{\varepsilon^2}\ Cf_\varepsilon \quad \text{on } (-a,a)\times(-1,1)\times\mathbb{R}^+, \\ f_\varepsilon(-a,\mu,t) = \alpha\ f_\varepsilon(-a,-\mu,t), \quad \mu > 0, \\ f_\varepsilon(a,-\mu,t) = \alpha\ f_\varepsilon(a,\mu,t), \qquad \mu > 0, \\ f_\varepsilon(x,\mu,0) = \varphi(x,\mu), \end{cases}$$

where $\varphi \in L^2\big((-a,a)\times(-1,1)\big)$. We apply here the method of change of functions introduced in [LoM] in a related framework. Let $h_\varepsilon = \alpha^{\mathrm{sgn}(\mu)\frac{x}{2a}+\frac{1}{2}} f_\varepsilon$. The new function h_ε satisfies the following transport equation with (conservative) specularly reflecting BCs:

$$\text{(5)}\qquad \begin{cases} \dfrac{\partial h_\varepsilon}{\partial t}(x,\mu,t) = -\dfrac{1}{\varepsilon}\mu\dfrac{\partial h_\varepsilon}{\partial x} + \dfrac{1}{\varepsilon}|\mu|\dfrac{\ln\alpha}{2a}\ h_\varepsilon - \dfrac{1}{\varepsilon^2}\ h_\varepsilon \\ \quad + \dfrac{1}{2\varepsilon^2}\displaystyle\int_{-1}^{1} \alpha^{\frac{x}{2a}\ (\mathrm{sgn}(\mu)-\mathrm{sgn}(\mu'))}\ h_\varepsilon(x,\mu',t)\ d\mu' \quad \text{on } (-a,a)\times(-1,1)\times\mathbb{R}^+, \\ h_\varepsilon(-a,\mu,t) = h_\varepsilon(-a,-\mu,t), \quad \mu > 0, \\ h_\varepsilon(a,-\mu,t) = h_\varepsilon(a,\mu,t), \qquad \mu > 0, \\ h_\varepsilon(x,\mu,0) = \alpha^{\mathrm{sgn}(\mu)\frac{x}{2a}+\frac{1}{2}}\ \varphi(x,\mu). \end{cases}$$

Let us multiply the transport equation (5) by h_ε, integrate by parts on $(-a,a)\times(-1,1)\times(0,t)$ for $t > 0$, and use the specularly reflecting BCs; then

$$\frac{1}{2}\ \|h_\varepsilon(t)\|^2_{L^2} \leq \frac{1}{2}\ \|h_\varepsilon(0)\|^2_{L^2} + \left(\frac{\ln\alpha}{2a\varepsilon} - \frac{1}{\varepsilon^2}\right)\int_0^t \|h_\varepsilon(s)\|^2_{L^2}\ ds$$
$$+ \frac{1}{2\varepsilon^2}\int_0^t\int_{-a}^{a}\int_{-1}^{1}\int_{-1}^{1} \alpha^{\frac{x}{2a}\ (\mathrm{sgn}(\mu)-\mathrm{sgn}(\mu'))}\ h_\varepsilon(x,\mu',s)\ h_\varepsilon(x,\mu,s)\ d\mu' d\mu dx ds.$$

By noticing that the function $x \longmapsto \alpha^{\frac{x}{a}} + \alpha^{-\frac{x}{a}}$ attains its maximum on $[-a,a]$ in $-a$ and in a, it follows that

$$\int_{-a}^{a}\int_{-1}^{1}\int_{-1}^{1} \alpha^{\frac{x}{2a}\ (\mathrm{sgn}(\mu)-\mathrm{sgn}(\mu'))}\ h_\varepsilon(x,\mu')\ h_\varepsilon(x,\mu)\ d\mu' d\mu dx \leq \left(\alpha + \frac{1}{\alpha}\right)\ \|h_\varepsilon\|^2_{L^2},$$

yielding

$$\frac{1}{2}\ \|h_\varepsilon(t)\|^2_{L^2} \le \frac{1}{2}\ \alpha^2\|\varphi\|^2_{L^2} + \left(\frac{\ln\alpha}{2a\varepsilon} + \frac{1}{2\varepsilon^2}\left(\alpha + \frac{1}{\alpha} - 2\right)\right)\int_0^t \|h_\varepsilon(s)\|^2_{L^2}\ ds.$$

By Gronwall's lemma we have $\|h_\varepsilon(t)\|^2_{L^2} \le \alpha^2\ \mathrm{e}^{2\omega(\alpha,\varepsilon)t}\ \|\varphi\|^2_{L^2}$, which together with the inequality $\|f_\varepsilon(t)\|_{L^2} \le \|h_\varepsilon(t)\|_{L^2}$ concludes the proof. ∎

3. Milne problem with reflecting boundary conditions. The problem of the boundary layer is related to the conservative Milne problem with reflecting BCs in bounded and semi-infinite geometries. The difference in the nature of these geometries leads to somewhat different analyses. First, we introduce the functional spaces which we shall use in what follows. Let $H_T := L^2(((-1,1),|\mu|d\mu);L^\infty(0,\infty))$ and $L^2_\mu(0,1)$ denote the spaces of measurable functions for which the norms $\|\varphi\|^2_{H_T} := \int_{-1}^1 |\mu|\sup_{x\in(0,\infty)}|\varphi(x,\mu)|^2\,d\mu$, $\|\varphi\|^2_{L^2_\mu(0,1)} := \int_0^1 \mu|\varphi(\mu)|^2 d\mu$ are, respectively, bounded.

3.1. Bounded geometry.

LEMMA 1. *Let $e^{2a} > \alpha > 0$ (this condition is satisfied for $\alpha \in [0,1]$), and let f and g be in $L^2_\mu(0,1)$. Then there exists an extension ϕ of the functions f and g in W^2 which satisfies*

$$\text{(6)}\qquad \begin{cases} \mu\dfrac{\partial\phi}{\partial x} + \phi = 0 \quad \text{on } (-a,a)\times(-1,1), \\ \phi(-a,\mu) = \alpha\ \phi(-a,-\mu) + f(\mu), \quad \mu > 0, \\ \phi(a,-\mu) = \alpha\ \phi(a,\mu) + g(\mu), \quad \mu > 0. \end{cases}$$

Moreover, the extension application is continuous, i.e., there exists a constant M independent of α and a, such that

$$\|\phi\|_{W^2} \le M\alpha^2\big(\|f\|_{L^2_\mu(0,1)} + \|g\|_{L^2_\mu(0,1)}\big).$$

Proof. The solution of (6) is given, for $\mu > 0$, by

$$\phi(x,\mu) = \left(\mathrm{e}^{\frac{2a}{\mu}} - \alpha^2\mathrm{e}^{-\frac{2a}{\mu}}\right)^{-1}\left(\alpha g(\mu)\ \mathrm{e}^{-\frac{a+x}{\mu}} + f(\mu)\ \mathrm{e}^{-\frac{x-a}{\mu}}\right),$$

$$\phi(x,-\mu) = \left(\mathrm{e}^{\frac{2a}{\mu}} - \alpha^2\mathrm{e}^{-\frac{2a}{\mu}}\right)^{-1}\left(g(\mu)\ \mathrm{e}^{\frac{a+x}{\mu}} + \alpha f(\mu)\ \mathrm{e}^{\frac{x-a}{\mu}}\right).$$

If $\alpha > 1$, the condition $2a > \ln\alpha$ ensures $\mathrm{e}^{\frac{2a}{\mu}} - \alpha^2\mathrm{e}^{-\frac{2a}{\mu}} > 0$ for all $0 < \mu < 1$. ∎

LEMMA 2. *Let T_R be the transport operator with specularly reflecting BCs:*

$$\begin{cases} T_R\ \varphi = -\mu\dfrac{\partial\varphi}{\partial x} - C\varphi, \\ D(T_R) = \Big\{\varphi \in W^2,\ \varphi(-a,\mu) = \varphi(-a,-\mu)\ \textit{and}\ \varphi(a,-\mu) = \varphi(a,\mu)\ \textit{for}\ \mu > 0\Big\}. \end{cases}$$

The value 0 is an algebraically simple eigenvalue of T_R, associated with the constant eigenvector. The same result holds for the adjoint of T_R.

Proof. It is obvious that the constant functions are eigenfunctions of T_R associated with the eigenvalue 0. Let φ denote an eigenfunction associated with 0. Upon integrating by parts, we obtain $(T_R\varphi,\varphi) = (C\varphi,\varphi) = 0$, where the bracket denotes

the scalar product in $L^2\big((-a,a)\times(-1,1)\big)$. Then $C\varphi = 0$ and φ does not depend on μ. Thus φ is constant because $\frac{\partial\varphi}{\partial x} = 0$.

Now let us note that $(T_R\varphi, 1) = (-\mu\frac{\partial\varphi}{\partial x}, 1) - (C\varphi, 1) = -(\varphi, C1) = 0$. If φ is a generalized eigenfunction such that $T_R\varphi = 1$, then $(T_R\varphi, 1) = 4a$. Therefore T_R does not admit other generalized eigenfunctions beyond the constant functions. The same proof holds for the adjoint of T_R defined by $T_R^\star\,\varphi = \mu\frac{\partial\varphi}{\partial x} - C\varphi,\ D(T_R^\star) = D(T_R)$. ∎

From Lemma 2 and from the theory of isolated singularities of the resolvent (see [Y]), we deduce the Fredholm alternative for the operator T_R and the following proposition.

PROPOSITION 2. *Let f and g be in $L^2_\mu(0,1)$. The system*

$$\text{(7)}\qquad \begin{cases} \mu\dfrac{\partial\varphi}{\partial x} + C\varphi = 0 \quad \text{on } (-a,a)\times(-1,1), \\ \varphi(-a,\mu) = \varphi(-a,-\mu) + f(\mu), \quad \mu > 0, \\ \varphi(a,-\mu) = \varphi(a,\mu) + g(\mu), \quad \mu > 0, \end{cases}$$

admits a solution in W^2 if and only if

$$\text{(8)}\qquad \int_0^1 \mu\ \big(g(\mu) + f(\mu)\big)\ d\mu = 0.$$

Moreover, if they exist, the solutions are defined up to a constant, and there is a unique solution in $\ker(T_R)^\perp$ *which satisfies*

$$\text{(9)}\qquad \int_{-a}^{a}\int_{-1}^{1} \varphi(x,\mu)\ dxd\mu = 0.$$

Proof. According to Lemma 2, 0 is an algebraically simple eigenvalue of T_R and of its adjoint, $T_R{}^\star$, associated with the constant eigenvector. The associated spectral projection, P, is bounded and $\ker(P) = \overline{R(-T_R)}$ (see [Y]). Thus the range of T_R, $R(T_R)$, is a closed subspace and $R(T_R) = \overline{R(T_R)} = \ker(T_R{}^\star)^\perp = \ker(T_R)^\perp$.

Now let ϕ be the extension of f and g as in Lemma 1. Then φ satisfies (7) if and only if the function $\psi = \varphi - \phi$ satisfies the system

$$\text{(10)}\qquad \begin{cases} \mu\dfrac{\partial\psi}{\partial x} + C\psi = T_R\psi = -\mu\dfrac{\partial\phi}{\partial x} - C\phi \quad \text{on } (-a,a)\times(-1,1), \\ \psi(-a,\mu) = \psi(-a,-\mu), \quad \mu > 0, \\ \psi(a,-\mu) = \psi(a,\mu), \quad \mu > 0. \end{cases}$$

Since $R(T_R) = \ker(T_R)^\perp$, the system (10) admits a solution if and only if

$$\text{(11)}\qquad \int_{-a}^{a}\int_{-1}^{1}\Big(\mu\frac{\partial\phi}{\partial x}(x,\mu) + C\phi(x,\mu)\Big)dxd\mu = 0.$$

Moreover, the solution ψ is unique in $\ker(T_R)^\perp$, which implies

$$\int_{-a}^{a}\int_{-1}^{1} \varphi(x,\mu)\ dxd\mu = \int_{-a}^{a}\int_{-1}^{1} \phi(x,\mu)\ dxd\mu.$$

But $\phi = -\mu\frac{\partial\phi}{\partial x}$, and an integration by parts leads to (9). Likewise, we can see that condition (11) is equivalent to condition (8). ∎

3.2. Semi-infinite geometry. We note that, in contradistinction with the bounded geometry case, 0 belongs to the spectrum of the transport operator in semi-infinite geometry with reflecting BCs

$$\widetilde{T_R}\ \varphi = -\mu\frac{\partial\varphi}{\partial x} - C\varphi,\ D(\widetilde{T_R})$$
$$= \Big\{\varphi \in W^2\big((0,\infty)\times(-1,1)\big),\ \varphi(0,\mu) = \varphi(0,-\mu) \text{ for } \mu > 0\Big\}$$

but is *not* an isolated eigenvalue (see [P]).

PROPOSITION 3. *The systems*

$$\begin{cases} \pm\mu\dfrac{\partial\varphi}{\partial x} + C\varphi = 0 \quad \text{on } (0,\infty)\times(-1,1), \\ \varphi(0,\mu) = \varphi(0,-\mu) + f(\mu), \quad \mu > 0, \\ \lim\limits_{x\to\infty}\varphi(x,\mu) = 0 \end{cases} \tag{12}$$

admit a unique solution φ *in* $H_T \cap L^1((-1,1);\ L^\infty(0,\infty))$ *if* $f \in L^2_\mu(0,1)$, *in* $L^p((-1,1);\ L^\infty(0,\infty))$ *for all* $1 \le p < \infty$ *if* $f \in L^2(0,1)$, *and in* $L^\infty((0,\infty)\times(-1,1))$ *if* $f \in L^\infty(0,1)$ *if and only if*

$$\int_0^1 \mu f(\mu)d\mu = 0. \tag{13}$$

Moreover, there exists a positive constant M *such that*

$$\|\varphi\|_{H_T} \le M\|f\|_{L^2_\mu(0,1)}, \tag{14}$$

$$\|\varphi\|_{L^1((-1,1);\ L^\infty(0,\infty))} \le M\|f\|_{L^2_\mu(0,1)}, \tag{15}$$

$$\|\varphi\|_{L^p((-1,1);\ L^\infty(0,\infty))} \le M\|f\|_{L^2(0,1)}, \tag{16}$$

$$\|\varphi\|_{L^\infty} \le M\|f\|_{L^\infty(0,1)}, \tag{17}$$

$$\|\varphi(0,.)\|_{L^2(0,1)} \le M\|f\|_{L^2(0,1)}. \tag{18}$$

Proof. We consider the problem

$$\begin{cases} \mu\dfrac{\partial\varphi}{\partial x} + C\varphi = 0 \quad \text{on } (0,\infty)\times(-1,1), \\ \varphi(0,\mu) = \varphi(0,-\mu) + f(\mu), \quad \mu > 0, \\ \lim\limits_{x\to\infty}\varphi(x,\mu) = 0. \end{cases} \tag{19}$$

First we notice that, by integrating (19) with respect to x and μ, we get $\int_0^1 \mu f(\mu)d\mu = 0$. Now let $f \in L^2_\mu(0,1)$. Integration of (19) with respect to the spatial variable leads to

$$\begin{cases} \varphi(x,\mu) = -\dfrac{1}{\mu}\displaystyle\int_x^\infty \mathrm{e}^{\frac{s-x}{\mu}} P\varphi(s)ds, \quad \mu < 0, \\ \varphi(x,\mu) = \dfrac{1}{\mu}\displaystyle\int_0^x \mathrm{e}^{\frac{s-x}{\mu}} P\varphi(s)ds + \dfrac{1}{\mu}\int_0^\infty \mathrm{e}^{-\frac{s+x}{\mu}} P\varphi(s)ds + f(\mu)\mathrm{e}^{-\frac{x}{\mu}}, \quad \mu > 0. \end{cases} \tag{20}$$

By integrating (20) with respect to μ, we obtain that $P\varphi$ satisfies the integral equation

$$P\varphi(x) = \int_0^\infty H(x,s)P\varphi(s)ds + \frac{1}{2}\int_0^1 f(\mu)\mathrm{e}^{-\frac{x}{\mu}}d\mu,$$

where

$$H(x,s) = \frac{1}{2}\int_0^1 \frac{1}{\mu}\mathrm{e}^{-\frac{|x-s|}{\mu}}d\mu + \frac{1}{2}\int_0^1 \frac{1}{\mu}\mathrm{e}^{-\frac{|x+s|}{\mu}}d\mu = \frac{1}{2}\int_1^\infty \frac{1}{t}\mathrm{e}^{-t|x-s|}dt + \frac{1}{2}\int_1^\infty \frac{1}{t}\mathrm{e}^{-t|x+s|}dt.$$

Let H be the operator which acts on $L^2(0,\infty)$ as

$$Hg(x) = \int_0^\infty H(x,s)g(s)ds,$$

and let

$$F(x) = \frac{1}{2}\int_0^1 f(\mu)\mathrm{e}^{-\frac{x}{\mu}}d\mu.$$

We introduce the subspace $X = \{h \in L^2(0,\infty) \mid |||h||| < \infty\}$, endowed with the norm $|||h|||^2 = \|h\|_{L^2}^2 + \big(\int_0^\infty x^2|h(x)|dx\big)^2$. We check easily that $|||F||| \le C\|f\|_{L^2_\mu(0,1)}$. Then $P\varphi$ is a solution g of the equation $g - Hg = F$, with $F \in X$.

Now, let us consider the even extension of the functions g and F to $\mathbb{R}$. Since the kernel H is an even function in each variable, it follows that

$$Hg(x) = \int_{-\infty}^\infty \frac{1}{2}\int_1^\infty \frac{1}{t}\mathrm{e}^{-t|x-s|}dt\ g(s)ds.$$

Thus, we can take advantage of the Fourier analysis, as in [P]. The preceding equation is equivalent to

$$\hat{g}(\xi)\Big(1 - \frac{\arctan(\xi)}{\xi}\Big) = \hat{F}(\xi), \tag{21}$$

where $\hat{h}(\xi) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty h(x)\ \mathrm{e}^{i\xi x}\ dx$ denotes the Fourier transform of a function h and $\xi \longmapsto \frac{\arctan(\xi)}{\xi}$ is the Fourier transform of the function $x \longmapsto \frac{1}{2}\int_1^\infty \frac{1}{t}\mathrm{e}^{-t|x|}dt$. Since F belongs to X, the Fourier transform of F, $\hat{F}$, belongs to $\mathcal{C}^2(\mathbb{R})$. Also, since the function $\xi \longmapsto (1 - \frac{\arctan(\xi)}{\xi})^{-1}$ has a quadratic singularity in $\xi = 0$, (21) admits a unique solution if and only if $\hat{F}(0) = 0$ and $\frac{d\hat{F}}{d\xi}(0) = 0$. While the condition

$$\frac{d\hat{F}}{d\xi}(0) = i\int_{-\infty}^\infty xF(x)dx = 0$$

is always satisfied due to the evenness of F, the condition

$$\hat{F}(0) = \int_{-\infty}^\infty F(x)dx = 2\int_0^\infty F(x)dx = 0$$

with $F(x) = \frac{1}{2}\int_0^1 f(\mu)\mathrm{e}^{-\frac{x}{\mu}}d\mu$ is implied by condition (13). We estimate now the $L^2(0,\infty)$-norm of the solution $g = P\varphi$ of (21). Let $\varepsilon > 0$; then according to Plancherel's formula, we get

$$\int_0^\infty |P\varphi(x)|^2dx = \int_0^\varepsilon \Big|\frac{\xi}{\xi - \arctan(\xi)}\ \hat{F}(\xi)\Big|^2 d\xi + \int_\varepsilon^\infty \Big|\frac{\xi}{\xi - \arctan(\xi)}\ \hat{F}(\xi)\Big|^2 d\xi.$$

On the one hand, by using the uniform boundedness on $\mathbb{R}^2$ of the function $(x,\xi)\in \mathbb{R}^2 \longmapsto \frac{e^{ix\xi}-1-ix\xi}{x^2\xi^2}$, we have

$$\begin{aligned}\int_0^\varepsilon \Big|\frac{\xi}{\xi-\arctan(\xi)}\ \hat F(\xi)\Big|^2 d\xi &\le C_2(\varepsilon)\int_0^\varepsilon \Big|\frac{\hat F(\xi)}{\xi^2}\Big|^2 d\xi\\ &= C_2(\varepsilon)\int_0^\varepsilon \Big|\frac{\hat F(\xi)-\hat F(0)-i\xi(\widehat{xF})(0)}{\xi^2}\Big|^2 d\xi\\ &= C_2(\varepsilon)\int_0^\varepsilon \Big|\int_0^\infty F(x)\ \frac{e^{ix\xi}-1-ix\xi}{\xi^2}\ dx\Big|^2 d\xi \le C_2(\varepsilon)\left(\int_0^\infty x^2|F(x)|\ dx\right)^2.\end{aligned}$$

On the other hand, it is easy to check that

$$\int_\varepsilon^\infty \Big|\frac{\xi}{\xi-\arctan(\xi)}\ \hat F(\xi)\Big|^2 d\xi \le C_1(\varepsilon)\int_\varepsilon^\infty \big|\hat F(\xi)\big|^2 d\xi \le C_1(\varepsilon)\|F\|_{L^2}^2.$$

Therefore, $\|P\varphi\|_{L^2(0,\infty)} \le |||F|||$ and $\|P\varphi\|_{L^2(0,\infty)} \le C\|f\|_{L^2_\mu(0,1)}$. Thus, we showed that $P\varphi$ is uniquely defined if condition (13) is satisfied, and (20) gives the solution of (19). To prove (14), we estimate each part of (20) as follows:

$$\int_0^1 \mu|f(\mu)|^2 \sup_{x\in(0,\infty)} |e^{-\frac{x}{\mu}}|^2 d\mu \ \le\ \|f\|^2_{L^2_\mu(0,1)},$$

$$\begin{aligned}\int_0^1 \frac{\mu}{\mu^2}\sup_{x\in(0,\infty)}\Big|\int_0^\infty e^{-\frac{s+x}{\mu}}P\varphi(s)ds\Big|^2 d\mu &= \int_0^1 \frac{d\mu}{\mu}\left(\sup_{x\in(0,\infty)} e^{-\frac{2x}{\mu}}\right)\Big|\int_0^\infty e^{-\frac{s}{\mu}}P\varphi(s)ds\Big|^2\\ &\le \int_0^1 \frac{d\mu}{\mu}\int_0^\infty e^{-\frac{2s}{\mu}}ds\int_0^\infty |P\varphi(s)|^2 ds \le M\|f\|^2_{L^2_\mu(0,1)},\end{aligned}$$

$$\begin{aligned}\int_0^1 \frac{\mu}{\mu^2}\sup_{x\in(0,\infty)}\Big|\int_0^x e^{\frac{s-x}{\mu}}P\varphi(s)ds\Big|^2 d\mu &\le \int_0^1 \frac{d\mu}{\mu}\sup_{x\in(0,\infty)}\int_0^x e^{2\frac{s-x}{\mu}}ds\int_0^x |P\varphi(s)|^2 ds\\ &\le \int_0^1 \frac{d\mu}{\mu}\ \frac{\mu}{2}\int_0^\infty |P\varphi(s)|^2 ds \le M\|f\|^2_{L^2_\mu(0,1)},\end{aligned}$$

$$\begin{aligned}\int_{-1}^0 \frac{|\mu|}{\mu^2}\sup_{x\in(0,\infty)}\Big|\int_x^\infty e^{\frac{s-x}{\mu}}P\varphi(s)ds\Big|^2 d\mu &\le \int_0^1 \frac{d\mu}{\mu}\sup_{x\in(0,\infty)}\int_x^\infty e^{2\frac{x-s}{\mu}}ds\int_x^\infty |P\varphi(s)|^2 ds\\ &\le \int_0^1 \frac{d\mu}{\mu}\ \frac{\mu}{2}\int_0^\infty |P\varphi(s)|^2 ds \le M\|f\|^2_{L^2_\mu(0,1)}.\end{aligned}$$

Inequality (15) is obtained using similar estimates.

We show now that $P\varphi$ is a bounded function if $f\in L^2(0,1)$; more precisely, there exists $M>0$ such that

$$\|P\varphi\|_{L^\infty(0,\infty)} \le M\|f\|_{L^2(0,1)}. \tag{22}$$

One first notices, by using the evenness of F, that

$$\hat{F}(\xi) = \frac{2}{\sqrt{2\pi}} \int_0^1 \frac{\mu f(\mu)}{1+\xi^2\mu^2}\, d\mu.$$

Then, as $P\varphi$ is the inverse Fourier transform of $\hat{g}$, it is sufficient to prove that $\hat{g}$ from (20) is bounded in $L^1(\mathbb{R})$. As before, we choose a sufficiently small $\varepsilon > 0$ and get

$$\int_0^\varepsilon \Big|\frac{\xi}{\xi - \arctan(\xi)}\, \hat{F}(\xi)\Big| d\xi = \int_0^\varepsilon \Big|\frac{\xi^3}{\xi - \arctan(\xi)}\, \frac{\hat{F}(\xi)}{\xi^2}\Big| d\xi \le C_2(\varepsilon) \left(\int_0^\varepsilon \Big|\frac{\hat{F}(\xi)}{\xi^2}\Big|^2 d\xi \right)^{1/2}$$
$$\le C_2(\varepsilon) \int_0^\infty x^2 |F(x)|\, dx \le C_2(\varepsilon)\, \|f\|_{L^2_\mu(0,1)} \;\le\; C_2(\varepsilon)\, \|f\|_{L^2(0,1)}.$$

On the other hand,

$$\int_\varepsilon^\infty \Big|\frac{\xi}{\xi - \arctan(\xi)}\, \hat{F}(\xi)\Big| d\xi \le C_1(\varepsilon) \int_\varepsilon^\infty |\hat{F}(\xi)| d\xi \le C_1(\varepsilon) \int_\varepsilon^\infty d\xi \int_0^1 \frac{\mu |f(\mu)|}{1+\xi^2\mu^2}\, d\mu$$
$$\le C_1(\varepsilon)\, \|f\|_{L^2(0,1)} \int_\varepsilon^\infty \left(\int_0^1 \frac{\mu^2}{\left(1+\xi^2\mu^2\right)^2}\, d\mu \right)^{1/2} d\xi \le C_1(\varepsilon)\, \|f\|_{L^2(0,1)}.$$

Thus estimate (22) is proven.

From (22) and (20), we easily obtain the estimates (16), (17), and (18). Finally, a similar result holds for the system

$$\begin{cases} -\mu \dfrac{\partial \varphi}{\partial x} + C\varphi = 0 \quad \text{on } (0,\infty)\times(-1,1), \\ \varphi(0,\mu) = \varphi(0,-\mu) + f(\mu), \quad \mu > 0, \\ \lim\limits_{x\to\infty} \varphi(x,\mu) = 0. \end{cases} \qquad \square$$

We consider now the Milne problem without the decay condition at infinity.

PROPOSITION 4. *Let $f \in L^2(0,1)$ and $1 \le p < \infty$. If $\int_0^1 \mu f(\mu) d\mu = 0$, then the systems*

$$\begin{cases} \pm\mu \dfrac{\partial \varphi}{\partial x} + C\varphi = 0 \quad \text{on } (0,\infty)\times(-1,1), \\ \varphi(0,\mu) = \varphi(0,-\mu) + f(\mu), \quad \mu > 0, \end{cases} \tag{23}$$

admit an infinity of solutions in $L^p((-1,1);\ L^\infty(0,\infty))$, or in $L^\infty((0,\infty)\times(-1,1))$ if $f \in L^\infty(0,1)$. These solutions are obtained by adding any constant to the unique solution of (12).

Proof. From the method introduced in [BSS], slightly modified in [DL2, Chap. XXI.3, Appendix, p. 335], we know (see the proof of [DL2, Thm. 3, p. 340]) that there exists a sequence $\{x_n\}_{n\in\mathbb{N}}$ such that $x_n \longrightarrow \infty$ and the solution φ of (23) satisfies

$$\int_0^{x_n} dx \int_{-1}^1 \big(\varphi(x,\mu) - P\varphi(x)\big)^2 d\mu \le \frac{1}{2} \int_{-1}^1 \mu \varphi^2(0,\mu) d\mu.$$

To prove the above estimate, the authors use only the fact that the solution is bounded in the space variable. Now if $f = 0$, we get

$$\int_{-1}^1 \mu\varphi^2(0,\mu) d\mu = \int_0^1 \mu\big(\varphi^2(0,\mu) - \varphi^2(0,-\mu)\big) d\mu = 0.$$

Therefore φ is independent of the velocity variable, and thus $\frac{\partial\varphi}{\partial x} = 0$ and φ is a constant. $\square$

4. Proof of the main result. In order to prove the existence of all the terms in the asymptotic expansion in the proof of Theorem 1, we need a few additional regularity results for the solution of the diffusion equation Pb(1.2). These results are warranted by the following lemma.

LEMMA 3. *If the data Pf_0 and PS are in $L^2(-a,a)$, then the problem* Pb(1.2) *admits a unique solution in $H^1(-a,a)$. Let h be a smooth extension function of the sources on the boundary $\overline{P}S^{\pm}$. Assuming that both functions Pf_0-h and $PS-\frac{1}{3}\frac{\partial^2 h}{\partial x^2}+\gamma h$ belong to $H^{10}(-a,a)$ and satisfy the BCs*

$$\frac{4}{3}\frac{\partial^{2i+1}}{\partial x^{2i+1}}\dot{}(\pm a)=\pm\beta\frac{\partial^{2i}}{\partial x^{2i}}\dot{}(\pm a),\quad i=0,\dots,4, \tag{24}$$

then the solution f of Pb(1.2) *is in $H^5(-a,a)$ and there exist two positive constants M and δ such that*

$$\left\|\frac{\partial^i f}{\partial x^i}(t)\right\|_{L^2(-a,a)}\le Me^{\delta t}\qquad \forall t\ge 0,\ i=1,\dots,5. \tag{25}$$

Remark 2. Notice that the exponential bound $\mathrm{e}^{\delta t}$ in estimate (25) is due to the fact that the Robin BCs are unstable. With the assumptions of Lemma 3, the solution f of the diffusion equation Pb(1.2) is in fact in $H^{10}(-a,a)$. The assumptions of Lemma 3 can probably be refined; here we focus only on the rigorous justification of the asymptotic expansion to prove the asymptotic equivalence as quantified by (1).

Proof. To simplify the proof we remove the constant 4/3 in the BCs. First, we use the superposition principle to eliminate the nonhomogeneous sources at the boundary, and define the (smooth) extension function

$$h(x)=\frac{\overline{P}S^+}{4a^2}(x+a)^2(x-a)+\frac{\overline{P}S^-}{4a^2}(x-a)^2(x+a).$$

Thus the function $g=f-h$ satisfies the following equation with homogeneous BCs:

$$\begin{cases}\dfrac{\partial g}{\partial t}(x,t)=\dfrac{1}{3}\dfrac{\partial^2 g}{\partial x^2}-\gamma g+\psi,\\ \dfrac{\partial g}{\partial n}=\beta\ g\quad \text{in } x=-a \text{ and } x=a,\\ g(x,0)=\varphi,\end{cases} \tag{26}$$

where $\varphi=Pf_0-h$ and $\psi=PS-\frac{1}{3}\frac{\partial^2 h}{\partial x^2}+\gamma h$.

We can now apply the variational theory (see, for example, [DL1, Chap. VII.2.3.2]) to deduce the existence of a solution in $H^1(-a,a)$ and that the operator

$$T_Df=\frac{1}{3}\frac{\partial^2 f}{\partial x^2}-\gamma f,\ D(T_D)=\left\{f\in H^2(-a,a),\ \frac{\partial f}{\partial x}(a)=\beta f(a),\ \frac{\partial f}{\partial x}(-a)=-\beta f(-a)\right\}$$

generates an analytic semigroup. Notice that the assumptions in Lemma 3 are equivalent to the fact that $\varphi\in D(T_D^5)$ and $\psi\in D(T_D^5)$. Then the semigroup theory (see [Pa]) implies that the solution g belongs to $D(T_D^5)$, namely, $g\in H^{10}(-a,a)$ and

$$\frac{\partial^{2i+1}g}{\partial x^{2i+1}}(\pm a)=\pm\beta\frac{\partial^{2i}g}{\partial x^{2i}}(\pm a),\quad i=0,\dots,4. \tag{27}$$

For sake of brevity, we omit here the rather technical proof of estimates (25); the interested reader can find them in [PT]. □

Proof of Theorem 1. We seek the solution f_ε in the form

$$f_\varepsilon = u_\varepsilon^0(x,\mu,t) + u_\varepsilon^i\left(x,\mu,\tau=\frac{t}{\varepsilon^2}\right) + u_\varepsilon^b\left(\xi=\frac{x}{\varepsilon},\mu,t\right) + w_\varepsilon(x,\mu,t),$$

where u_ε^0, u_ε^i, u_ε^b, and w_ε denote, respectively, the interior, initial layer, boundary layer, and remainder terms. Each one of these first three terms is assumed to satisfy exactly the transport equation and is written as an asymptotic expansion in ε, namely,

$$u_\varepsilon^0 = \sum_{n=0}^{\infty}\varepsilon^n u_n^0, \quad u_\varepsilon^i = \sum_{n=0}^{\infty}\varepsilon^n u_n^i, \quad u_\varepsilon^b = \sum_{n=0}^{\infty}\varepsilon^n u_n^b.$$

In what follows we expand u_ε^0 and u_ε^b to order 2 and u_ε^i to order 0 in ε. The initial layer term is assumed to satisfy

$$u_0^i(\tau) \longrightarrow 0 \ (\tau\to\infty). \tag{28}$$

The boundary layer terms are written as

$$u_j^b\left(\frac{x}{\varepsilon},\mu,t\right) = u_j^{b+}\left(\xi_+=\frac{a-x}{\varepsilon},\mu,t\right)\mathbb{1}_{(0,a)}(x) + u_j^{b-}\left(\xi_-=\frac{a+x}{\varepsilon},\mu,t\right)\mathbb{1}_{(-a,0)}(x)$$
$$(j=0,1,2).$$

Each term u_j^{b+} and u_j^{b-} $(j=0,1,2)$ are in fact defined on $\xi_\pm\in(0,\infty)$ and assumed to satisfy

$$u_j^{b\pm}(\xi_\pm) \longrightarrow 0 \ (\xi_\pm\to\infty) \quad (j=0,1,2). \tag{29}$$

In the initial and boundary layer terms we perform the change of variables $\tau = \frac{t}{\varepsilon^2}$ and $\xi = \frac{x}{\varepsilon}$, respectively. By replacing each term in Pb(1.1) and equating the coefficients of corresponding powers in ε, we get

$$Cu_0^0 = 0, \tag{30}$$

$$Cu_1^0 = -\mu\frac{\partial u_0^0}{\partial x}, \tag{31}$$

$$Cu_2^0 = -\Big(\frac{\partial u_0^0}{\partial t} + \mu\frac{\partial u_1^0}{\partial x} + \gamma u_0^0 - S\Big), \tag{32}$$

$$\frac{\partial u_0^i}{\partial\tau} = -Cu_0^i, \tag{33}$$

$$\mu\frac{\partial u_0^b}{\partial\xi} + Cu_0^b = 0, \tag{34}$$

$$\mu\frac{\partial u_1^b}{\partial\xi} + Cu_1^b = 0, \tag{35}$$

$$\mu\frac{\partial u_2^b}{\partial \xi} + Cu_2^b = -\gamma u_0^b - \frac{\partial u_0^b}{\partial t}. \tag{36}$$

The same procedure applied to the initial data and the BCs leads to

$$u_0^0(x,\mu,0) + u_0^i(x,\mu,0) = f_0(x,\mu), \tag{37}$$

and for $\mu > 0$

$$\begin{cases} u_0^0(-a,\mu) + u_0^b\left(-\frac{a}{\varepsilon},\mu\right) = u_0^0(-a,-\mu) + u_0^b\left(-\frac{a}{\varepsilon},-\mu\right), \\ u_0^0(a,-\mu) + u_0^b\left(\frac{a}{\varepsilon},-\mu\right) = u_0^0(a,\mu) + u_0^b\left(\frac{a}{\varepsilon},\mu\right), \end{cases} \tag{38}$$

(39)
$$\begin{cases} u_1^0(-a,\mu) + u_1^b\left(-\frac{a}{\varepsilon},\mu\right) = u_1^0(-a,-\mu) + u_1^b\left(-\frac{a}{\varepsilon},-\mu\right) \\ \qquad\qquad + \beta u_0^0(-a,-\mu) + \beta u_0^b\left(-\frac{a}{\varepsilon},-\mu\right) + S^-, \\ u_1^0(a,-\mu) + u_1^b\left(\frac{a}{\varepsilon},-\mu\right) = u_1^0(a,\mu) + u_1^b\left(\frac{a}{\varepsilon},\mu\right) + \beta u_0^0(a,\mu) + \beta u_0^b\left(\frac{a}{\varepsilon},\mu\right) + S^+, \end{cases}$$

(40)
$$\begin{cases} u_2^b\left(-\frac{a}{\varepsilon},\mu\right) + u_2^0(-a,\mu,t) = u_2^b\left(-\frac{a}{\varepsilon},-\mu\right) + \beta u_1^b\left(-\frac{a}{\varepsilon},-\mu\right) \\ \qquad\qquad + u_2^0(-a,-\mu,t) + \beta u_1^0(-a,-\mu,t), \\ u_2^b\left(\frac{a}{\varepsilon},-\mu\right) + u_2^0(a,-\mu,t) = u_2^b\left(\frac{a}{\varepsilon},\mu\right) + \beta u_1^b\left(\frac{a}{\varepsilon},\mu\right) + u_2^0(a,\mu,t) + \beta u_1^0(a,\mu,t). \end{cases}$$

Equations (30)–(33) are treated in a standard manner (see [Sa], [BSS], and [DL2, Chap. XXI-5]). Equation (30) implies that u_0^0 belongs to $\ker(C)$, i.e., $u_0^0(x,\mu,t) = u_0^0(x,t)$. We solve (31) and (32) using the solvability condition $R(C) = \ker(C)^\perp$, implied by the Fredholm alternative, since C is self-adjoint. Let us note that the solvability condition for (31) is trivially satisfied, and

$$u_1^0(x,\mu,t) = -\mu\frac{\partial u_0^0}{\partial x}(x,t) + c_1(x,t).$$

In what follows we shall take $c_1 = 0$. The solvability condition for (32),

$$\int_{-1}^{1}\Big(\frac{\partial u_0^0}{\partial t}(x,t) + \mu\frac{\partial u_1^0}{\partial x}(x,\mu,t) + \gamma u_0^0(x,t) - S\Big)\,d\mu = 0,$$

leads to the *diffusion equation* in Pb(1.2), namely,

$$\frac{\partial u_0^0}{\partial t} = \frac{1}{3}\frac{\partial^2 u_0^0}{\partial x^2} - \gamma u_0^0 + PS.$$

From (33), the initial layer term is given by

$$u_0^i\left(x,\mu,\frac{t}{\varepsilon^2}\right) = \frac{1}{2}\int_{-1}^{1} u_0^i(x,\mu',0)\,d\mu' \;+\; \mathrm{e}^{-\frac{t}{\varepsilon^2}}\;\left(u_0^i(x,\mu,0) - \frac{1}{2}\int_{-1}^{1} u_0^i(x,\mu',0)\,d\mu'\right).$$

Condition (28) yields $u_0^i(x,\mu,0) \in \ker(C)^\perp = \ker(Id - C)$, namely,

$$u_0^i(x,\mu,0) = f_0(x,\mu) - \frac{1}{2}\int_{-1}^{1} f_0(x,\mu')\ d\mu',$$

hence

$$u_0^i\left(x,\mu,\frac{t}{\varepsilon^2}\right) = \mathrm{e}^{-\frac{t}{\varepsilon^2}}\ \left(f_0(x,\mu) - \frac{1}{2}\int_{-1}^{1} f_0(x,\mu')\ d\mu'\right).$$

Consequently, (37) implies the *initial condition* for the diffusion equation in Pb(1.2):

$$u_0^0(x,0) = \frac{1}{2}\int_{-1}^{1} f_0(x,\mu')\ d\mu'.$$

To complete the analysis and show that u_0^0 is indeed the solution of Pb.(1.2), we treat now the boundary layer problem. Since u_0^0 does not depend on μ, and taking into account (34) and (39), the term u_0^b satisfies

$$\begin{cases} \mu\dfrac{\partial u_0^b}{\partial \xi} + Cu_0^b = 0 \quad \text{on } \left(-\dfrac{a}{\varepsilon},\dfrac{a}{\varepsilon}\right)\times(-1,1), \\ u_0^b\left(-\dfrac{a}{\varepsilon},\mu\right) = u_0^b\left(-\dfrac{a}{\varepsilon},-\mu\right), \quad \mu>0, \\ u_0^b\left(\dfrac{a}{\varepsilon},-\mu\right) = u_0^b\left(\dfrac{a}{\varepsilon},\mu\right), \quad \mu>0. \end{cases}$$

Then Lemma 2 implies that u_0^b is a constant, and condition (29) implies that $u_0^b = 0$. According to (35) and (38), the term u_1^b satisfies

(41)
$$\begin{cases} \mu\dfrac{\partial u_1^b}{\partial \xi} + Cu_1^b = 0 \quad \text{on } \left(-\dfrac{a}{\varepsilon},\dfrac{a}{\varepsilon}\right)\times(-1,1), \\ u_1^b\left(-\dfrac{a}{\varepsilon},\mu\right) = u_1^b\left(-\dfrac{a}{\varepsilon},-\mu\right) + 2\mu\dfrac{\partial u_0^0}{\partial x}(-a,t) + \beta u_0^0(-a,t) + S^-(\mu), \quad \mu>0, \\ u_1^b\left(\dfrac{a}{\varepsilon},-\mu\right) = u_1^b\left(\dfrac{a}{\varepsilon},\mu\right) - 2\mu\dfrac{\partial u_0^0}{\partial x}(a,t) + \beta u_0^0(a,t) + S^+(\mu), \quad \mu>0. \end{cases}$$

We note that u_1^b depends on the parameter t, which appears in the nonhomogeneous boundary terms. Let us define, for $\mu > 0$, the functions f and g as

$$f(\mu,t) = 2\mu\frac{\partial u_0^0}{\partial x}(-a,t) + \beta u_0^0(-a,t) + S^-(\mu), \quad g(\mu,t) = -2\mu\frac{\partial u_0^0}{\partial x}(a,t) + \beta u_0^0(a,t) + S^+(\mu).$$

According to Proposition 2, (41) admits some solutions u_1^b in $L^2\big((-a,a)\times(-1,1)\big)$ if and only if

$$\int_0^1 \mu\Big(f(\mu,t) + g(\mu,t)\Big)\ d\mu = 0. \tag{42}$$

Condition (42) is equivalent to

$$-\frac{2}{3}\frac{\partial u_0^0}{\partial x}(a,t) + \frac{\beta}{2}u_0^0(a,t) + \frac{2}{3}\frac{\partial u_0^0}{\partial x}(-a,t) + \frac{\beta}{2}u_0^0(-a,t) + \frac{1}{2}\overline{P}S^+ + \frac{1}{2}\overline{P}S^- = 0,$$

which leads to the *nonhomogeneous Robin BCs* in Pb(1.2). The boundedness of u_1^b in ε can be obtained from the solution of the Milne problem in semi-infinite geometry. The change of variables $\xi_\pm = \frac{a \mp x}{\varepsilon}$ implies

$$
\left\{
\begin{array}{l}
-\mu \dfrac{\partial u_1^{b+}}{\partial \xi_+} + C u_1^{b+} = 0 \quad \text{on } (0,\infty)\times(-1,1), \\
u_1^{b+}(0,-\mu) = u_1^{b+}(0,\mu) - 2\mu \dfrac{\partial u_0^0}{\partial x}(a,t) + \beta u_0^0(a,t) + S^+(\mu), \quad \mu > 0, \\
\lim\limits_{\xi_+ \to \infty} u_1^{b+}(\xi_+,\mu) = 0,
\end{array}
\right. \tag{43}
$$

$$
\left\{
\begin{array}{l}
\mu \dfrac{\partial u_1^{b-}}{\partial \xi_-} + C u_1^{b-} = 0 \quad \text{on } (0,\infty)\times(-1,1), \\
u_1^{b-}(0,\mu) = u_1^{b-}(0,-\mu) + 2\mu \dfrac{\partial u_0^0}{\partial x}(-a,t) + \beta u_0^0(-a,t) + S^-(\mu), \ \mu > 0, \\
\lim\limits_{\xi_- \to \infty} u_1^{b-}(\xi_-,\mu) = 0.
\end{array}
\right. \tag{44}
$$

Proposition 3 yields that (43) and (44) admit, respectively, a unique solution in $L^2((-1,1);L^\infty(0,\infty))$, and give separately the previous nonhomogeneous Robin BCs at a and $-a$. We deduce from

$$
\left\| u_1^b\left(\frac{x}{\varepsilon},\mu\right)\right\|^2_{L^2((-a,a)\times(-1,1))} \le a\Big(\|u_1^{b_+}\|^2_{L^2((-1,1);L^\infty(0,\infty))} + \|u_1^{b_-}\|^2_{L^2((-1,1);L^\infty(0,\infty))}\Big) \tag{45}
$$

that u_1^b is uniformly bounded in ε in $L^2((-a,a)\times(-1,1))$. Indeed

$$
\begin{aligned}
\int_{-1}^1 \int_0^a \left| u_1^{b+}\left(\frac{a-x}{\varepsilon},\mu\right)\right|^2 dx d\mu &= \int_{-1}^1 \varepsilon \int_0^{\frac{a}{\varepsilon}} \left|u_1^{b+}(\xi_+,\mu)\right|^2 d\xi_+ d\mu \\
&\le a \int_{-1}^1 \left(\sup_{\xi\in(0,\infty)} \left|u_1^{b+}(\xi,\mu)\right|\right)^2 d\mu.
\end{aligned}
$$

The same estimate holds for u_1^{b-}.

We notice that, according to Lemma 2 (see also Proposition 4), the term u_1^b is defined up to an arbitrary function of time that we denote $\sigma_\varepsilon(t)$. Then we have to replace in the asymptotic expansion the term u_1^b by $u_1^b + \sigma_\varepsilon(t)$, where u_1^b is uniquely defined with u_1^{b+} and u_1^{b-}. According to (36) and (40), the term u_2^b satisfies

$$
\left\{
\begin{array}{l}
\mu \dfrac{\partial u_2^b}{\partial \xi} + C u_2^b = 0 \quad \text{on } \left(-\dfrac{a}{\varepsilon},\dfrac{a}{\varepsilon}\right)\times(-1,1), \\
u_2^b\left(-\dfrac{a}{\varepsilon},\mu\right) = u_2^b\left(-\dfrac{a}{\varepsilon},-\mu\right) + \beta u_1^b\left(-\dfrac{a}{\varepsilon},-\mu\right) + \beta\sigma_\varepsilon + \beta\mu \dfrac{\partial u_0^0}{\partial x}(-a) \\
\qquad\qquad + u_2^0(-a,-\mu) - u_2^0(-a,\mu), \quad \mu > 0, \\
u_2^b\left(\dfrac{a}{\varepsilon},-\mu\right) = u_2^b\left(\dfrac{a}{\varepsilon},\mu\right) + \beta u_1^b\left(\dfrac{a}{\varepsilon},\mu\right) + \beta\sigma_\varepsilon - \beta\mu \dfrac{\partial u_0^0}{\partial x}(a) \\
\qquad\qquad + u_2^0(a,\mu) - u_2^0(a,-\mu), \quad \mu > 0.
\end{array}
\right. \tag{46}
$$

Of course, the nonhomogeneous boundary terms depend also on the time t. According to Proposition 2, (46) admits some solutions in $L^2\big((-a,a)\times(-1,1)\big)$ if and only if

$$\sigma_\varepsilon(t) = -\frac{1}{\beta}\int_0^1 \mu\left(\beta\mu\frac{\partial u_0^0}{\partial x}(-a,t) + \beta u_1^b\left(-\frac{a}{\varepsilon},-\mu,t\right) + u_2^0(-a,-\mu,t) - u_2^0(-a,\mu,t)\right.$$
$$\left. - \beta\mu\frac{\partial u_0^0}{\partial x}(a,t) + \beta u_1^b\left(\frac{a}{\varepsilon},\mu,t\right) + u_2^0(a,\mu,t) - u_2^0(a,-\mu,t)\right)\, d\mu.$$

As before, the boundedness of u_2^b in ε is related to the Milne problem in semi-infinite geometry. With $\sigma_\varepsilon = \sigma_\varepsilon^+ + \sigma_\varepsilon^-$, the terms u_2^{b+} and u_2^{b-} satisfy

$$\text{(47)}\qquad \begin{cases} -\mu\dfrac{\partial u_2^{b+}}{\partial \xi_+} + C u_2^{b+} = 0 \quad \text{on } (0,\infty)\times(-1,1), \\ u_2^{b+}(0,-\mu) = u_2^{b+}(0,\mu) + \beta u_1^b\left(\dfrac{a}{\varepsilon},\mu\right) + \beta\sigma_\varepsilon^+ - \beta\mu\dfrac{\partial u_0^0}{\partial x}(a) \\ \qquad\qquad + u_2^0(a,\mu) - u_2^0(a,-\mu), \quad \mu>0, \\ \lim\limits_{\xi_+\to\infty} u_2^{b+}(\xi_+,\mu) = 0, \end{cases}$$

$$\text{(48)}\qquad \begin{cases} \mu\dfrac{\partial u_2^{b-}}{\partial \xi_-} + C u_2^{b-} = 0 \quad \text{on } (0,\infty)\times(-1,1), \\ u_2^{b-}(0,\mu) = u_2^{b-}(0,-\mu) + \beta u_1^b\left(-\dfrac{a}{\varepsilon},-\mu\right) + \beta\sigma_\varepsilon^- + \beta\mu\dfrac{\partial u_0^0}{\partial x}(-a) \\ \qquad\qquad + u_2^0(-a,-\mu) - u_2^0(-a,\mu), \quad \mu>0, \\ \lim\limits_{\xi_-\to\infty} u_2^{b-}(\xi_-,\mu) = 0. \end{cases}$$

Proposition 3 yields that (47) and (48) admit, respectively, a unique solution in $L^2((-1,1);L^\infty(0,\infty))$ if and only if the two following conditions are satisfied:

$$\sigma_\varepsilon^+(t) = -\frac{1}{\beta}\int_0^1 \mu\left(-\beta\mu\frac{\partial u_0^0}{\partial x}(a,t) + \beta u_1^b\left(\frac{a}{\varepsilon},\mu,t\right) + u_2^0(a,\mu,t) - u_2^0(a,-\mu,t)\right)\, d\mu,$$
$$\sigma_\varepsilon^-(t) = -\frac{1}{\beta}\int_0^1 \mu\left(\beta\mu\frac{\partial u_0^0}{\partial x}(-a,t) + \beta u_1^b\left(-\frac{a}{\varepsilon},-\mu,t\right) + u_2^0(-a,-\mu,t) - u_2^0(-a,\mu,t)\right)\, d\mu.$$

As before, the estimate

$$\text{(49)}\qquad \left\| u_2^b\left(\frac{x}{\varepsilon},\mu\right)\right\|^2_{L^2((-a,a)\times(-1,1))} \le a\Big(\|u_2^{b+}\|^2_{L^2((-1,1);L^\infty(0,\infty))} + \|u_2^{b-}\|^2_{L^2((-1,1);L^\infty(0,\infty))}\Big)$$

yields that u_2^b is uniformly bounded in ε in $L^2((-a,a)\times(-1,1))$, where u_2^b is uniquely defined.

At this stage, all the functions which appear in the asymptotic expansion are well defined if u_0^0 is the solution of Pb(1.2). We estimate now the remainder w_ε, which satisfies the equation

$$\text{(50)}\qquad \begin{cases} \dfrac{\partial w_\varepsilon}{\partial t}(x,\mu,t) = -\varepsilon^{-1}\mu\dfrac{\partial w_\varepsilon}{\partial x} - \varepsilon^{-2}C w_\varepsilon - \gamma w_\varepsilon + \Psi_\varepsilon^0 \quad \text{on } (-a,a)\times(-1,1)\times\mathbb{R}^+, \\ w_\varepsilon(-a,\mu,t) = (1+\beta\varepsilon)\, w_\varepsilon(-a,-\mu,t) + \Psi_\varepsilon^{b-}(\mu,t), \quad \mu>0, \\ w_\varepsilon(a,-\mu,t) = (1+\beta\varepsilon)\, w_\varepsilon(a,\mu,t) + \Psi_\varepsilon^{b+}(\mu,t), \quad \mu>0, \\ w_\varepsilon(x,\mu,0) = \Psi_\varepsilon^i(x,\mu), \end{cases}$$

where

$$\Psi^0_\varepsilon(x,\mu,t) = -\varepsilon\left(\frac{\partial u_1^0}{\partial t} + \varepsilon\frac{\partial u_2^0}{\partial t} + \mu\frac{\partial u_2^0}{\partial x} + \gamma u_1^0 + \varepsilon\gamma u_2^0\right) - \gamma u_0^i\left(\frac{t}{\varepsilon^2}\right) - \frac{\mu}{\varepsilon}\frac{\partial u_0^i}{\partial x}\left(\frac{t}{\varepsilon^2}\right)$$
$$- \varepsilon\left(\frac{\partial u_1^b}{\partial t}\left(\frac{x}{\varepsilon}\right) + \gamma u_1^b\left(\frac{x}{\varepsilon}\right) + \frac{\partial \sigma_\varepsilon}{\partial t}(t) + \gamma\sigma_\varepsilon(t) + \varepsilon\frac{\partial u_2^b}{\partial t}\left(\frac{x}{\varepsilon}\right) + \varepsilon\gamma u_2^b\left(\frac{x}{\varepsilon}\right)\right),$$

$$\Psi^{b-}_\varepsilon(\mu,t) = \varepsilon^3\beta u_2^0(-a,-\mu,t) + \varepsilon^3\beta u_2^b\left(-\frac{a}{\varepsilon},-\mu,t\right),$$

$$\Psi^{b+}_\varepsilon(\mu,t) = \varepsilon^3\beta u_2^0(a,\mu,t) + \varepsilon^3\beta u_2^b\left(\frac{a}{\varepsilon},\mu,t\right),$$

$$\Psi^i_\varepsilon(x,\mu) = -\varepsilon\left(u_1^0(x,\mu,0) + \sigma_\varepsilon(0) + \varepsilon u_2^0(x,\mu,0) + u_1^b\left(\frac{x}{\varepsilon},\mu,0\right) + \varepsilon u_2^b\left(\frac{x}{\varepsilon},\mu,0\right)\right).$$

Let $\phi_\varepsilon(t)$ be an extension given by Lemma 1 of $\varepsilon^{-3}\Psi^{b-}_\varepsilon$ and $\varepsilon^{-3}\Psi^{b+}_\varepsilon$. We show in what follows that the extension ϕ_ε is uniformly bounded in ε in W^2. The function $W_\varepsilon = w_\varepsilon - \varepsilon^3\phi_\varepsilon$ satisfies

(51)
$$\begin{cases} \dfrac{\partial W_\varepsilon}{\partial t}(x,\mu,t) = -\varepsilon^{-1}\mu\dfrac{\partial W_\varepsilon}{\partial x} - \varepsilon^{-2}CW_\varepsilon - \gamma W_\varepsilon + \overline{\Psi}^0_\varepsilon \quad \text{on } (-a,a)\times(-1,1)\times\mathbb{R}^+, \\ W_\varepsilon(-a,\mu,t) = (1+\beta\varepsilon)\ W_\varepsilon(-a,-\mu,t), \quad \mu>0, \\ W_\varepsilon(a,-\mu,t) = (1+\beta\varepsilon)\ W_\varepsilon(a,\mu,t), \quad \mu>0, \\ W_\varepsilon(x,\mu,0) = \overline{\Psi}^i_\varepsilon(x,\mu), \end{cases}$$

where $\overline{\Psi}^i_\varepsilon(x,\mu) = \Psi^i_\varepsilon(x,\mu) - \varepsilon^3\phi_\varepsilon(x,\mu,0)$, and

$$\overline{\Psi}^0_\varepsilon = \Psi^0_\varepsilon - \varepsilon^3\frac{\partial\phi_\varepsilon}{\partial t} - \varepsilon^2\mu\frac{\partial\phi_\varepsilon}{\partial x} - \varepsilon C\phi_\varepsilon - \varepsilon^3\gamma\phi_\varepsilon = \Psi^0_\varepsilon - \varepsilon^3\frac{\partial\phi_\varepsilon}{\partial t} - \varepsilon C\phi_\varepsilon - \varepsilon^2\phi_\varepsilon\big(\varepsilon\gamma - 1\big).$$

We note that the semigroup generated by $T_\varepsilon - \gamma$, where T_ε is defined in (2), with a restitution coefficient $\alpha = 1+\beta\varepsilon$, is exponentially stable, due to the absorption γ. Indeed, in the limit $\varepsilon\to 0$, the exponential bound $w(1+\beta\varepsilon,\varepsilon)$ given in Proposition 1 is dominated by the absorption coefficient γ (see Remark 1). Finally, the estimate (3) and the Duhamel formula yield

$$\|W_\varepsilon\|_{L^2} \le (1+\beta)\|\overline{\Psi}^i_\varepsilon\|_{L^2} + (1+\beta)\int_0^t \|\overline{\Psi}^0_\varepsilon(s)\|_{L^2}ds.$$

Last, lengthy but straightforward calculations yield the estimates (see [PT])

$$\|\overline{\Psi}^i_\varepsilon\|_{L^2} \le \varepsilon M, \quad \|\overline{\Psi}^0_\varepsilon(t)\|_{L^2} \le \varepsilon M\mathrm{e}^{\delta t} \quad \forall t\ge 0,$$

which conclude the proof. ∎

5. Discussion. Theorem 1 remains true when the BCs are infinitesimally partially absorbing. In this case, $\beta<0$ and we can take $\gamma=0$ and obtain a diffusion equation with *stable* Robin BCs. While the proof is simpler, due to the fact that the transport operator (2) is now dissipative, the procedure to solve the boundary layer

problem is similar to the one presented here. In particular, we still need to expand the boundary layer term to the second order, because we are not able, due to the BCs, to obtain an estimate of the remainder directly from (50). This difficulty is overcome by using the superposition principle and considering an extension of the boundary sources. To control the quadratic singularity in ε the equation, we need the boundary sources to be of order three in ε, which can be obtained with an asymptotic expansion of the boundary layer term up to order two.

If the transport problem has finite (unscaled) partially absorbing BCs ($\alpha < 1$), the diffusion limit with Dirichlet BCs is obtained without any of the boundary layer complications.

Similar results can be proven for the transport problem set in $L^p((-a,a)\times(-1,1))$, $1 \le p < \infty$, by using the same scaling of the BCs. The case $p = 1$ is still open, but—at least formally—it appears that in order to prove the result for $p = 1$ one would need to resort to a different scaling of the BCs, namely, $\alpha = 1 + \beta\, \varepsilon^2$. This aspect of the problem is under investigation.

Finally, we remark on a specific feature of this problem within the P_1-approximation framework. In this approximation, which is widely used in neutron transport theory, the BCs are written as the difference between the (P_1-approximated values of the) incoming and outgoing currents at the boundaries. This leads to certain inconsistencies in the BCs for the diffusion limit, in which the notion of current is not well defined. To avoid these inconsistencies, one has to use the exact formulation and impose the (exact) BCs by using the distribution function itself or its integral (particle density), without resorting to the notion of (approximate) current. We are indebted to Profs. M. M. R. Williams and E. Larsen for independently raising the issue of this discrepancy between the exact and P_1-approximation results.

Acknowledgment. We thank M. Mokhtar-Kharroubi for helpful discussions.

REFERENCES

[BM] J. Banasiak and J. R. Mika, *Singularly Perturbed Evolution Equations with Applications to Kinetic Theory*, Adv. Math. Appl. Sci. 34, World Scientific, Singapore, 1995.

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

[BP] R. Beals and V. Protopopescu, *Abstract time dependent transport equations*, J. Math. Anal. Appl., 121 (1987), pp. 370–405.

[BLP] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Boundary layers and homogenization of transport processes*, Publ. Res. Inst. Math. Sci., 15 (1979), pp. 53–157.

[BT] G. Borgioli and S. Totaro, *3D-streaming operator with multiplying boundary conditions: Semigroup generation properties*, Semigroup Forum, 55 (1997), pp. 110–117.

[B1] M. Boulanouar, *Un modèle de Rotemberg avec la loi a mémoire parfaite*, C. R. Acad. Sci. Paris Sér. I Math., 327 (1998), pp. 965–968.

[B2] M. Boulanouar, *Le transport neutronique avec des conditions aux limites générales*, C. R. Acad. Sci. Paris Sér. I Math., 329 (1999), pp. 121–124.

[B3] M. Boulanouar, *L'opérateur d'advection.* I. *Existence d'un C_0 semi-groupe*, Transp. Theory Statist. Phys., 31 (2002), pp. 169–176.

[BE] M. Boulanouar and H. Emamirad, *A transport equation in cell population dynamics*, Differential Integral Equations, 13 (2000), pp. 125–144.

[DL1] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology. Vol.* 2, Springer-Verlag, Berlin, 1988.

[DL2] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology. Vol.* 6, Springer-Verlag, Berlin, 1993.

[GMP] W. Greenberg, C. van der Mee, and V. Protopopescu, *Boundary Value Problems in Abstract Kinetic Theory*, Birkhäuser-Verlag, Basel, 1987.

[HM] G. J. Habetler and B. J. Matkowsky, *Uniform asymptotic in transport theory with small mean free path and the diffusion approximation*, J. Math. Phys., 16 (1975), pp. 846–854.

[K] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1980.

[KL] J. B. Keller and E. W. Larsen, *Asymptotic solution of neutron transport problems for small mean free paths*, J. Math. Phys., 15 (1974), pp. 75–81.

[LaM] K. Latrach and M. Mokhtar-Kharroubi, *Spectral analysis and generation results for streaming operators with multiplying boundary conditions*, Positivity, 3 (1999), pp. 273–296.

[LoM] B. Lods and M. Mokhtar-Kharroubi, *On the theory of growing cell populations with zero minimum cycle length*, J. Math. Anal. Appl., 266 (2002), pp. 70–99.

[MaT] S. Mancini and S. Totaro, *Particle transport problems with general multiplying boundary conditions*, Transport Theory Statist. Phys., 27 (1998), pp. 159–176.

[M1] M. Mokhtar-Kharroubi, *On the Diffusion Approximation for Neutron Transport*, manuscript.

[M2] M. Mokhtar-Kharroubi, *Mathematical Topics in Neutron Transport. New Aspects*, Ser. Adv. Math. Appl. Sci. 46, World Scientific, Singapore, 1997.

[MoT] M. Mokhtar-Kharroubi and L. Thevenot, *On the diffusion theory of neutron transport on the torus*, Asymptot. Anal., 30 (2002), pp. 273–300.

[MPT] M. Mokhtar-Kharroubi, V. Protopopescu, and L. Thevenot, *On the singular limit of a model transport semigroup*, Math. Methods Appl. Sci., 23 (2000), pp. 1301–1322.

[Pa] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[P] V. Protopopescu, *On the spectrum of the linear transport operator in a semi-infinite medium*, J. Phys. A, 9 (1976), pp. 1925–1937.

[PT] V. Protopopescu and L. Thevenot, *Diffusion Approximation for a Neutron Transport Equation with Multiplying Boundary Conditions*, Preprint 2003:10, Chalmers University of Technology, Göteborg, Sweden, 2003.

[R] M. Rotemberg, *Transport theory for growing cell populations*, J. Theoret. Biol., 103 (1983), pp. 181–199.

[Sa] R. Santos, *Analyse asymptotique des équations de transport dans le cas d'évolution*, Proc. Indian Acad. Sci. Math. Sci., 90 (1981), pp. 219–227.

[Se] R. Sentis, *Approximation and homogenization of a transport process*, SIAM J. Appl. Math., 39 (1980), pp. 134–141.

[V] J. Voigt, *Functional Analytic Treatment of the Initial Boundary Value Problem for Collisionless Gases*, Habilitationsschrift, Universität München, 1981.

[Y] K. Yosida, *Functional Analysis*, Springer-Verlag, Berlin, 1974.

SIAM J. APPL. MATH.
Vol. 65, No. 5, pp. 1677–1707

REGULARIZATION BY KINETIC UNDERCOOLING OF BLOW-UP IN THE ILL-POSED STEFAN PROBLEM*

J. R. KING† AND J. D. EVANS‡

Abstract. It is well known that the one-dimensional supercooled Stefan problem possesses solutions that blow up in finite time. The asymptotics of such solutions have been analyzed by Herrero and Velazquez [*European J. Appl. Math.*, 7 (1996), pp. 119–150]. Here we consider the effect of kinetic undercooling as a regularizing mechanism to prevent the formation of such singularities and study the continuation of the solution through the "near blow-up" regime. The asymptotics of solutions and interfaces are described for small values of the kinetic undercooling parameter. It is shown that, in this limit, the interface jumps over an interval determined by the latent heat and by the initial data. Specifically, in dimensionless variables, if the temperature profile at blow-up is denoted by $u(x, t_c^-)$, where t_c is the finite blow-up time, then the interface jumps over the interval in which $u(x, t_c^-) < -\lambda$, where λ is the latent heat.

Key words. Stefan problem, kinetic undercooling, matched asymptotic expansions

AMS subject classifications. 35B40, 35R35, 80A22

DOI. 10.1137/04060528X

1. Introduction.

The classical Stefan problem arises as an important model in conductive heat transfer problems that involve two phases separated by an unknown phase boundary. In its usual context, the governing equations for temperature are parabolic in both phases, and at the phase boundary the temperature is usually specified as being constant, with a release of latent heat on melting or an uptake on solidification; this latter condition is commonly referred to as the Stefan condition. Further, there is usually a sign requirement (relating to well-posedness) on the initial and/or boundary data that the temperature of the material in the solid phase be negative and that in the liquid phase be positive. Although the problem will be described in the notation and context of heat transfer, mathematically equivalent problems also arise in mass transfer applications. Accordingly, an appropriate generalization of the Stefan condition at the phase boundary will be considered in view of its relevance to such problems (see, for example, [2] or [27] and the references therein). In this context existence, uniqueness, and well-posedness, in both the classical and weak sense, are known (see, for example, Fasano and Primicerio [16], [17], [18] and Elliott and Ockendon [13]). However, as shown by Sherman [33] and characterized by Fasano, Primicerio, and Lacey [20], if the sign requirement on the temperature is violated, with the liquid being sufficiently supercooled (or the solid sufficiently superheated), then a solution may still exist locally in time, but now blow-up can occur in a finite time (i.e., the temperature develops an unbounded first derivative at the phase boundary (interface) or, equivalently, the speed of the interface becomes unbounded).

Several authors have considered the nature of the singularity which appears as blow-up occurs and the possible continuation of solutions after blow-up. In particular,

*Received by the editors March 16, 2004; accepted for publication (in revised form) December 13, 2004; published electronically July 13, 2005.
http://www.siam.org/journals/siap/65-5/60528.html
†Theoretical Mechanics Section, University of Nottingham, Nottingham, NG7 2RD, UK (john.king@nottingham.ac.uk).
‡Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK (masjde@maths.bath.ac.uk).

Herrero and Velazquez [25] have obtained asymptotic behaviors for the interface and blow-up profile, while Gurtin [24] and Götz and Zaltzman [23] argue that under certain conditions the interface jumps at the blow-up time. Regularization of the model is required in order for the phase change to be described to completion. Here we consider the one-phase, one-dimensional problem and hence consider kinetic undercooling as the regularizing mechanism (in higher space dimensions, surface energy effects could be included). An alternative (less physical) regularization approach is discussed in Fasano et al. [21], where a Baiocchi-type transformation is used to give the Crank–Gupta (oxygen diffusion-consumption) moving boundary problem; the constraint of solution nonnegativity for this problem provides the required regularizing mechanism.

The layout of the paper is as follows. In section 2, the statement of the one-phase problem is carefully derived as a limit case of the two-phase formulation. In section 3 the problem statement is given, together with integral statements and the conditions for blow-up. Section 4 derives the same blow-up asymptotics as those obtained by Herrero and Velazquez [25], albeit through the use of an alternative (in our case less rigorous but perhaps more transparent and comprehensive) approach. The regularization asymptotics are then described in section 5, with the main timescale for the interface jump described in section 5.2 and the consequent recovery of the classical Stefan problem in section 5.3. The case of zero segregation coefficient is discussed in section 6, where kinetic undercooling is shown not to prevent blow-up. Finally, in section 7, the asymptotics are compared to numerical solutions of the full problem.

2. Derivation of the one-phase Stefan problem. We first describe the derivation of the one-phase Stefan problem in multiple dimensions with kinetic undercooling and surface tension effects included. The governing equations for the two-phase multidimensional problem may be written as

$$\frac{\partial}{\partial t}(\rho c_i u_i) = \nabla.(K_i \nabla u_i), \quad \mathbf{x} \in D_i(t), \quad i = 1, 2, \quad D = D_1 \cup D_2,$$

where $i = 1$ denotes the liquid phase, $i = 2$ the solid phase, and $u_i(\mathbf{x}, t), K_i$, and c_i the temperature, conductivity, and specific heat, respectively, in the corresponding phase. The density ρ is taken to be the same in both phases. We present the problem in terms of heat transfer but, as indicated in (for instance) Evans and King [14], the same formulation arises in problems of mass transfer for which the one-phase limit discussed below is more often relevant.

The conditions on the moving phase boundary $F(\mathbf{x}, t) = 0$ are the Stefan condition

$$\left[K_i \frac{\partial u_i}{\partial n}\right]_{i=2}^{i=1} = -\rho v_n \left([c_i u_i]_{i=2}^{i=1} + L\right) \tag{2.1}$$

and temperature continuity (in view of the possibility of segregation in mass transfer applications, we allow greater generality by introducing a constant "partition coefficient" μ), together with kinetic undercooling and surface tension effects, i.e.,

$$u_1 = \frac{u_2}{\mu} = \alpha\kappa + \beta v_n. \tag{2.2}$$

Here $\mathbf{n}$ denotes the outward normal (from the liquid phase into the solid), v_n is the velocity of the phase boundary in that direction, $\partial/\partial n$ the outward normal derivative, α the surface tension coefficient, κ the mean curvature, β the kinetic undercooling coefficient, and L the latent heat per unit mass at the equilibrium temperature u_m

(cf. Charach, Zaltzman, and Götz [7], for example), and we choose units such that the equilibrium melting temperature is zero. We assume that appropriate initial conditions and boundary conditions on the fixed boundary ∂D are given to complete the problem statement.

We now introduce the dimensionless variables

$$\tilde{\mathbf{x}} = \frac{\mathbf{x}}{\ell}, \quad \tilde{t} = \frac{K_1}{\rho c_1 \ell^2} t, \quad \tilde{u}_i = \frac{u_i}{U} \quad \tilde{n} = \frac{n}{\ell}, \quad \tilde{\kappa} = \ell\kappa, \quad \tilde{v}_n = \frac{\rho c_1 \ell}{K_1} v_n,$$

where ℓ is a typical length scale and U is a representative temperature scale. We define the dimensionless parameters

$$\lambda = \frac{L}{c_1 U}, \quad \epsilon = \frac{\beta K_1}{\rho c_1 U \ell}, \quad \tilde{\alpha} = \frac{\alpha}{U\ell}, \quad c = \frac{c_2}{c_1}, \quad K = \frac{K_2}{K_1};$$

λ is commonly termed the Stefan number. For simplicity we have taken ρ, c_i, and K_i to be constant. Dropping the tildes then gives the dimensionless formulation

$$\frac{\partial u_1}{\partial t} = \nabla^2 u_1, \quad \mathbf{x} \in D_1(t), \quad \frac{\partial u_2}{\partial t} = \frac{K}{c}\nabla^2 u_2, \quad \mathbf{x} \in D_2(t), \tag{2.3}$$

$$\text{on} \quad F(\mathbf{x}, t) = 0 \qquad u_1 = \frac{u_2}{\mu} = \alpha\kappa + \epsilon v_n, \tag{2.4}$$

$$\frac{\partial u_1}{\partial n} - K\frac{\partial u_2}{\partial n} = -v_n\left((1 - \mu c)u_1 + \lambda\right) \tag{2.5}$$

subject to suitable nondimensionalized initial and fixed boundary conditions, in particular $u_2(\mathbf{x}, 0) = U_2(\mathbf{x})$. When $\epsilon \neq 0$, a truly one-phase problem is not possible in either the freezing or melting regime, since u_2 is nonzero on the moving boundary, leading to temperature variations in the solid. Moreover, even in the absence of kinetic undercooling and surface tension ($\epsilon = \alpha = 0$), only the melting regime $v_n > 0$ can be reduced to a bona fide one-phase problem, whereby $U_2 \equiv 0$ implies $u_2 \equiv 0$ and (2.10) below holds exactly. This contrasts with the freezing regime $v_n < 0$ when temperature variations in the solid result even when $U_2 \equiv 0$, as described below. Nevertheless, we can obtain a one-phase formulation asymptotically by considering the limit $K \to 0$, being careful to distinguish the cases of melting and freezing. To leading order in the solid, away from the moving interface, we then have, if $\mathbf{x}$ has not been visited by the interface, that

$$u_2 = U_2(\mathbf{x}), \tag{2.6}$$

where $U_2(\mathbf{x})$ is the initial temperature. However, there also exists an interior layer near the moving interface in which significant temperature changes occur in the solid. The scaling for this interior layer is

$$\mathbf{x} = \mathbf{x}_0(t) + \theta\hat{n}\mathbf{n},$$

where $F(\mathbf{x}_0, t) = 0$, so $\mathbf{x}_0$ lies on the moving boundary. At leading order within this layer in the solid we obtain from (2.3) that

$$\frac{\partial^2 u_2}{\partial \hat{n}^2} = -cv_n\frac{\partial u_2}{\partial \hat{n}} \tag{2.7}$$

subject to (in view of (2.4))

$$u_2 = \mu u_1(\mathbf{x}_0, t) \qquad \text{at } \hat{n} = 0,$$

which gives for $\hat{n} = O(1)$ that

$$u_2 \sim \begin{cases} (\mu u_1(\mathbf{x}_0, t) - U_2(\mathbf{x}_0))e^{-cv_n\hat{n}} + U_2(\mathbf{x}_0) & \text{if} \quad v_n > 0, \\ \mu u_1(\mathbf{x}_0, t) & \text{if} \quad v_n \leq 0, \end{cases} \tag{2.8}$$

where the sign in (2.7) implies that we can impose $u_2 \to U_2(\mathbf{x}_0)$ as $\hat{n} \to \infty$ (to match with (2.6)) in the former case but not in the latter (from which we have to exclude the possibility of exponential growth). This highlights the fact that the regimes $v_n > 0$ and $v_n < 0$ must be carefully distinguished. We do not consider cases in which the interface reverses direction; if it does, then the requirement of matching with (2.6) may need modification due to an earlier visit by the moving boundary. For $v_n < 0$, the leading-order equation in D_2, in the limit $K \to 0$ with $|\mathbf{x} - \mathbf{x}_0| = O(1)$, namely,

$$\frac{\partial u_2}{\partial t} = 0,$$

implies on matching with (2.8) that

$$u_2 = \mu u_1(\mathbf{x}, \omega(\mathbf{x})) \quad \text{for } \mathbf{x} \in D_2(t) \backslash D_2(0), \tag{2.9}$$

where we write $F(\mathbf{x}, t) = 0$ in the form $t = \omega(\mathbf{x})$ (of course, $D_2(t) \backslash D_2(0)$ being non-empty corresponds to $v_n < 0$). Since $\frac{\partial}{\partial \hat{n}} = K \frac{\partial}{\partial n}$, the leading-order Stefan condition (2.5) for the liquid becomes

$$\frac{\partial u_1}{\partial n} + v_n u_1 = v_n(cU_2 - \lambda) \quad \text{if} \quad v_n > 0 \tag{2.10}$$

and

$$\frac{\partial u_1}{\partial n} + v_n(1 - \mu c)u_1 = -\lambda v_n \quad \text{if} \quad v_n < 0. \tag{2.11}$$

In [14] we were concerned with the particular case $U_2 \equiv 0$ in the melting regime $v_n > 0$ (and it should be emphasized that the one-phase model derivation given there applies only when $v_n > 0$), in which the Stefan condition (2.10) simplifies to

$$\frac{\partial u_1}{\partial n} + v_n u_1 = -\lambda v_n. \tag{2.12}$$

Here we consider the freezing regime $v_n < 0$. For brevity we henceforth take $c = 1$, or equivalently redefine μ, so the Stefan condition (2.11) reads

$$\frac{\partial u_1}{\partial n} + v_n(1 - \mu)u_1 = -\lambda v_n. \tag{2.13}$$

In the case $\mu = 0$ (complete segregation into the fluid), this reduces to (2.12) and, as is to be expected, conservation of total heat within the fluid (sensible and latent) then follows, as discussed in [14]. For $\mu = 1$ (u_i continuous at the interface), however, the condition reduces to

$$\frac{\partial u_1}{\partial n} = -\lambda v_n. \tag{2.14}$$

It is the freezing (supercooled liquid) regime that has received the most atten-

tion in the literature, as it is this case which is ill-posed in the absence of kinetic undercooling and surface tension effects (as of course is the melting regime of a superheated solid). The effect of kinetic undercooling has been considered by numerous authors. Visintin [35] and Xie [36] established existence and uniqueness of the two-phase problem in one space dimension. Dewynne et al. [11] investigated the behavior of similarity (including traveling wave) solutions of the one-phase problem. Stability was investigated by Coriell and Parker [9] and Coriell and Sekerka [10]. Charach and Zaltzman [5], [6] and Charach, Zaltzman, and Götz [7] developed large-time asymptotic solutions.

The effects of surface tension have received even more attention. For example, Mullins and Sekerka [29] analyzed the stability of supersaturated solutes and supercooled melts in a spherical geometry. Chadam and Ortoleva [3] discussed the stability properties of planar melting and solidification (with and without surface tension). Zhu, Peirce, and Chadam [38] extended these linear stability results to earlier times. Chadam, Howison, and Ortoleva [4] considered the radially symmetric case and showed existence and uniqueness if the supersaturation is not too large, but nonexistence in finite time if the far-field supersaturation is sufficiently pronounced. Linear stability of a radially symmetric similarity solution is also examined. Existence and uniqueness results for small surface tension in the two-phase problem are discussed in Friedman and Reitich [22], while Scianna [32] showed global existence and determined the large-time asymptotics in the radially symmetric case.

The simultaneous effects of surface tension and kinetic undercooling have been considered by several authors. Schaefer and Glicksman [30] considered stability of the spherical supercooled case, while the two-dimensional case was considered by Umantsev and Davis [34]. Doole [12] examined the stability of similarity and traveling wave solutions for the two-dimensional, one-phase problem. Existence and uniqueness results were obtained by Chen and Reitich [8]. Yi [37] demonstrated existence of solutions for the Hele–Shaw problem (obtained from the Stefan problem in the limit of small specific heat) in the one-phase supercooled case. Surface tension is often incorporated through a Gibbs–Thomson term, as in (2.2). However, a term of the form $e^{\sigma\kappa}$, based on Nernst's law, was treated by Abergel, Hilhorst, and Issard-Roch [1] and Scheid [31]. Analogous considerations for nonlinear kinetic undercooling were discussed in Evans and King [15].

Of particular relevance to the work presented here is that of Gurtin [24], who postulates a jump in the position of the phase interface when blow-up occurs in the supercooled one-phase, one-dimensional case in the absence of kinetic undercooling. Götz and Zaltzman [23] studied the one-dimensional two-phase problem with kinetic undercooling and showed that as the kinetic modulus $\beta \to 0$ (in our notation), the limiting solution contains a jump exactly equal to the interval in which $u_1(x, t_c^-) < -\lambda$, where t_c is the blow-up time. It is the asymptotics of this situation that is of most interest to us here.

3. Problem formulation. The one-dimensional, one-phase supercooled Stefan problem with linear kinetic undercooling can be stated in dimensionless terms as follows:

$$\text{in } 0 < x < s(t) \qquad \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \tag{3.1}$$

$$\text{on } x = s(t) \qquad u = \epsilon \dot{s}(t), \quad \frac{\partial u}{\partial x} + \dot{s}(t)(1-\mu)u = -\lambda \dot{s}(t), \tag{3.2}$$

$$\text{on } x = 0 \qquad \frac{\partial u}{\partial x} = 0, \tag{3.3}$$

$$\text{at } t = 0 \qquad s = 1, \quad u = u_{in}(x) \leq 0 \quad \text{for} \quad 0 \leq x \leq 1. \tag{3.4}$$

The suffix 1 on the temperature variable from the previous section has been dropped for brevity, and a Neumann condition has been taken on the fixed boundary $x = 0$. The length scale ℓ has been taken to be the width of the initial region occupied by the liquid, while a characteristic reference temperature U is provided by the initial temperature profile in the liquid.

The important dimensionless parameters are the kinetic undercooling parameter ϵ, the Stefan number λ, and the initial supercooling parameter Q, defined by

$$Q = \int_0^1 (u_{in}(x) + \lambda) dx. \tag{3.5}$$

In the absence of kinetic undercooling (i.e., for $\epsilon = 0$), $Q < 0$ is a necessary and sufficient condition for blow-up (see Sherman [33] and Fasano, Primicerio, and Lacey [20]). By this we mean that there exists a $t_c < \infty$ such that

$$\lim_{t \to t_c^-} s > 0 \quad \text{and} \quad \lim_{t \to t_c^-} \dot{s} = -\infty.$$

We are interested in describing the asymptotics of this case in the limit $\epsilon \to 0^+$, where kinetic undercooling has a regularizing effect.

The following integral statements hold:

$$\int_0^{s(t)} (u(x,t) + \lambda) dx = \mu\epsilon \int_0^t \dot{s}(\tau)^2 d\tau + Q \tag{3.6}$$

(the significance of positivity or otherwise of Q in the unregularized case $\epsilon = 0$ may be apparent from this expression) and

$$\int_0^{s(t)} x(u(x,t) + \lambda) dx = \int_0^1 x(u_{in}(x) + \lambda) dx + \int_0^t u(0,\tau) d\tau + \mu\epsilon \int_0^t s\dot{s}^2 d\tau - \epsilon(s-1). \tag{3.7}$$

It can be seen from (3.6) that in the case $\mu = 0$, kinetic undercooling will not prevent blow-up if $Q < 0$.

4. Blow-up of the unregularized problem. Results equivalent to those of Herrero and Velazquez [25] (and Herrero, Medina, and Velazquez [26], where the full asymptotic structure in the one-dimensional case is completed) are derived here through an alternative derivation using a Baiocchi-type transformation; they will be required subsequently for the analysis of the regularized case. Denoting the moving boundary by $t = \omega(x)$ (so that $\omega(s(t)) = t$) and setting $\epsilon = 0$, the transformation

$$w(x,t) = \int_{\omega(x)}^t u(x,t') dt' \tag{4.1}$$

gives the system

$$\text{in } 0 < x < s(t) \quad \frac{\partial w}{\partial t} = \frac{\partial^2 w}{\partial x^2} - \lambda, \tag{4.2}$$

$$\text{on } x = s(t) \qquad w = 0, \quad \frac{\partial w}{\partial x} = 0, \tag{4.3}$$

$$\text{on } x = 0 \qquad \frac{\partial w}{\partial x} = -Q, \tag{4.4}$$

$$\text{at } t = 0 \qquad s = 1, \quad w = \int_x^1 (x' - x)(u_{in}(x') + \lambda)dx' \quad \text{for} \quad 0 \le x \le 1. \tag{4.5}$$

Taking blow-up to occur at the location $x = x_c \in (0, 1)$ and the time $t = t_c < \infty$, the rescalings

$$y = \frac{(x - x_c)}{(t_c - t)^{\frac{1}{2}}}, \quad \tau = -\log(t_c - t), \quad L(\tau) = \frac{s(t) - x_c}{(t_c - t)^{\frac{1}{2}}} \tag{4.6}$$

(where L here is not to be confused with the latent heat which appeared earlier), with

$$w(x, t) = (t_c - t)\Phi(y, \tau), \tag{4.7}$$

transform (4.2)–(4.5) into the system

$$\text{in } -x_c e^{\tau/2} < y < L(\tau) \quad \frac{\partial \Phi}{\partial \tau} = \frac{\partial^2 \Phi}{\partial y^2} - \frac{y}{2}\frac{\partial \Phi}{\partial y} + \Phi - \lambda, \tag{4.8}$$

$$\text{on } y = L(\tau) \qquad \Phi = 0, \quad \frac{\partial \Phi}{\partial y} = 0, \tag{4.9}$$

$$\text{on } y = -x_c e^{\tau/2} \qquad \frac{\partial \Phi}{\partial y} = -Qe^{\tau/2}. \tag{4.10}$$

The asymptotic structure of this problem as $\tau \to \infty$ can be subdivided into two main regions, as follows. First, for $y = O(1)$ we introduce the expansion

$$\Phi \sim \lambda + A(\tau)\Phi_0(y) + \dot{A}(\tau)\Phi_1(y) \qquad \text{as } \tau \to \infty, \tag{4.11}$$

where $A(\tau) \equiv \Phi(0, \tau) - \lambda$ remains to be determined, with $\dot{A} \ll A \ll 1$ as $\tau \to \infty$. Thus

$$\frac{\partial^2 \Phi_0}{\partial y^2} - \frac{y}{2}\frac{\partial \Phi_0}{\partial y} + \Phi_0 = 0,$$

$$\frac{\partial^2 \Phi_1}{\partial y^2} - \frac{y}{2}\frac{\partial \Phi_1}{\partial y} + \Phi_1 = \Phi_0,$$

so that, without loss of generality (by excluding exponential growth as $y \to \infty$),

$$\Phi_0 = 1 - \frac{y^2}{2} \tag{4.12}$$

and

$$\frac{d}{dy}\left(e^{-y^2/4}\left(\Phi_0 \frac{d\Phi_1}{dy} - \Phi_1 \frac{d\Phi_0}{dy}\right)\right) = e^{-y^2/4}\Phi_0^2,$$

which implies

$$\Phi_0 \frac{d\Phi_1}{dy} - \Phi_1 \frac{d\Phi_0}{dy} = -\frac{1}{2}y^3 - y + 4e^{y^2/4}\int_{-\infty}^{y/2} e^{-\sigma^2} d\sigma,$$

and hence

$$\Phi_1 \sim -\frac{16\sqrt{\pi}}{y^3} e^{y^2/4} \qquad \text{as } y \to \infty. \tag{4.13}$$

Second, the scalings for the interior layer $z = O(1)$ at the moving boundary are

$$y = L + \frac{z}{L}, \qquad \Phi = \frac{1}{L^2}\Psi(z,\tau),$$

with $\Psi \sim \Psi_0(z)$ as $\tau \to \infty$ and

$$\frac{d^2\Psi_0}{dz^2} - \frac{1}{2}\frac{d\Psi_0}{dz} - \lambda = 0,$$

implying that

$$\Psi_0 = -2\lambda z - 4\lambda\left(1 - e^{z/2}\right). \tag{4.14}$$

Leading-order matching of (4.11)–(4.12) with (4.14) (in the limit $z \to -\infty$) yields

$$A \sim \frac{2\lambda}{L^2},$$

while matching of the exponential in (4.14) with (4.13) requires

$$\frac{16\sqrt{\pi}}{L^3} e^{L^2/4}\dot{A} \sim -\frac{4\lambda}{L^2},$$

so that

$$\frac{16\sqrt{\pi}}{L^4} e^{L^2/4}\dot{L} \sim 1. \tag{4.15}$$

This then gives

$$\frac{32\sqrt{\pi}}{L^5} e^{L^2/4} \sim \tau \qquad \text{as } \tau \to \infty$$

and thus

$$L \sim 2\left(\ln\tau\right)^{1/2}, \qquad A \sim \lambda/2\ln\tau \qquad \text{as } \tau \to \infty. \tag{4.16}$$

Since

$$u = \frac{\partial\Phi}{\partial\tau} + \frac{y}{2}\frac{\partial\Phi}{\partial y} - \Phi = \frac{\partial^2\Phi}{\partial y^2} - \lambda,$$

we have

$$u \sim -\lambda - A(\tau) \quad \text{for } y = O(1), \qquad u \sim -\lambda(1 - e^{z/2}) \quad \text{for } z = O(1). \tag{4.17}$$

Denoting the blow-up profile $u(x,t_c)$ by $u_c(x)$, it follows using (4.6) that

$$u_c(x) \sim -\lambda - \frac{\lambda}{2\ln\left(-\ln(x_c - x)\right)} \qquad \text{as } x \to x_c^-. \tag{4.18}$$

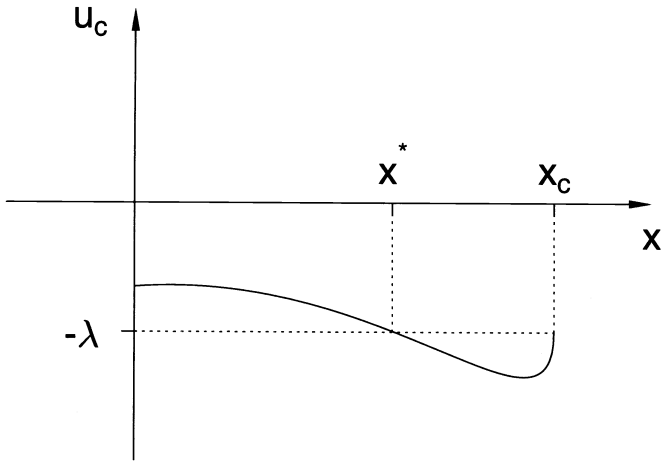


FIG. 1. *A schematic illustration of the generic blow-up profile* $u_c(x) = u(x, t_c)$ *which gives* $Q < 0$ *in* (4.19).

The profile is shown schematically in Figure 1. From (3.6), the blow-up profile satisfies

$$Q = \int_0^{x_c} (u_c(x) + \lambda)dx = \int_0^{x^*} (u_c(x) + \lambda)dx + Q_c, \tag{4.19}$$

where $u_c(x^*) = -\lambda$ (see Figure 1) and

$$Q_c = \int_{x^*}^{x_c} (u_c(x) + \lambda)dx.$$

As we shall see, $Q_c < 0$ represents the quantity of heat that the solid needs to absorb before the interface can be continued in the classical sense, the remainder, $Q - Q_c$, then being nonnegative (at least for profiles of the form shown in Figure 1). It is in effect the manner in which Q_c is removed from the liquid phase that is investigated in subsequent sections.

Nongeneric blow-up scenarios are also possible (see also [25]) and, since we shall need to exclude such a possibility in a particular situation in section 6 below, we record the first of these here. The analysis is in some respects simpler than that of the generic case above since there is no logarithmic dependence on τ. The expansion (4.11) is replaced by

$$\Phi \sim \lambda + e^{-\tau/2}\hat{\Phi}(y), \tag{4.20}$$

where

$$\hat{\Phi}(y) = A\left(y - \frac{y^3}{6}\right)$$

(the various modes can be identified as corresponding to solutions which are polynomials in y) and the value of the positive constant A depends on the initial data; the interior layer scalings then read

$$y = L + \frac{z}{L}, \qquad w = (t_c - t)^{4/3}\hat{\Psi}(z, \tau)$$

with

$$L \sim \left(\frac{6\lambda}{A}\right)^{1/3} e^{\tau/6}, \qquad \hat{\Psi} \sim \left(\frac{A}{6\lambda}\right)^{2/3} \hat{\Psi}_0$$

and

$$\frac{d^2\hat{\Psi}_0}{dz^2} - \frac{1}{3}\frac{d\hat{\Psi}_0}{dz} - \lambda = 0,$$

so that

$$\hat{\Psi}_0 = -3\lambda z - 9\lambda\left(1 - e^{z/3}\right).$$

It follows in particular that

$$u \sim -\lambda + A(x_c - x) \tag{4.21}$$

for $y = O(1)$, with (4.21) also furnishing the local behavior of $u_c(x)$ as $x \to x_c^-$, and that

$$s(t) \sim x_c + \left(\frac{6\lambda}{A}\right)^{1/3} (t_c - t)^{1/3} \qquad \text{as } t \to t_c^-. \tag{4.22}$$

5. The regularized problem with $\mu > 0$. The blow-up described in the previous section is moderated and then suppressed by the (small) kinetic undercooling term, as we now describe. There are three timescales of relevance, namely, the slow down timescale $t = t_c + O\left(\epsilon^2 \ln^3 \ln(1/\epsilon)\right)$, on which the blow-up rate in $\dot{s}$ is mitigated somewhat; the turnaround timescale $t = t_c + O(\epsilon)$, during which the interface speed goes through a minimum and the interface moves so as to release the required amount of latent heat (i.e., the interface moves quickly to the next location x^* at which $u_c(x) = -\lambda$); and, finally, the timescale $t = t_c + (\mu/3\nu)\epsilon \ln(1/\epsilon) + O(\epsilon)$ (where ν is a constant defined in section 5.2), at the end of which the classical Stefan problem is recovered.

5.1. The slow down timescale. This is the shortest of the timescales. As indicated by (4.17) the solution as blow-up is approached is almost of the self-similar form

$$u \sim U\left((x - x_c)/(t_c - t)^{\frac{1}{2}}\right),$$

so we first rescale via

$$t = t_c(\epsilon) + \epsilon^2 T, \qquad x = x_c(\epsilon) + \epsilon X, \qquad s = x_c(\epsilon) + \epsilon S(T)$$

to obtain in the first instance the full balance

$$\frac{\partial u}{\partial T} = \frac{\partial^2 u}{\partial X^2}, \tag{5.1}$$

$$\text{at } X = S(T) \qquad u = \dot{S}(T), \qquad \frac{\partial u}{\partial X} + (1 - \mu)\dot{S}u = -\lambda\dot{S}, \tag{5.2}$$

where $\dot{}$ denotes d/dT. However, as indicated by (4.17) (and indeed by (5.14) below, which implies "almost traveling-wave" behavior in matching as $T \to +\infty$), this

complete elimination of ϵ is in fact spurious; the scales appropriate to describing the transition are in fact

$$T = \delta^{-3}\hat{T}, \qquad S = \delta^{-2}\hat{S}, \qquad X = \delta^{-2}\hat{S} + \delta^{-1}\hat{Z},$$

where $\delta = 1/\ln\ln(1/\epsilon)$ (so that ϵ has to be minute for δ to be genuinely small in practice; fortunately this issue does not arise on the other timescales). The dependence on ϵ is thus remarkably weak.

Hence

$$\delta\frac{\partial u}{\partial \hat{T}} - \frac{d\hat{S}}{d\hat{T}}\frac{\partial u}{\partial \hat{Z}} = \frac{\partial^2 u}{\partial \hat{Z}^2}, \tag{5.3}$$

$$\text{on } \hat{Z} = 0 \qquad u = \delta\frac{d\hat{S}}{d\hat{T}}, \qquad \frac{\partial u}{\partial \hat{Z}} + (1-\mu)\frac{d\hat{S}}{d\hat{T}}u = -\lambda\frac{d\hat{S}}{d\hat{T}}, \tag{5.4}$$

so, writing

$$u \sim \hat{U}_0(\hat{Z},\hat{T}) + \delta\hat{U}_1(\hat{Z},\hat{T}), \qquad \hat{S} \sim \hat{S}_0(\hat{T}) + \delta\hat{S}_1(\hat{T}) \quad \text{as } \delta \to 0,$$

we find that

$$\hat{U}_0 = -\lambda\left(1 - \exp\left(-\frac{d\hat{S}_0}{d\hat{T}}\hat{Z}\right)\right), \tag{5.5}$$

$$\begin{aligned}\hat{U}_1 = \frac{\lambda\frac{d^2\hat{S}_0}{d\hat{T}^2}}{\left(\frac{d\hat{S}_0}{d\hat{T}}\right)^3}&\left(\left(\frac{1}{2}\left(\frac{d\hat{S}_0}{d\hat{T}}\right)^2\hat{Z}^2 + \frac{d\hat{S}_0}{d\hat{T}}\hat{Z} + 1\right)\exp\left(-\frac{d\hat{S}_0}{d\hat{T}}\hat{Z}\right) - 1\right)\\ &+ (1-\mu)\frac{d\hat{S}_0}{d\hat{T}}\exp\left(-\frac{d\hat{S}_0}{d\hat{T}}\hat{Z}\right) + \mu\frac{d\hat{S}_0}{d\hat{T}} - \lambda\frac{d\hat{S}_1}{d\hat{T}}\hat{Z}\exp\left(-\frac{d\hat{S}_0}{d\hat{T}}\hat{Z}\right).\end{aligned} \tag{5.6}$$

Since $d\hat{S}_0/d\hat{T} < 0$, we thus have to these orders that

$$u \sim -\lambda + \delta\left(\mu\frac{d\hat{S}_0}{d\hat{T}} - \frac{\lambda\frac{d^2\hat{S}_0}{d\hat{T}^2}}{\left(\frac{d\hat{S}_0}{d\hat{T}}\right)^3}\right) \qquad \text{as } \hat{Z} \to -\infty$$

as the condition required in matching with (4.18), so that

$$\mu\frac{d\hat{S}_0}{d\hat{T}} - \frac{\lambda\frac{d^2\hat{S}_0}{d\hat{T}^2}}{\left(\frac{d\hat{S}_0}{d\hat{T}}\right)^3} = -\frac{\lambda}{2}.$$

It follows (by suitable choice of $x_c(\epsilon)$ at $O(\epsilon\delta^{-2})$, i.e., up to a translation of $\hat{S}_0$) that we have

$$\mu\hat{S}_0 + \frac{\lambda}{2\left(\frac{d\hat{S}_0}{d\hat{T}}\right)^2} = -\frac{\lambda}{2}\hat{T}. \tag{5.7}$$

Finally, writing

$$\hat{S}_0 = -\frac{\lambda}{2\mu}\hat{T} - W^2 \tag{5.8}$$

yields

$$\frac{2W^2}{\left(\left(\frac{\lambda}{2\mu}\right)^{\frac{1}{2}} - \left(\frac{\lambda}{2\mu}\right)W\right)}\frac{dW}{d\hat{T}} = 1$$

so that, by suitable choice of $t_c(\epsilon)$ at $O(\epsilon^2\delta^{-3})$ (i.e., up to a translation of $\hat{T}$),

$$W^2 + 2\left(\frac{2\mu}{\lambda}\right)^{\frac{1}{2}} W + \frac{4\mu}{\lambda}\ln\left(W - \left(\frac{2\mu}{\lambda}\right)^{\frac{1}{2}}\right) = -\frac{\lambda}{2\mu}\hat{T}. \tag{5.9}$$

It follows from (5.9) that

$$W \sim \left(\frac{-\lambda\hat{T}}{2\mu}\right)^{1/2} - \left(\frac{2\mu}{\lambda}\right)^{1/2} \qquad \text{as } \hat{T} \to -\infty, \tag{5.10}$$

which gives the required behavior

$$\hat{S}_0 \sim 2\left(-\hat{T}\right)^{1/2} \qquad \text{as } \hat{T} \to -\infty \tag{5.11}$$

to match with (4.16) and also

$$W \sim \left(\frac{2\mu}{\lambda}\right)^{1/2} + e^{-\lambda^2\hat{T}/8\mu^2} \qquad \text{as } \hat{T} \to +\infty, \tag{5.12}$$

which gives the behavior

$$\hat{S}_0 \sim -\frac{\lambda}{2\mu}\hat{T} - \frac{2\mu}{\lambda} - 2\left(\frac{2\mu}{\lambda}\right)^{1/2} e^{-\lambda^2\hat{T}/8\mu^2} \qquad \text{as } \hat{T} \to +\infty, \tag{5.13}$$

furnishing a matching condition for the next timescale. The behaviors of $\hat{S}_0$ and $d\hat{S}_0/d\hat{T}$ are shown in Figure 2 using (5.7) and (5.11).

5.2. The turnaround timescale. Setting

$$t = t_c(\epsilon) + \epsilon\hat{t}, \qquad x = s(\hat{t};\epsilon) + \epsilon\hat{z}$$

in the interior layer located at the moving boundary, we have

$$\epsilon\frac{\partial u}{\partial \hat{t}} - \frac{ds}{d\hat{t}}\frac{\partial u}{\partial \hat{z}} = \frac{\partial^2 u}{\partial \hat{z}^2}, \tag{5.14}$$

$$\text{at } \hat{z} = 0 \qquad u = \frac{ds}{d\hat{t}}, \quad \frac{\partial u}{\partial \hat{z}} + (1-\mu)\frac{ds}{d\hat{t}}u = -\lambda\frac{ds}{d\hat{t}}, \tag{5.15}$$

while in the outer region $x = O(1)$

$$\frac{\partial u}{\partial \hat{t}} = \epsilon\frac{\partial^2 u}{\partial x^2},$$

so that

$$u \sim u_c(x), \tag{5.16}$$

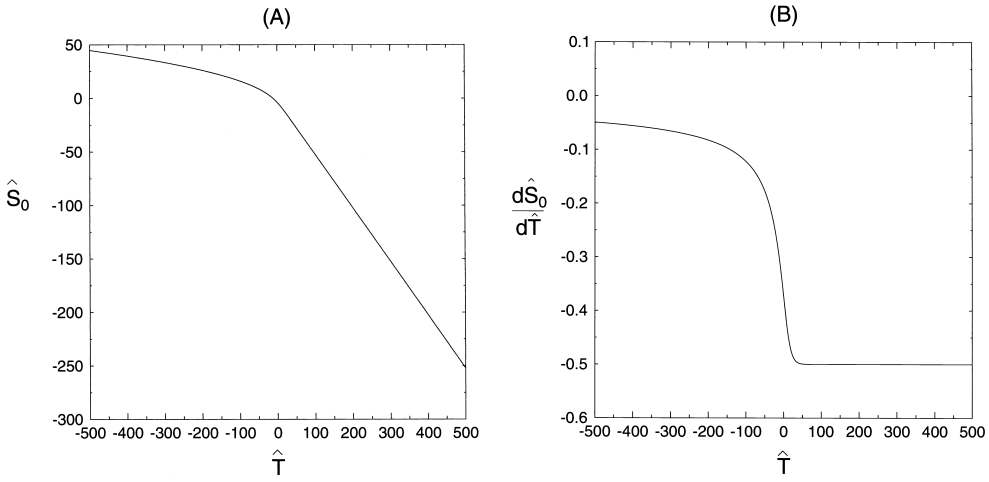


FIG. 2. *Illustration of the leading-order asymptotic solution showing* (A) $\hat{S}_0$ *and* (B) $d\hat{S}_0/d\hat{T}$ *using* (5.7)–(5.9) *with* $\lambda = \mu = 1$. *The far-field matching behaviors are clearly exhibited, particularly in* (B), *where* $d\hat{S}_0/d\hat{T} \sim -(-\hat{T})^{-1/2}$ *as* $\hat{T} \to -\infty$ *and* $d\hat{S}_0/d\hat{T} \sim -\lambda/2\mu$ *as* $\hat{T} \to +\infty$.

where, as before, $u_c(x) = u_0(x, t_c)$, u_0 being the solution to the classical ($\epsilon = 0$) Stefan problem, as discussed above. Expanding in (5.14)–(5.15) in the form

$$u \sim \hat{u}_0(\hat{z}, \hat{t}\,), \qquad s \sim \hat{s}_0(\hat{t}\,) \qquad \text{as } \epsilon \to 0$$

yields

$$\hat{u}_0 = -\lambda \left(1 - e^{-\dot{\hat{s}}_0 \hat{z}}\right) + \dot{\hat{s}}_0 \left(\mu + (1-\mu) e^{-\dot{\hat{s}}_0 \hat{z}}\right), \tag{5.17}$$

where $\dot{}$ now denotes $d/d\hat{t}$. Matching (5.17) as $\hat{z} \to -\infty$ with (5.16) requires that $\hat{s}_0$ be given by

$$\mu \dot{\hat{s}}_0 = u_c(\hat{s}_0) + \lambda \qquad \text{with } \hat{s}_0 = x_c \text{ at } \hat{t} = 0. \tag{5.18}$$

This first-order ordinary differential equation governs the dynamics of the interface motion over the current timescale.

In the one-phase limit, we have from (2.9) that

$$u_2 = \mu u_1(x, \omega(x)) = \epsilon \mu \frac{ds}{dt}\left(\omega(x)\right),$$

which implies that

$$u_2 \sim u_c(x) + \lambda$$

is, to leading order, independent of μ. Thus, as expected (cf. Gurtin [24]), the temperature left behind in the second phase during this rapid transition is given by that in the first phase modified by the latent heat λ in a manner which accounts correctly for the conservation of energy on change of phase.

It remains to discuss the behavior of the ordinary differential equation (5.18). First, it follows from (4.18) that the initial behavior is governed by

$$\mu \dot{\hat{s}}_0 \sim -\frac{\lambda}{2 \ln\left(-\ln(x_c - \hat{s}_0)\right)} \tag{5.19}$$

and hence

$$\hat{s}_0 \sim x_c - \frac{\lambda}{2\mu}\frac{\hat{t}}{\ln\left(-\ln\hat{t}\right)} \qquad \text{as } \hat{t} \to 0^+. \tag{5.20}$$

Since $\hat{t} = \epsilon\delta^{-3}\hat{T}$, this gives the matching condition

$$\hat{s}_0 \sim x_c - \frac{\lambda}{2\mu}\epsilon\delta^{-2}\hat{T},$$

consistent with (5.13).

Elsewhere, $u_c(x)$ depends on the initial data. The key observation regarding (5.18) is that the interface slows down as it approaches the point $x^* < x_c$ at which $u_c(x^*) = -\lambda$; we generically have

$$u_c \sim -\lambda - \nu(x - x^*) \tag{5.21}$$

as $x \to x^*$ for some positive constant ν and hence

$$\hat{s}_0 \sim x^* + e^{-\nu(\hat{t}-\hat{t}_c)/\mu} \qquad \text{as } \hat{t} \to +\infty \tag{5.22}$$

for some constant $\hat{t}_c$.

It is worth remarking that the nongeneric behavior

$$u_c \sim -\lambda - \nu(x - x^*)^{2m+1} \tag{5.23}$$

as $x \to x^*$ for some positive constant ν and integer power $m > 0$ gives instead the algebraic decay behavior

$$\hat{s}_0 \sim x^* + \left(\frac{\mu}{2m\nu\hat{t}}\right)^{1/2m} \qquad \text{as } \hat{t} \to +\infty. \tag{5.24}$$

5.3. The timescale of recovery of the classical Stefan problem. The relevant timescale corresponds to $s = x_c + O(\epsilon^{1/3})$, so by (5.22) the scalings

$$\hat{t} = \frac{\mu}{3\nu}\ln(1/\epsilon) + \hat{t}_c + \bar{T}, \qquad s(\bar{T};\epsilon) = x^* + \epsilon^{1/3}\bar{S}, \qquad x = s(\bar{T};\epsilon) + \epsilon^{2/3}\frac{Z}{\dot{\bar{S}}}$$

pertain, where $\dot{}$ is now $d/d\bar{T}$. Thus in the interior layer $Z = O(1)$

$$\frac{\epsilon^{1/3}}{\dot{\bar{S}}^2}\frac{\partial u}{\partial \bar{T}} + \epsilon^{1/3}\frac{\ddot{\bar{S}}Z}{\dot{\bar{S}}^3}\frac{\partial u}{\partial Z} - \frac{\partial u}{\partial Z} = \frac{\partial^2 u}{\partial Z^2} \tag{5.25}$$

$$\text{at } Z = 0 \qquad u = \epsilon^{1/3}\dot{\bar{S}}, \quad \frac{\partial u}{\partial Z} + (1-\mu)u = -\lambda. \tag{5.26}$$

Expanding in the form

$$u \sim U_0(Z) + \epsilon^{1/3}U_1(Z,\bar{T}), \qquad \bar{S} \sim \bar{S}_0(\bar{T}) \qquad \text{as } \epsilon \to 0$$

yields

$$U_0 = -\lambda\left(1 - e^{-Z}\right)$$

and hence

$$-\lambda\frac{\ddot{\bar{S}}_0 Z}{\dot{\bar{S}}_0^3}e^{-Z}-\frac{\partial U_1}{\partial Z}=\frac{\partial^2 U_1}{\partial Z^2}, \tag{5.27}$$

$$\text{at } Z=0 \qquad U_1=\dot{\bar{S}}_0, \quad \frac{\partial U_1}{\partial Z}=-(1-\mu)\dot{\bar{S}}_0, \tag{5.28}$$

so that

$$U_1=\frac{\lambda\ddot{\bar{S}}_0}{\dot{\bar{S}}_0^3}\left(\left(\frac{1}{2}Z^2+Z+1\right)e^{-Z}-1\right)+(1-\mu)\dot{\bar{S}}_0 e^{-Z}+\mu\dot{\bar{S}}_0.$$

Since (5.16) still applies in $x<s$, matching as $Z\to+\infty$ (noting that $\dot{\bar{S}}<0$) with (5.22) requires that $\bar{S}_0$ be given by the ordinary differential equation

$$\mu\dot{\bar{S}}_0-\lambda\frac{\ddot{\bar{S}}_0}{\dot{\bar{S}}_0^3}=-\nu\bar{S}_0, \tag{5.29}$$

with

$$\bar{S}_0\sim e^{-\nu\bar{T}/\mu}-\frac{\lambda\mu}{3\nu^2}e^{2\nu\bar{T}/\mu} \qquad \text{as } \bar{T}\to-\infty. \tag{5.30}$$

As $\bar{T}\to+\infty$, the first term in (5.29) becomes negligible and we obtain

$$\bar{S}_0\sim-\left(\frac{6\lambda}{\nu}\right)^{1/3}\bar{T}^{1/3} \quad \text{as } \bar{T}\to+\infty; \tag{5.31}$$

the formulation (5.29) indicates how the interface s passes through x^*. A numerical solution of (5.29) as an initial value problem is shown in Figure 3, using the two-term behavior (5.30) at $\bar{T}=-\bar{T}_{\text{init}}$ (for suitably large $\bar{T}_{\text{init}}$) to start the scheme.

Finally, from (5.31) we can deduce that for $t=t_{c0}+O(1)$ we recover at leading order the classical Stefan problem

$$\begin{aligned}
&\text{in } 0<x<s_0(t) && \frac{\partial u_0}{\partial t}=\frac{\partial^2 u_0}{\partial x^2},\\
&\text{on } x=0 && \frac{\partial u_0}{\partial x}=0,\\
&\text{on } x=s_0(t) && u_0=0, \quad \frac{\partial u_0}{\partial x}=-\lambda\frac{ds_0}{dt},\\
&\text{at } t=t_{c0} && u_0=u_c(x), \quad s_0=x^*,
\end{aligned}$$

whose small $t-t_{c0}$ behavior can be shown (essentially as above) to take the form

$$s_0(t)\sim x^*-\left(\frac{6\lambda}{\nu}\right)^{1/3}(t-t_{c0})^{1/3} \qquad \text{as } t\to t_{c0}^+,$$

thereby matching with (5.31).

The discussion so far has assumed that the blow-up profile has only two roots for $u_c(x)+\lambda=0$ with $0\le x\le x_c$. The case in which there are more than two roots

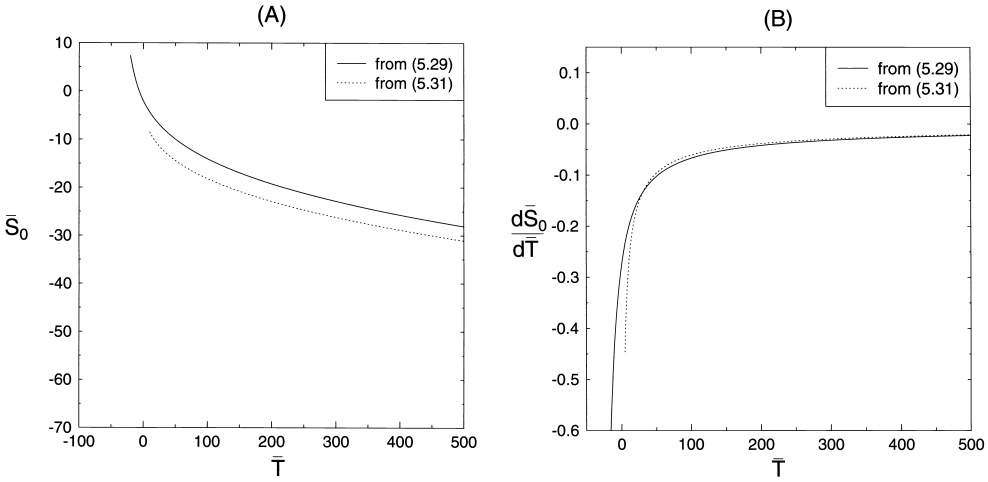


FIG. 3. (A) *and* (B) *show* $\bar{S}_0$ *and* $\dot{\bar{S}}_0$*, respectively, against* $\bar{T}$ *calculated from* (5.29)*. The values* $\nu = 0.1, \lambda = \mu = 1$ *were taken for illustrative purposes, and the scheme started from* $\bar{T} = -20$*. The asymptotic behavior* (5.31) *is also shown in both* (A) *and* (B)*.*

gives rise to the possibility of repeated blow-up in the unregularized problem. Such scenarios depend upon the blow-up profile satisfying

$$Q - Q_c = \int_0^{x^*} (u_c(x) + \lambda) dx < 0.$$

The case in which $u_c < -\lambda$ for all x is remarked upon in the next subsection.

5.4. Final behavior. In this section we address the possible ultimate outcomes of (3.1)–(3.4) for arbitrary ϵ. One possibility for any μ is that $s \to s_\infty$ as $t \to \infty$ for some positive constant s_∞ (which will depend on the initial data), with

$$\text{on } x = s_\infty \qquad \epsilon \frac{\partial u}{\partial x} \sim -\lambda u$$

giving the asymptotic boundary condition, implying that

$$u \sim -U_\infty e^{-\kappa^2 t} \cos(\kappa x) \qquad \text{as } t \to \infty \tag{5.32}$$

for some positive constant U_∞, where κ is (generically) the smallest root of

$$\epsilon \kappa \tan \kappa s_\infty = \lambda. \tag{5.33}$$

However, for $\mu < 1$ it is also possible that s drops to zero in finite time, a scenario which can be identified by assuming that the left-hand side of (3.1) is asymptotically negligible (the self-consistency of this assumption being readily checked a posteriori), implying that

$$u \sim U(t), \qquad \frac{\partial u}{\partial x} \sim \dot{U}(t) x$$

for some function $U(t)$. The moving boundary conditions then yield

$$U \sim \epsilon \dot{s}, \qquad \dot{U} s + \dot{s}(1 - \mu) U \sim -\lambda \dot{s},$$

so for $\mu \neq 1$ we obtain

$$\dot{s} \sim A s^{\mu-1} - \frac{\lambda}{\epsilon(1-\mu)} \tag{5.34}$$

for some constant A. If $A > 0$ and $\mu < 1$, then s tends to $s_\infty \sim (\lambda/\epsilon(1-\mu)A)^{1/(\mu-1)}$ as $t \to \infty$, and the behavior of (5.34) in this case can easily be seen to be consistent with the small s_∞ limit of (5.32)–(5.33). If $\mu > 1$, then A is necessarily less than zero (since $\dot{s} < 0$ must hold for $u < 0$) and the same conclusion applies (the equivalent result can easily be seen always to follow when $\mu = 1$ as well). However, for $\mu < 1$ with $A < 0$ we have that

$$s \sim ((2-\mu)(-A)(t_e - t))^{\frac{1}{2-\mu}}, \qquad U \sim \epsilon(-A)\left((2-\mu)(-A)(t_e-t)\right)^{-\frac{1-\mu}{2-\mu}} \qquad \text{as } t \to t_e^- \tag{5.35}$$

for some finite extinction time t_e which will depend on the initial data, with u becoming unbounded in this limit. Thus, interestingly, there are two possible generic outcomes for $\mu < 1$: first, $s \to s_\infty > 0$ with (5.32), and second, (5.35), the borderline between the two being given by the nongeneric case $A = 0$ in (5.34), whereby

$$s \sim \frac{\lambda}{\epsilon(1-\mu)}(t_e - t), \qquad U \sim -\frac{\lambda}{1-\mu}. \tag{5.36}$$

For small ϵ, when $u_c(\hat{s}_0) < -\lambda$ for all $\hat{s}_0$ the result (5.18) does not in general hold uniformly as $\hat{s}_0$ approaches zero. We shall not address the corresponding small ϵ asymptotics here, but it is plausible that for $\mu < 1$ the scenario (5.35) will typically ultimately apply, whereas for $\mu \geq 1$ we necessarily have $s_\infty > 0$, so the interface does not impinge upon the boundary, though as ϵ tends to zero s_∞ presumably also does so.

6. The special case $\mu = 0$. We stress that the parameter μ plays no role in the zero kinetic undercooling case, so it does not feature as a special case in the unregularized problem. It does, however, require separate treatment for $\epsilon > 0$ since, as already noted, even with kinetic undercooling bona fide blow-up can occur for $\mu = 0$. Here we treat the behavior close to the blow-up time $t = t_c$ for arbitrary $\epsilon > 0$. Two reformulations of (3.1)–(3.4) are helpful in this case (it is no coincidence that these reformulations are available only in the special case $\mu = 0$, which is also exceptional with regard to its blow-up properties); we first recall from (3.6) that

$$\int_0^{s(t)} (u(x,t) + \lambda)\,dx = Q \tag{6.1}$$

holds for all time when $\mu = 0$. Again introducing w as in (4.1), we obtain (4.2) subject to

$$\text{on } x = s(t) \qquad w = 0, \quad \frac{\partial w}{\partial x} = -\epsilon, \tag{6.2}$$

$$\text{on } x = 0 \qquad \frac{\partial w}{\partial x} = -Q - \epsilon, \tag{6.3}$$

$$\text{at } t = 0 \qquad s = 1, \quad w = \epsilon(1-x) + \int_x^1 (x' - x)(u_0(x') + \lambda)\,dx', \tag{6.4}$$

the conditions (6.3) and (6.4) following from (6.2) and

$$\text{at } t = 0 \qquad \frac{\partial^2 w}{\partial x^2} = u_0(x) + \lambda.$$

Second, introducing

$$v = -\int_x^{s(t)} (u(x',t) + \lambda)dx', \tag{6.5}$$

so that

$$v = \epsilon + \frac{\partial w}{\partial x}, \qquad u = \frac{\partial w}{\partial t} = \frac{\partial v}{\partial x} - \lambda, \tag{6.6}$$

leads to

$$\frac{\partial v}{\partial t} = \frac{\partial^2 v}{\partial x^2}, \tag{6.7}$$

$$\text{on } x = s(t) \qquad v = 0, \quad \frac{\partial v}{\partial x} = \lambda + \epsilon \dot{s}(t), \tag{6.8}$$

$$\text{on } x = 0 \qquad v = -Q, \tag{6.9}$$

$$\text{at } t = 0 \qquad s = 1, \quad v = -\int_x^1 (u_0(x') + \lambda)dx'. \tag{6.10}$$

If the λ terms are negligible, then (6.7)–(6.10) reduces to the classical Stefan problem with Stefan number ϵ. The boundary condition we have adopted on $x = 0$, (3.3), differs in form from (6.9), but provided $x_c > 0$ the analysis of section 4 carries over because it is local to $x = x_c$; thus identifying u with $-v$ and ϵ with λ, it follows from (4.17) that

$$v \sim \epsilon + A(\tau) \quad \text{for } y = O(1), \qquad v \sim \epsilon(1 - e^{z/2}) \quad \text{for } z = O(1), \tag{6.11}$$

immediately prior to blow-up, where (4.6) remains valid, as does $y = L + z/L$. The expressions (4.16) become

$$L \sim 2\,(\ln \tau)^{1/2}, \qquad A \sim \epsilon/2 \ln \tau \qquad \text{as } \tau \to \infty$$

in the current notation; since

$$\dot{s} \sim -\frac{1}{2} L(\tau)(t_c - t)^{-1/2} \qquad \text{as } t \to t_c^-$$

the λ term in (6.8) is indeed negligible, as required for self-consistency. From (6.6) and (4.18) (which now gives the local behavior of v) we have

$$u_c(x) \sim -\frac{\epsilon}{2\ln^2(-\ln(x_c - x))\,(-\ln(x_c - x))\,(x_c - x)} \quad \text{as } x \to x_c^-, \tag{6.12}$$

so (unlike (4.18)) u becomes unbounded as the blow-up time is approached.

For $u < 0$, we have from (6.6) that v is monotonic decreasing with x, and hence (6.11) is consistent with (6.9) only if $-Q > \epsilon$ (that this inequality should arise is perhaps not surprising in the light of (6.3)). A second, distinct, blow-up scenario thus necessarily pertains when this inequality is not satisfied; this has $x_c = 0$, so the form of the boundary condition on $x = 0$ is important in this case, and this form of blow-up can most conveniently be derived directly from the original formulation (3.1)–(3.4), with $\mu = 0$ and with the λ term again negligible in the limit $t \to t_c^-$. We thus seek self-similar behavior of the form

$$u \sim \frac{1}{(t_c - t)^{1/2}} f(\eta), \quad \eta = \frac{x}{(t_c - t)^{1/2}}, \quad s(t) \sim \sigma(t_c - t)^{1/2} \quad \text{as } t \to t_c^-, \tag{6.13}$$

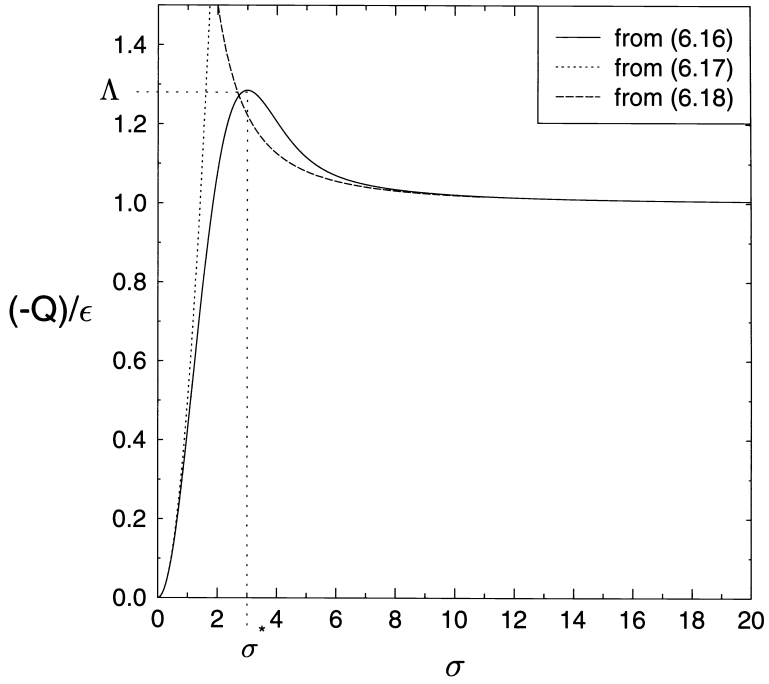


FIG. 4. *Plot of the solutions to the transcendental equation* (6.14) *together with the asymptotic behaviors* (6.16) *and* (6.17). *We have that* $\sigma^* \approx 3.0045$ *and* $\Lambda \approx 1.28475$.

implying

$$f = -Be^{\eta^2/4}$$

for some constant B, with σ given by the transcendental equation

$$\frac{1}{2}\sigma e^{-\sigma^2/4}\int_0^\sigma e^{\eta^2/4}d\eta = \frac{(-Q)}{\epsilon} \tag{6.14}$$

and with

$$B = \frac{\epsilon}{2}\sigma e^{-\sigma^2/4}; \tag{6.15}$$

these follow from (3.2) and (6.1). Thus as $(-Q)/\epsilon \to 0^+$ we have

$$\sigma \sim (2(-Q)/\epsilon)^{1/2}, \qquad B \sim ((-Q)\epsilon/2)^{1/2}, \tag{6.16}$$

and as $(-Q)/\epsilon \to 1^+$,

$$\sigma \sim (2/\left((-Q)/\epsilon - 1\right))^{1/2} \tag{6.17}$$

with B exponentially small.

We plot in Figure 4 the solutions to (6.14), the nonmonotonicity following from (6.16)–(6.17). There is thus a range of $(-Q)/\epsilon > 1$ (which we denote $1 < (-Q)/\epsilon < \Lambda$, as in Figure 4) for which two solutions are possible, and we speculate that those with

$\sigma \leq \sigma^*$ are stable (where $\sigma = \sigma^*$ corresponds to the maximum value of $(-Q)/\epsilon$ from (6.14)) and those beyond the fold, in $\sigma > \sigma^*$, are unstable. In the range of $1 < (-Q)/\epsilon < \Lambda$ there are thus two generic (stable) blow-up scenarios (one with $x_c = 0, \sigma < \sigma^*$ and another with $x_c > 0$ depending on the initial data, for which (6.12) applies), separated by a nongeneric (unstable) scenario of the form (6.13) with $\sigma > \sigma^*$. Finally, in the borderline case $(-Q)/\epsilon = 1$ it is in principle possible that a form of blow-up which would be nongeneric for arbitrary parameter values may occur generically, so it is worthwhile to confirm that the situation described at the end of section 4 cannot pertain with $x_c = 0$; thus, while in the current notation (4.21) becomes

$$v \sim \epsilon + Ax,$$

which satisfies (6.9), it also follows that v is increasing with x, contrary to the assumption that $u < 0$. Thus we conjecture that (6.13) provides the only generic form of blow-up for $(-Q)/\epsilon \leq 1$ and (6.17) for $(-Q)/\epsilon > 1$.

For completeness we also note that when $Q > 0$ (so that blow-up does not occur even without kinetic undercooling) we have $s \sim Q/\lambda$ as $t \to \infty$, with u decaying exponentially with t, as dictated by the relevant linear diffusion problem subject to

$$\text{on } x = Q/\lambda \qquad \epsilon \frac{\partial u}{\partial x} \sim -\lambda u.$$

Finally, for $Q = 0$ we have

$$u \sim -\lambda, \qquad s \sim \frac{\lambda}{\epsilon}(t_c - t) \qquad \text{as } t \to t_c^-,$$

which in fact furnishes the exact solution to (3.1)–(3.4) for $\mu = 0, u_0 \equiv -\lambda$. This exact solution generalizes to arbitrary $\mu < 1$ to

$$u = -\frac{\lambda}{1-\mu}, \qquad s = \frac{\lambda}{\epsilon(1-\mu)}(t_c - t), \tag{6.18}$$

a solution which is consistent with (5.18) and which has played a role above in (5.36); moreover, (6.18) generalizes further to

$$u = -\lambda - \epsilon\mu q + (\lambda - \epsilon(1-\mu)q)e^{qz}, \qquad s = q(t_c - t)$$

for arbitrary wavespeed q (cf. (5.17)).

For small ϵ, the relevant regime is of course that leading to (6.12). The transition to this from the blow-up behavior described in section 4 is rather complex, and we shall not give anything like full details; the initial stages can be described by naively incorporating the ϵ term in the boundary condition in $x = s$ into the interior layer analysis of section 4, in which case the ordinary differential equation (4.15) generalizes to

$$\frac{16\sqrt{\pi}}{L^4} e^{L^2/4} \dot{L} \sim 1 - \frac{\epsilon L}{2\lambda} e^{\tau/2},$$

which itself has a rather delicate asymptotic structure in the limit $\epsilon \to 0$ over the timescales on which the second term on the right-hand side comes to dominate the first. Subsequently, there is a phase in which the λ terms are asymptotically negligible (i.e., the evolution is governed by the zero latent heat problem), from which the blow-up behavior (6.12) ultimately emerges.

7. Numerical results. The method of lines may be used to solve the problem (3.1)–(3.4) numerically. The method adopted is described in detail in the appendix and follows Fasano, Meyer, and Primicerio [19], with the modification discussed in Meyer [28] used to accommodate the Neumann condition on the fixed boundary. The scheme may be used for both cases $\epsilon = 0$ and $\epsilon > 0$, with the minor modifications needed for the former case also described in the appendix.

For the numerical simulations, the initial profile adopted was the following piecewise linear function:

$$u_{in}(x) = \begin{cases} -a, & x < \hat{\alpha}, \\ -a + a\dfrac{(x-\hat{\alpha})(x-\hat{\beta})}{(1-\hat{\alpha})(1-\hat{\beta})}, & \hat{\alpha} < x \leq 1, \end{cases} \tag{7.1}$$

where $\hat{\alpha} = 0.2$ and $\hat{\beta} = 0.9$ were fixed in all of the following calculations.

Figure 5 shows the full numerical solution for the interface position and speed in the case with no kinetic undercooling, $\epsilon = 0$. Taking $a = 0.4, 0.5, 0.6, 0.8$ gives $Q = 0.33, 0.167, 1.3 \times 10^{-7}, -0.33$, respectively, from (3.5). The increase in the interface speed as Q decreases is clearly illustrated. Blow-up would be expected to occur in the case $Q = -0.33$. Not surprisingly, such behavior is not fully reproduced in the numerics, with truncation errors presumably serving to regularize the problem.

Figures 6 and 7 show the time development of the full numerical solution for the interface position and speed as the two key parameters, the segregation coefficient μ and the kinetic undercooling parameter ϵ, vary independently; the initial profile (7.1) with $a = 0.8$, corresponding to $Q = -0.33$, was adopted. Figure 6 shows the approach to blow-up behavior as $\mu \to 0^+$ for $\epsilon = 1$, while Figure 7 illustrates the corresponding behavior as $\epsilon \to 0^+$ for $\mu = 1$. As already noted, kinetic undercooling does not prevent blow-up when $\mu = 0$, with Figure 6 illustrating that this limit is approached relatively quickly as μ is decreased (the numerical scheme being particularly sensitive to the value of μ, as can be seen from the scalar equation (A.9) for determination of the position of the interface at the next time level). For comparison in Figure 7, the asymptotic blow-up behavior (4.16) of the unregularized problem

$$\dot{s} \sim -\left(\frac{1}{\ln(t_c - t)\sqrt{\ln(-\ln(t_c - t))}}\right)(t_c - t)^{-1/2} \quad \text{as } t \to t_c^- \tag{7.2}$$

is also plotted in Figure 7(B) (the estimate $t_c = 0.0096$ being used for the blow-up time, this being determined from the full numerical solution in the $\epsilon = 0$ case with the same initial profile). The value $\epsilon = 10^{-4}$ represents the smallest value accessible numerically with reasonable accuracy.

Figures 8 and 9 compare the full numerical solution (shown in both as (A) and (B)) with the asymptotics (shown in both as (C) and (D)) of sections 4 and 5 for the specific case $\epsilon = 10^{-3}, \mu = 1, Q = -0.33$. The same initial profile was used and the time variable $-\ln|t - t_c|$ was chosen to illustrate the timescales involved. The asymptotics require estimates of certain values from the blow-up profile which were obtained from the full numerical scheme in the $\epsilon = 0$ case with the same initial profile. This yielded the approximate values $x_c = 0.56, x^* = 0.15, t_c = 0.0096, \nu = 1.75$, with

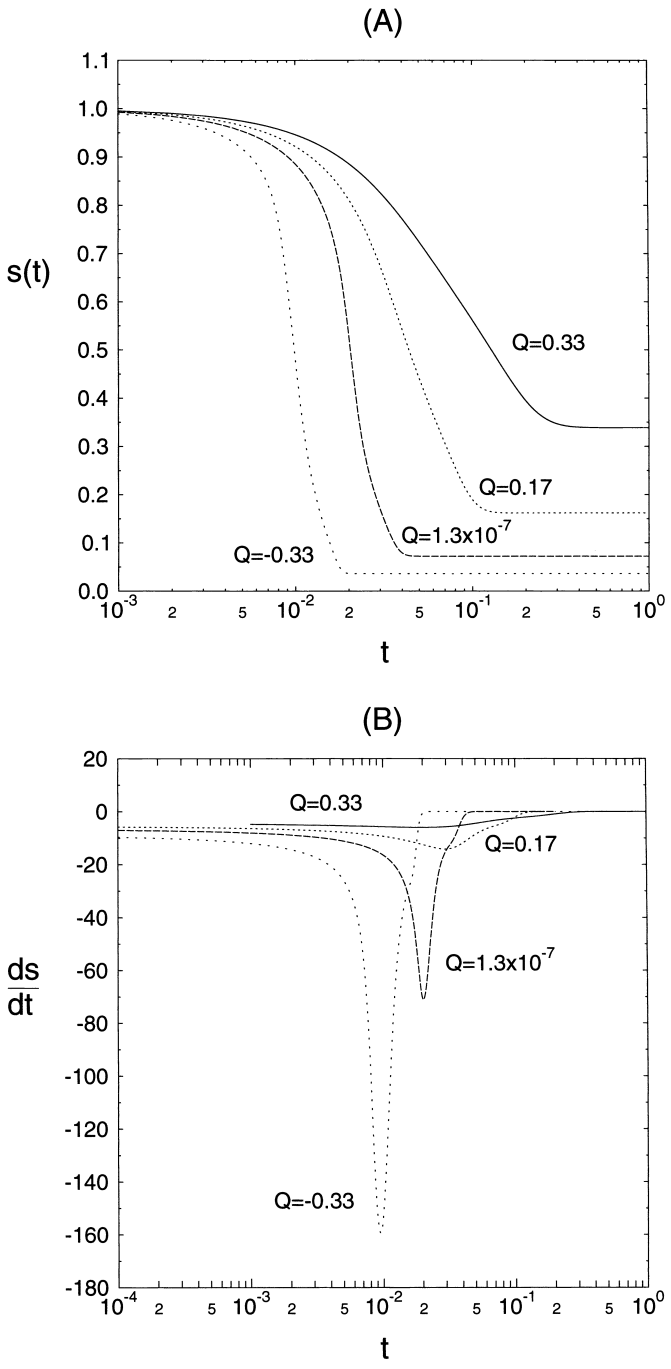


FIG. 5. *Results from the method of lines numerical scheme on* (3.1)–(3.4) *in the unregularized case* $\epsilon = 0$. *The initial profile* (7.1) *was taken with the parameter* a *varied to give the values of initial supercooling parameter* Q *shown. The interface location* s *is shown in* (A) *and its speed* $\dot{s}$ *in* (B).

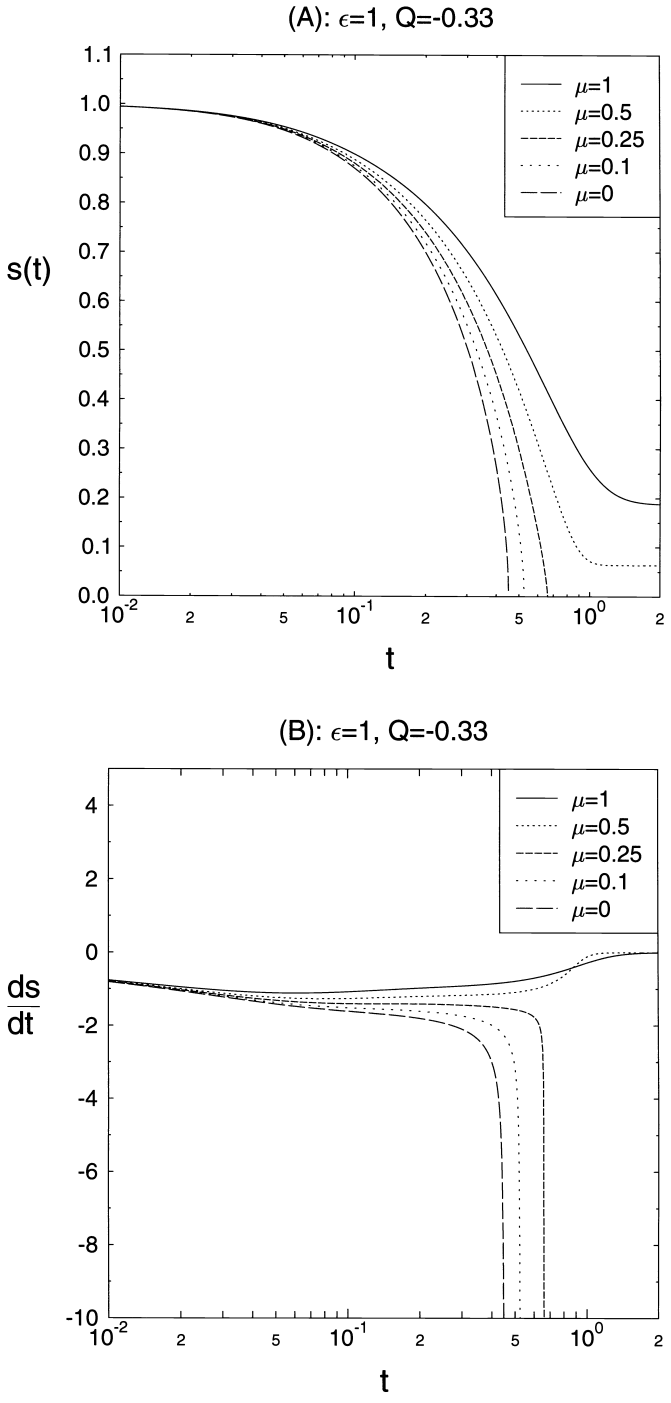


FIG. 6. *Results from the method of lines numerical scheme for* (3.1)–(3.4) *in the regularized case* $\epsilon = 1, Q = -0.33$ *for selected values of the segregation coefficient* μ*;* (A) *shows* s *and* (B) $\dot{s}$*. The initial profile* (7.1) *was taken with the parameter* $a = 0.8$*. Our analysis implies bona fide blow-up occurs for* $\mu = 0$*, and we observe that for* $\mu = 0.1$ *and* $\mu = 0.25$ *also the regularizing term seems insufficient to prevent the numerical scheme from breaking down in a manner akin to such blow-up.*

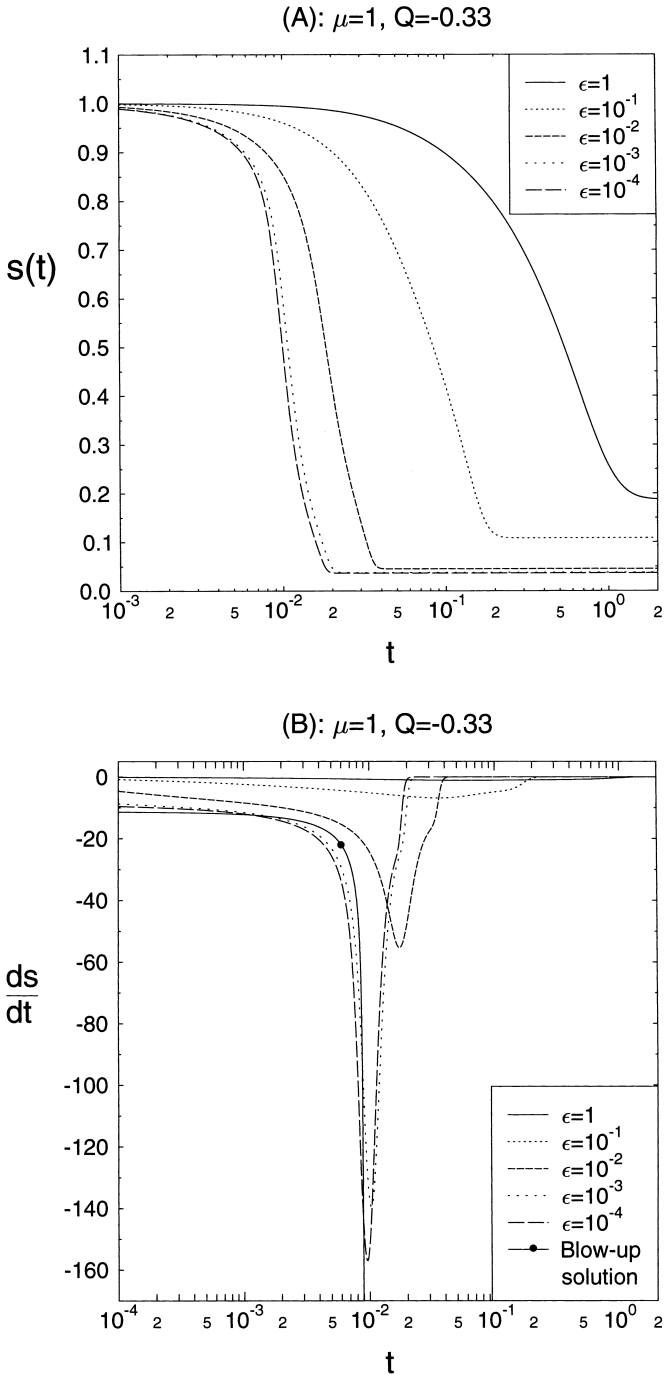


FIG. 7. *Results from the method of lines numerical scheme for* (3.1)–(3.4) *in the regularized case with* $\mu = 1, Q = -0.33$ *for selected values of the kinetic parameter* ϵ*;* (A) *shows* s *and* (B) $\dot{s}$*. The initial profile* (7.1) *was taken with the parameter* $a = 0.8$ *fixed. Also shown for comparison in* (B) *is the asymptotic blow-up behavior* (7.2).

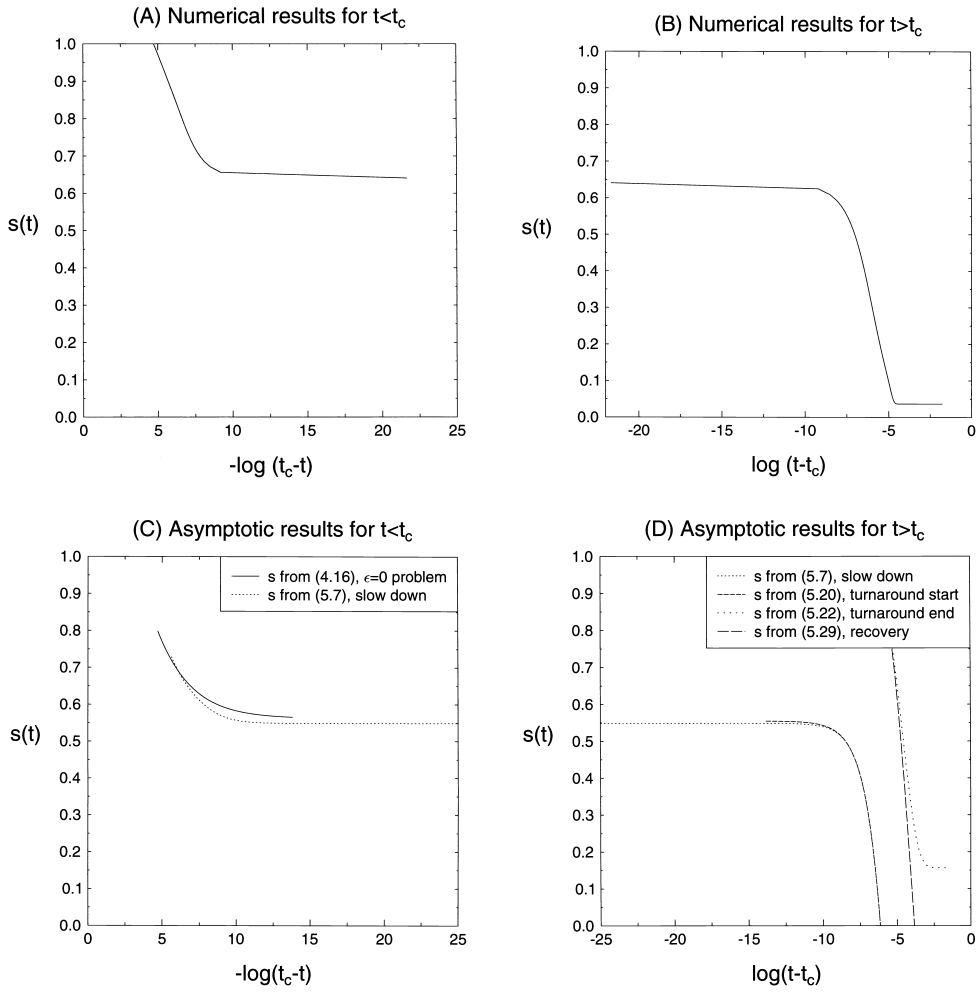


FIG. 8. *Comparison of numerical and asymptotic solutions for the interface* $s(t)$. *The parameter values taken were* $\epsilon = 10^{-3}, \mu = 1$. (A) *and* (B) *show* $s(t)$ *from the method of lines numerical scheme for* (3.1)–(3.4) *with* $a = 0.8$ *in the initial profile* (7.1), *so that* $Q = -0.33$. *The value* $t_c = 0.0096$ *was used in* (A) *and* (B) *for the transformed timescale, which was estimated using the numerical scheme on the same initial profile but with the parameter value* $\epsilon = 0$ *(this scheme also gave the estimates* $x_c = 0.56, x^* = 0.15$*). The upper curve gives* $s(t)$ *for* $t < t_c$ *and the lower gives* $s(t)$ *for* $t > t_c$. (C) *and* (D) *illustrate the asymptotic approximations derived for* $s(t)$ *in section* 5 *on the three main timescales. The parameter values for* ϵ *and* μ *are used to express these asymptotic approximations for* s *in terms of* $t - t_c$. *Since the asymptotics are for times relative to* t_c, *an approximation for* t_c *is not necessary, but the estimates* $x_c = 0.56, x^* = 0.15$ *from the numerical scheme with* $\epsilon = 0$ *were used. For comparison in* (C) *the blow-up asymptotic solution* (4.16) *is also shown. In* (D), *only the initial and final behaviors* (5.20) *and* (5.22) *on the turnaround timescale are shown since behavior over the whole of this timescale is controlled by* (5.18), *which requires the full profile* $u_c(x)$. *The interval over which the interface effectively "jumps" is represented in* (D) *by the vertical distance between horizontal asymptote at* $s \approx 0.56$ *and the end of the timescale that leads to the recovery of the classical Stefan problem at* $s \approx 0.15$.

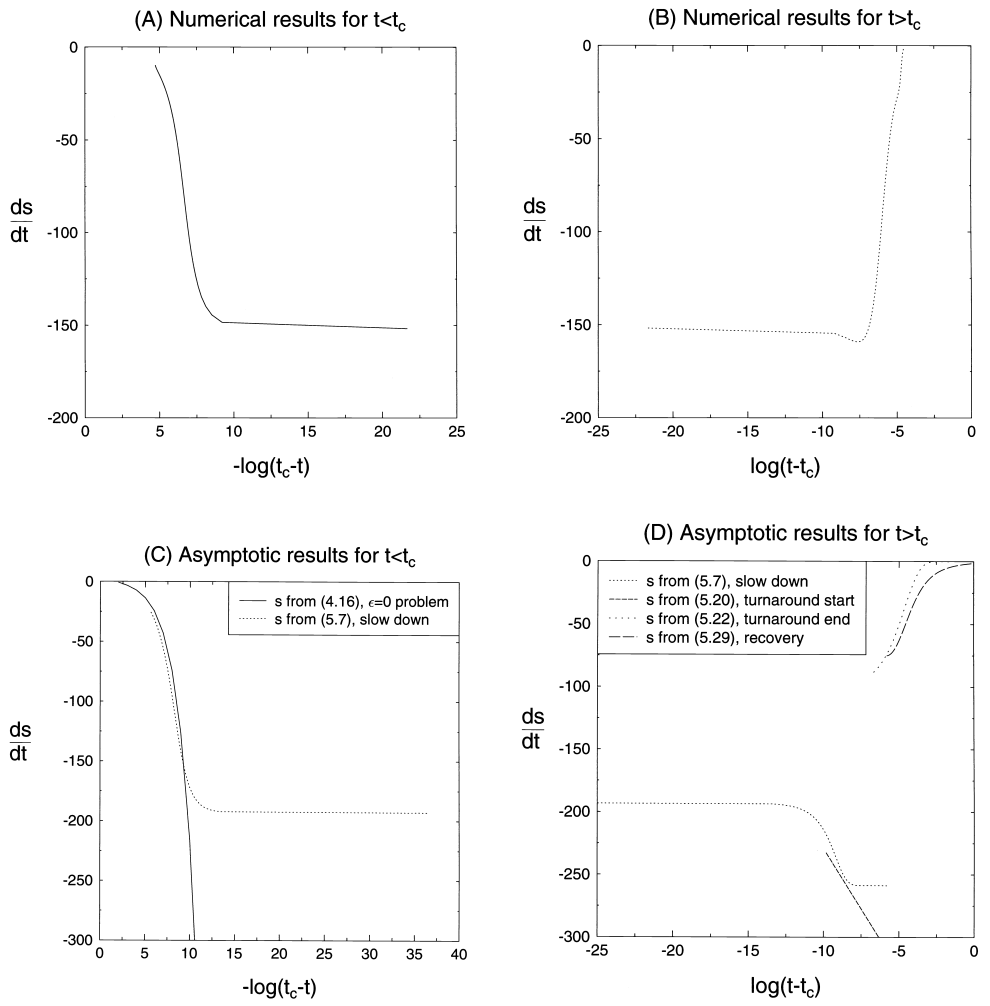


FIG. 9. *Comparison of numerical and asymptotic solutions for the interface speed ds/dt in the example described in Figure* 8. *Results analogous to those shown for s in Figure* 8 *are given here for ds/dt. Results from the method of lines numerical scheme are shown in* (A) *and* (B), *while the asymptotic approximations derived in section* 5 *are shown in* (C) *and* (D). *We note in* (D) *a gap in the asymptotic solution for the turnaround timescale where only the start and end behaviors are shown. The intermediate behavior requires the blow-up profile as dictated by* (5.18).

Figure 10 showing the initial profile $u_0(x)$ and an approximation to the blow-up profile $u_c(x)$. These values are very sensitive to the approximation to the blow-up profile chosen and are recorded here for illustrative purposes only. Even though $\epsilon = 10^{-3}$ is relatively large (as far as the asymptotics are concerned, since $\delta \approx 0.52$ is still not particularly small), only qualitative comparisons can be made. Although the initial behavior on the turnaround timescale of section 5.2 cannot really be observed for the ϵ chosen, the emergence of the predicted timescales is supported. On the turnaround timescale, the asymptotics suggest that ds/dt will decrease according to (5.13) before subsequently increasing to eventually give (5.22), with profiles of the form shown in Figure 1 indicating from (5.18) the existence of a local minimum for ds/dt.

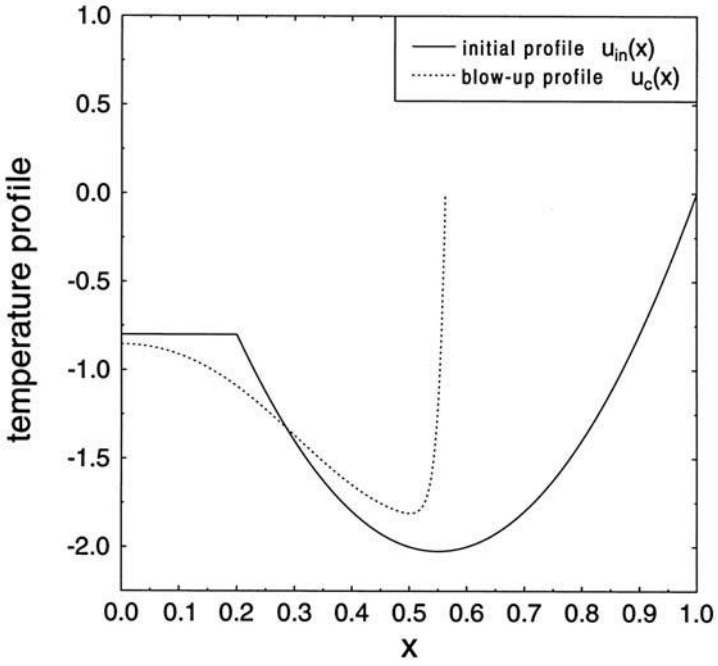


FIG. 10. *Illustration of the initial and blow-up profiles from the method of lines numerical scheme in the case* $\epsilon = 0, Q = -0.33$, *where* $a = 0.8$ *in the initial profile* (7.1). *The blow-up profile is of course approximate and has the value* $Q = -0.3$, *illustrating that approximately* 10% *of* Q *has been lost by the numerical scheme. This approximation to the blow-up profile was used to give the estimates* $x_c = 0.56, x^* = 0.15, \nu = 1.75$ *with* $t_c = 0.0096$. *These values are used in both the numerical scheme and asymptotic approximations shown in Figures* 8 *and* 9.

8. Discussion. The regularization of the supercooled Stefan problem by kinetic undercooling in the one-dimensional, one-phase case has been discussed both asymptotically and numerically. The derivation of the one-phase model from a two-phase formulation highlights the necessity to distinguish between the cases in which the interface advances and retreats. The asymptotics of the main transition timescales for regularization by kinetic undercooling support the proposal of Gurtin [24] that an appropriate $\epsilon = 0$ formulation which allows continuation through blow-up involves an instantaneous jump in the interface location (to a position which the $\epsilon \to 0^+$ asymptotics uniquely identifies, clarifying an issue raised in [24]) at the blow-up time. This abrupt jump becomes a smooth but rapid one in the regularized problem with $0 < \epsilon \ll 1$. For $\mu = 0$, the regularization becomes ineffective in preventing singularity formation (as follows from the conserved quantity (6.1)), and some intriguing blow-up behavior arises.

One way of characterizing concisely the asymptotic results is to note the power laws and so forth which describe the various intermediate asymptotic regimes. Here we note, in particular, the $(-\hat{T})^{1/2}$ behavior in (5.11), the near-linear nature of (5.20), the exponential decay in (5.22), and the $\bar{T}^{1/3}$ rate of decrease in (5.31). Other noteworthy features include the cases in which two distinct generic behaviors occur (so that which is realized depends on the initial data); these arise in section 5.4 for $0 < \mu < 1$ (related behavior for $\mu = 0$ is described at the end of section 6, but the conservation law (6.1) which pertains in this case means that the dependence on the initial conditions can be characterized explicitly through the value of $Q \geq -\epsilon$) and in section 6 for $\mu = 0$ with

$1 < (-Q)/\epsilon \leq \Lambda$, wherein the existence of multiple similarity solutions is particularly striking.

Obvious extensions of this work are to consider the analogous two-phase and higher-dimensional problems.

Appendix. Method of lines numerical scheme. The continuous time problem is discretized and solved at successive time levels as a sequence of free boundary problems for the ordinary differential equations which arise in this way.

At time level $t = t_n$ with $t_n - t_{n-1} = \Delta t$ the solution $\{u_n(x), s_n\}$ is computed as the solution of the discretized equations

$$\text{in } 0 < x < s_n \qquad u_n'' - \frac{1}{\Delta t}(u_n - u_{n-1}) = 0, \tag{A.1}$$

$$u_n'(0) = 0, \tag{A.2}$$

$$\frac{s_n - s_{n-1}}{\Delta t} = f(u_n(s_n)), \qquad s_0 = s(0) = 1, \tag{A.3}$$

$$u_n'(s_n) = -\frac{s_n - s_{n-1}}{\Delta t}\left(\lambda + (1-\mu)u_n(s_n)\right). \tag{A.4}$$

Here, we have allowed for a general kinetic condition of the form $\dot{s} = f(u)$, where $f(u)$ satisfies the conditions $f'(u) > 0$ and $f(0) = 0$. The power law form $f(u) = u^{1/n}/\epsilon$ is common, with $n = 1$ being the case considered in the main text. It is assumed that the function $u_{n-1}(x)$ (defined over $[0, s_{n-1}]$) and s_{n-1} are both known. The free boundary problem (A.1)–(A.4) may be solved by the method of invariant embedding (or the sweep method), as described in [28].

Writing (A.1)–(A.4) as a first-order system over $(0, s_n)$

$$u_n' = v_n, \qquad v_n' = \frac{1}{\Delta t}(u_n - u_{n-1}), \tag{A.5}$$

the Riccati transformation

$$v_n(x) = R(x)u_n(x) + W_n(x) \tag{A.6}$$

relates u_n and v_n, where

$$R' = \frac{1}{\Delta t} - R^2, \qquad R(0) = 0, \tag{A.7}$$

and

$$W_n' = -R(x)W_n - \frac{u_{n-1}(x)}{\Delta t}, \qquad W_n(0) = 0. \tag{A.8}$$

(A.3) and (A.6) show that s_n is a root of the scalar equation

$$\sigma_n(x) \equiv \frac{(x - s_{n-1})}{\Delta t} - f\left(\frac{-(W_n(x) + \lambda\frac{(x-s_{n-1})}{\Delta t})}{R(x) + (1-\mu)\left(\frac{x-s_{n-1}}{\Delta t}\right)}\right) = 0. \tag{A.9}$$

Given s_n, we have

$$u_n(s_n) = -\frac{\lambda\dot{s}_n + W_n(s_n)}{R(s_n) + (1-\mu)\dot{s}_n}, \tag{A.10}$$

where

$$\dot{s}_n = \frac{s_n - s_{n-1}}{\Delta t} = f(u_n(s_n)) \tag{A.11}$$

and

$$u_n'(s_n) = v_n(s_n) = -\dot{s}_n \left(\frac{\lambda R(s_n) - (1-\mu) W_n(s_n)}{R(s_n) + (1-\mu)\dot{s}_n} \right). \tag{A.12}$$

The triple $\{u_n(s_n), v_n(s_n), s_n\}$ is an exact solution of (A.3), (A.4), and (A.6).

Once s_n is determined, then v_n can be found by integrating backward over $[0, s_n)$ the reverse sweep equation

$$v_n' = \frac{1}{\Delta t} \left(R(x) v_n + W_n(x) - u_{n-1}(x) \right), \tag{A.13}$$

with $v_n(s_n)$ given in (A.12).

The above algorithm was implemented as follows. A time-independent mesh with grid points $\{x_i\}_{i=0}^N$ was imposed on the interval $[0,1]$, the grid being uniform with $x_i = i/N$. Nonuniform grids with clustering of grid points where the free boundary moves quickly may also be used. The time step Δt is variable. The Riccati equation (A.7) has the closed form solution

$$R(x, \Delta t) = \frac{1}{\sqrt{\Delta t}} \tanh\left(\frac{x}{\sqrt{\Delta t}} \right), \tag{A.14}$$

while the linear equation (A.8) is integrated to give

$$W_n(x) \cosh\left(\frac{x}{\sqrt{\Delta t}} \right) = -\int_0^x \frac{u_{n-1}(\tau)}{\Delta t} \cosh\left(\frac{\tau}{\sqrt{\Delta t}} \right) d\tau$$

and evaluated using a suitable quadrature rule (here the simple trapezoidal rule was used). The function $\sigma_n(x)$ is evaluated at successive mesh points of the grid before s_{n-1} until it changes sign between, say, x_ℓ and $x_{\ell-1}$. The free boundary s_n is now determined as the root of the quadratic interpolant through $\sigma(x_{\ell-1}), \sigma(x_\ell)$, and $\sigma(x_{\ell+1})$. Equation (A.6) is also integrated to give

$$u_n(x) = u_n(s_n) \frac{\cosh\left(\frac{x}{\sqrt{\Delta t}} \right)}{\cosh\left(\frac{s_n}{\sqrt{\Delta t}} \right)} + \cosh\left(\frac{x}{\sqrt{\Delta t}} \right) \int_{s_n}^x \frac{W_n(\tau)}{\cosh\left(\frac{\tau}{\sqrt{\Delta t}} \right)} d\tau$$

and evaluated by the trapezoidal rule, first from s_n to $x_{\ell-1}$ and then backward over the fixed mesh.

In the special case $\epsilon = 0$, (A.3) and (A.4) reduce to

$$u_n(s_n) = 0, \qquad s_0 = s(0) = 1, \tag{A.15}$$

$$u_n'(s_n) = -\lambda \frac{(s_n - s_{n-1})}{\Delta t}, \tag{A.16}$$

with (A.9) replaced with

$$\sigma_n(x) = W_n(x) + \lambda \frac{(x - s_{n-1})}{\Delta t} = 0 \tag{A.17}$$

and (A.10)–(A.12) modified to

$$u_n(s_n) = 0, \qquad \dot{s}_n = \frac{s_n - s_{n-1}}{\Delta t}, \qquad u_n'(s_n) = v_n(s_n) = W_n(s_n) = -\lambda \dot{s}_n.$$

The algorithm for this special case proceeds as before.

Acknowledgment. Financial support from the Leverhulme Trust is gratefully acknowledged by J. R. King.

REFERENCES

[1] F. Abergel, D. Hilhorst, and F. Issard-Roch, *On a dissolution-growth problem with surface tension in the neighborhood of a stationary solution*, SIAM J. Math. Anal., 24 (1993), pp. 299–316.

[2] D. V. Alexandrov, *On the theory of the formation of the two-phase concentration-supercooling region*, Dokl. Phys., 48 (2003), pp. 481–486.

[3] J. Chadam and P. Ortoleva, *The stabilizing effect of surface tension on the development of the free boundary in a planar, one-dimensional, Cauchy-Stefan problem*, IMA J. Appl. Math., 30 (1983), pp. 57–66.

[4] J. Chadam, S. D. Howison, and P. Ortoleva, *Existence and stability for spherical crystals growing in a supersaturated solution*, IMA J. Appl. Math., 39 (1987), pp. 1–15.

[5] Ch. Charach and B. Zaltzman, *Planar solidification from an undercooled melt: Asymptotic solutions to a continuum model with interfacial kinetics*, Phys. Rev. E, 47 (1993), pp. 1230–1234.

[6] Ch. Charach and B. Zaltzman, *Analytic model for planar growth of a solid germ from an undercooled melt*, Phys. Rev. E, 49 (1994), pp. 4322–4327.

[7] Ch. Charach, B. Zaltzman, and I. G. Götz, *Interfacial kinetics effect in planar solidification problems without initial undercooling*, Math. Models Methods Appl. Sci., 4 (1994), pp. 331–354.

[8] X. Chen and F. Reitich, *Local existence and uniqueness of solutions of the Stefan problem with surface tension and kinetic undercooling*, J. Math. Anal. Appl., 164 (1992), pp. 350–362.

[9] S. R. Coriell and R. L. Parker, *Interface kinetics and the stability of the shape of a solid sphere growing from the melt*, in Proceedings of the International Conference on Crystal Growth, Boston, H. S. Peiser, ed., Pergamon Press, New York, 1966, pp. 20–24.

[10] S. R. Coriell and R. F. Sekerka, *Oscillatory morphological instabilities due to non-equilibrium segregation*, J. Crystal Growth, 61 (1983), pp. 499–508.

[11] J. N. Dewynne, S. D. Howison, J. R. Ockendon, and W. Xie, *Asymptotic behaviour of solutions to the Stefan problem with a kinetic condition at the free boundary*, J. Austral. Math. Soc. Ser. B, 31 (1989), pp. 81–96.

[12] S. H. Doole, *A Stefan-like problem with a kinetic condition and surface tension effects*, Math. Comput. Modelling, 23 (1996), pp. 55–67.

[13] C. M. Elliott and J. R. Ockendon, *Weak and Variational Methods for Moving Boundary Problems*, Res. Notes in Math. 59, Pitman, Boston, London, 1982.

[14] J. D. Evans and J. R. King, *Asymptotic results for the Stefan problem with kinetic undercooling*, Quart. J. Mech. Appl. Math., 53 (2000), pp. 449–473.

[15] J. D. Evans and J. R. King, *The Stefan problem with nonlinear kinetic undercooling*, Quart. J. Mech. Appl. Math., 56 (2003), pp. 139–161.

[16] A. Fasano and M. Primicerio, *General free boundary problems for the heat equation*. I, J. Math. Anal. Appl., 57 (1977), pp. 694–723.

[17] A. Fasano and M. Primicerio, *General free boundary problems for the heat equation*. II, J. Math. Anal. Appl., 58 (1977), pp. 202–231.

[18] A. Fasano and M. Primicerio, *General free boundary problems for the heat equation*. III, J. Math. Anal. Appl., 59 (1977), pp. 1–14.

[19] A. Fasano, G. H. Meyer, and M. Primicerio, *On a problem in the polymer industry: Theoretical and numerical investigation of swelling*, SIAM J. Math. Anal., 17 (1986), pp. 945–960.

[20] A. Fasano, M. Primicerio, and A. A. Lacey, *New results on some classical parabolic free boundary problems*, Quart. Appl. Math., 38 (1981), pp. 439–460.

[21] A. Fasano, M. Primicerio, S. D. Howison, and J. R. Ockendon, *Some remarks on the regularization of supercooled one-phase Stefan problems in one dimension*, Quart. Appl. Math., 48 (1990), pp. 153–168.

[22] A. Friedman and F. Reitich, *The Stefan problem with small surface tension*, Trans. Amer. Math. Soc., 328 (1991), pp. 465–515.

[23] I. G. Götz and B. Zaltzman, *Two-phase Stefan problem with supercooling*, SIAM J. Math. Anal., 26 (1995), pp. 694–714.

[24] M. E. Gurtin, *Thermodynamics and the supercritical Stefan equations with nucleations*, Quart. Appl. Math., 52 (1994), pp. 133–155.

[25] M. A. Herrero and J. J. L. Velazquez, *Singularity formation in the one-dimensional supercooled Stefan problem*, European J. Appl. Math., 7 (1996), pp. 119–150.
[26] M. A. Herrero, E. Medina, and J. J. L. Velazquez, *The birth of a cusp in the two-dimensional, undercooled Stefan problem*, Quart. Appl. Math., 58 (2000), pp. 473–494.
[27] O. J. Ilegbusi and M. D. Mat, *A review of the modelling of multi-phase phenomena in materials processing.* I. *Solid-liquid systems*, J. Materials Processing & Manufacturing Sci., 8 (2000), pp. 188–217.
[28] G. H. Meyer, *One-dimensional parabolic free boundary problems*, SIAM Rev., 19 (1977), pp. 17–34.
[29] W. W. Mullins and R. F. Sekerka, *Morphological stability of a particle growing by diffusion or heat flow*, J. Appl. Phys., 34 (1963), pp. 323–329.
[30] R. J. Schaefer and M. E. Glicksman, *Fully time-dependent theory for the growth of spherical crystal nuclei*, J. Crystal Growth, 5 (1969), pp. 44–58.
[31] J. F. Scheid, *A dissolution-growth problem with surface tension: Local existence and uniqueness*, Appl. Math. Lett., 8 (1995), pp. 91–95.
[32] G. Scianna, *Global existence and asymptotic behavior for the radially symmetric case of a two-phase Stefan problem with interfacial energy*, J. Math. Anal. Appl., 211 (1997), pp. 1–29.
[33] B. Sherman, *A general one-phase Stefan problem*, Quart. Appl. Math., 28 (1970), pp. 377–382.
[34] A. Umantsev and S. H. Davis, *Growth from a hypercooled melt near absolute stability*, Phys. Rev. A, 45 (1992), pp. 7195–7201.
[35] A. Visintin, *Stefan problem with a kinetic condition at the free boundary*, Ann. Mat. Pura Appl., 146 (1987), pp. 97–122.
[36] W. Q. Xie, *The Stefan problem with a kinetic condition at the free boundary*, SIAM J. Math. Anal., 21 (1990), pp. 362–373.
[37] F. Yi, *Asymptotic behaviour of the solutions of the supercooled Stefan problem*, Proc. Roy. Soc. Edinburgh Sect. A, 127 (1997), pp. 181–190.
[38] Q. Zhu, A. Peirce, and J. Chadam, *Initiation of shape instabilities of free boundaries in planar Cauchy-Stefan problems*, European J. Appl. Math., 4 (1993), pp. 419–436.

SIAM J. APPL. MATH.
Vol. 65, No. 5, pp. 1708–1725

WEAKLY NONLINEAR AND NUMERICAL ANALYSES OF DYNAMICS IN A SOLID COMBUSTION MODEL*

L. K. GROSS† AND J. YU‡

Abstract. This paper contains qualitative and quantitative comparisons between a weakly nonlinear analysis and direct numerical simulations of a free-boundary problem. The former involves modulating the most linearly unstable mode, taking a small perturbation of the neutrally stable value ν_c of a parameter ν related to the activation energy. Analogously, we perform the direct numerical computations near the marginally unstable value, namely, $\nu = \nu_c - \epsilon^2$, where ϵ is rather small.

We delineate the role of a different parameter σ (related to the Arrhenius kinetics) in the combustion dynamics when $\nu = \nu_c - \epsilon^2$. In particular, the numerics show that varying σ produces a period-doubling scenario when ϵ lies approximately between 0.08 and 0.12. We describe the σ intervals within which complex dynamics occur for various values of ϵ and for ν fixed at $\nu_c - \epsilon^2$. When ϵ drops to approximately 0.06, the asymptotic and numerical solutions agree well for all physical values of σ.

Key words. free-boundary problems, condensed-phase combustion, weak instability, asymptotic expansions, Crank–Nicolson method, Fourier transforms

AMS subject classifications. 35R35, 80A25, 80M35, 80M25, 42A38

DOI. 10.1137/S0036139904439508

1. Introduction.

In this article, we study the nonuniform dynamics of front propagation in a free-boundary model of solid combustion, through both weakly nonlinear analysis and direct simulations. We do the first quantitative comparison of the two methods.

The asymptotic technique applies to the weakly unstable setting. In particular, we fix the bifurcation parameter ν related to the activation energy to within a rather small number ϵ^2 of the neutrally stable value ν_c. By solving numerically in the same regime, we closely investigate the role of a parameter σ associated with the Arrhenius kinetics. In particular, period-doubling and eventual chaos develop as the kinetics parameter σ decreases (and the bifurcation parameter ν remains at a deviation of ϵ^2 from its critical value).

Weakly nonlinear analysis involves modulating the most linearly unstable mode. Within quite a small neighborhood of the neutral stability boundary, Fourier spectra of the numerical quasi-steady-state solutions indicate a regime in which a single mode dominates, as well as complex regimes of front propagation.

As the bifurcation parameter ν approaches ever closer to the neutrally stable value, the range of the parameter σ for which period-doubling and other strongly nonlinear dynamics occur shrinks. Sufficiently near the stability threshold (ϵ approximately 0.06), numerical solutions for all values of σ agree closely with the weakly

*Received by the editors January 6, 2004; accepted for publication (in revised form) November 3, 2004; published electronically July 13, 2005.

http://www.siam.org/journals/siap/65-5/43950.html

†Department of Theoretical and Applied Mathematics, The University of Akron, Akron, OH 44325-4002 (gross@math.uakron.edu). The research of this author was supported by the National Science Foundation under grant DMS-0074965, the Ohio Board of Regents (Research Challenge Grant), and the Buchtel College of Arts and Sciences at The University of Akron (matching grant).

‡Department of Mathematics and Statistics, The University of Vermont, Burlington, VT (jun.yu@uvm.edu).

nonlinear solutions. By varying ϵ, we quantify the domain of applicability of the weakly nonlinear analysis.

The problem under consideration models, for example, solid combustion, in which a chemical reaction converts a solid fuel directly into solid products with no intermediate gas phase formation. For instance, in self-propagating high-temperature synthesis (SHS), a flame wave advancing through powdered ingredients leaves high-quality ceramic materials or metallic alloys in its wake. (See, for instance, [15, 17, 20].)

The propagation results from the interplay between heat generation and heat diffusion in the medium. A balance exists between the two in some parametric regimes, producing a constant burning rate. In other cases, competition between the reaction and the diffusion results in a wide variety of nonuniform behaviors, some leading to chaos.

Shkadinsky, Khaikin, and Merzhanov [18] predicted the simplest oscillatory regimes through numerical simulation on reaction-diffusion partial differential equations (PDEs). The system contains Arrhenius-kinetics terms that account for chemical conversion throughout the spatial domain.

Various works have explored numerically the dynamics of models that employ approximations to the Arrhenius kinetics. For instance, in [1], Arrhenius kinetics with a cutoff was used to observe chaotic pulsations, following a number of period-doubling bifurcations.

Other approximations exploit the narrowness of the reaction zone. A point-source model has an exact traveling-wave solution and is more amenable to analysis than one with the full Arrhenius kinetics. Matkowsky and Sivashinsky [14] studied a concentrated-kinetics model in the case of large activation energy. The δ-function kinetics follow from an analysis similar to that of [19].

This free-interface problem has been studied numerically in [4]. For a sufficiently large activation energy, the work showed transitions to chaos via a period-doubling solution and highly irregular relaxational oscillations. The authors attributed a lack of sequential secondary bifurcations to the difference between the point-source and distributed-kinetics models (as in [1]). Later, however, in [9], the entire spectrum of behavior was observed for the free-interface model, as previously had been seen for distributed kinetics.

In [9], the authors performed numerical computations on a second model of solid combustion as well. They motivate it by noting that both the reaction-diffusion model as in [18] and the free-interface model in [14] assume a constant value of thermal diffusivity. However, some problems manifest a clear dependence of this parameter on degree of conversion. In fact, when the burnt product is a foam-like substance, heat diffusion in the product region is negligible. For such cases, they consider a model that includes the heat equation on a semi-infinite domain ahead of the reaction and a nonlinear kinetic condition imposed on the moving boundary. The present paper uses this free-boundary problem.

Note that both the free-interface (two-sided) model and the free-boundary (one-sided) model stem from reaction-diffusion PDEs with full Arrhenius kinetics. To emphasize, the one-sided model is not an adaptation of the two-sided model; rather each of them is a viable derivative of the reaction-diffusion model. The two-sided model assumes a single constant conductivity throughout the reactant and product zones. The one-sided model assumes zero conductivity in the burned region. In some cases the first approximation is more appropriate, in others the second.

Belyaev and Komkova discovered a pulsating regime in the burning of a chrome-

magnesium thermite in 1950 [2]. A planar front may have oscillated with a constant frequency in their experiments, but they did not observe the process in detail. Later Merzhanov, Filonenko, and Borovinskaya [16] observed experimentally both the periodic propagation of a flat front in SHS as well as spinning waves, showing a fuller understanding of the behaviors. All the models discussed in this literature review exhibit the same spectrum of dynamics as experiments. Specifically, we refer to computed solutions of (i) the reaction-diffusion system governed by the full Arrhenius kinetics (e.g., [5]), (ii) the reaction-diffusion system with Arrhenius kinetics with a cutoff (e.g., [1]), and models that use point-source kinetics like (iii) the free-interface ("two-sided") model with constant heat diffusivity (e.g., [9]), as well as (iv) the free-boundary ("one-sided") model, in which heat transfer behind the flame front (in the burned matter) is qualitatively unimportant (e.g., [9]).

Simulations on all these models show the same dynamical behaviors as one pushes the bifurcation parameters deeper into the instability regions. In particular, numerical simulations and analysis in [9] show that dynamics of the two-sided and one-sided problems agree extremely closely.

In the present work, we fix the bifurcation parameter ν within ϵ^2 of the neutrally stable value and vary the kinetics parameter σ, rather than exploring regimes more and more strongly unstable in ν. In addition, we vary ϵ, thereby also changing ν, and study the impact on the dynamics with respect to the kinetics parameter σ. We will point out the agreement with dynamical scenarios described in previous studies, which use a variety of models.

The stability thresholds for uniformly propagating fronts generally differ for all of the different kinetics mentioned, however. Distributed kinetics have only the numerical approximate bifurcation values. Intricate bifurcation analyses [13, 10] of instabilities for the point-source models have also classified the interactions of clockwise and counterclockwise spinning waves on the surface of a cylinder. Margolis's review paper [13] includes a thorough discussion of resonance phenomena, treating sample radii that yield close, as well as equal, eigenvalues. Also, Booty, Margolis, and Matkowsky [3] predicted cascades of bifurcations from a double eigenvalue of a linearized model of condensed-phase combustion in a long cylindrical sample. They show that the inclusion of melting in the model makes the neutral-stability threshold more accessible. A bifurcation parameter ν in the present work is restricted to a smaller neighborhood of the value corresponding to a single neutrally stable eigenvalue. A different parameter σ is varied to produce period-doubling behaviors numerically.

Combustion in two dimensions can be described by a one-dimensional model when the only unstable mode corresponds to the dynamics with no spatial variation in the transverse direction. For example, the linear stability analysis in [12] shows that for a free-boundary model, a flat front dominates the behavior for the case of a sufficiently narrow strip of material with insulated edges.

In particular, to satisfy the boundary conditions, the wave numbers are integer multiples of π/a, where a is the strip width [12]. If $a < \pi$, all modes are stable for $\nu > 1/3$. Exactly one mode (the zeroth mode) loses stability at $\nu_c = 1/3$. The zero mode corresponds to the dynamics with no spatial variation in the transverse direction (i.e., to the one-dimensional case). If, on the other hand, $\pi < a < 2\pi$, then, as we decrease ν, the first mode π/a loses stability prior to the flat mode, namely, at a value of $\nu > 1/3$. In both cases ($a < \pi$ and $\pi < a < 2\pi$) the weakly nonlinear analysis shows that the evolution is governed by a complex Landau–Stuart ordinary differential equation [11]. (See what follows for the narrow-strip analysis.)

At $a = \pi$, the flat mode and the first wavy mode both lose stability at $\nu_c = 1/3$, while the other modes remain stable. The nonlinear interaction of the flat and curvy modes is the subject of the weakly nonlinear analysis in [12], which culminates in the derivation of a system of two complex Landau–Stuart equations. (Notice that if the width a is infinite, a continuum of modes goes unstable, and the evolution is governed by Ginzburg–Landau PDEs.)

In the remainder of this section, we introduce the governing equations and, for convenience, summarize a linear stability analysis. Because we consider the case in which a zero-wavenumber mode is the most unstable, we present the model in one space dimension. In [21], we do a full linear stability study for the two-dimensional problem formulated as an initial-value problem.

Section 2 contains a weakly nonlinear analysis, and section 3 shows simulations in the marginally unstable regime. In computations, the dynamics unfold as the parameter σ associated with the Arrhenius kinetics decreases (while the bifurcation parameter ν remains fixed within ϵ^2 of its neutrally stable value).

Section 4 presents quantitative comparisons of the asymptotic solutions and computed solutions. Some qualitative comparisons for a similar problem—involving competing flat and wavy (two-dimensional) modes—appear in [6] (together with numerics that venture into more strongly unstable regimes than in the present paper). Here we investigate the numerical solutions for marginally unstable values of the activation energy, allowing a full range of kinetics-parameter values.

Specifically, we perform the computations with ν fixed near the marginally unstable value, namely, $\nu = \nu_c - \epsilon^2$, where ϵ is fairly small. For ϵ smaller than about 0.12, we see the smooth periodic solutions that the weakly nonlinear analysis predicts, provided σ has an appropriate value. In particular, Fourier transforms of the numerical data illustrate the ranges of σ in which the analysis accurately predicts the quantitative behavior of solutions.

The data simultaneously reveal the development of complex dynamics in various kinetics-parameter regimes (with the inverse activation energy ν held at ϵ^2 units below the stability threshold), when ϵ exceeds about 0.06. When ϵ drops below this value, the σ intervals of strongly nonlinear dynamics disappear.

In the model, we seek the temperature distribution $u(x,t)$ in one spatial dimension and the interface position $\Gamma(t) = \{x | x = f(t)\}$ that satisfy the appropriately nondimensionalized free-boundary problem

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad x > f(t), \quad t > 0, \tag{1.1}$$

$$V = G\left(u\big|_{\Gamma}\right), \quad t > 0, \tag{1.2}$$

$$\left.\frac{\partial u}{\partial x}\right|_{\Gamma} = -V, \quad t > 0. \tag{1.3}$$

Here V is the velocity of the rightward-traveling interface, i.e.,

$$V = \frac{df}{dt}.$$

In addition, the temperature satisfies the condition

$$u \to 0 \text{ as } x \to \infty; \tag{1.4}$$

that is, the ambient temperature is normalized to zero at infinity.

To model solid combustion, we take the Arrhenius function as the kinetics function G in the nonequilibrium interface condition (1.2) [4, 17]. Then, with appropriate nondimensionalization, the velocity of propagation relates to the interface temperature as

$$V = \exp\left[\left(\frac{1}{\nu}\right)\frac{u-1}{\sigma+(1-\sigma)u}\right] \tag{1.5}$$

at the interface Γ. Here ν is inversely proportional to the activation energy of the exothermic chemical reaction that occurs at the interface, and $0 < \sigma < 1$ is the ambient temperature nondimensionalized by the adiabatic temperature of combustion products. (See [8].)

Inverting the Arrhenius function (1.5), we reexpress the boundary condition (1.2) in the form

$$u|_\Gamma = 1 + \nu K(V;\sigma,\nu), \tag{1.6}$$

where

$$K(V;\sigma,\nu) = \frac{\ln(V)}{1-(1-\sigma)\nu\ln(V)}. \tag{1.7}$$

Note the function $K(V)$ has been introduced to have the convenient properties $K(1) = 0$, $K'(1) = 1$.

For ease of subsequent asymptotic and numerical analysis, we reformulate the problem in the front-attached coordinate frame:

$$\eta = x - f(t), \quad \tau = t.$$

Problem (1.1)–(1.6) then takes the form

$$\frac{\partial u}{\partial \tau} = \frac{\partial^2 u}{\partial^2 \eta} + V\frac{\partial u}{\partial \eta}, \quad \eta > 0,\ \tau > 0, \tag{1.8}$$

$$u|_\Gamma = u(0,\tau) = 1 + \nu K(V), \tag{1.9}$$

$$\left.\frac{\partial u}{\partial \eta}\right|_\Gamma = \left.\frac{\partial u}{\partial \eta}\right|_{(0,\tau)} = -V, \tag{1.10}$$

$$\lim_{\eta\to\infty} u = 0. \tag{1.11}$$

The free-boundary problem (1.8)–(1.11) admits a traveling-wave solution

$$u_0(\eta,\tau) = \exp(-\eta), \quad f_0(\tau) = \tau. \tag{1.12}$$

The problem linearized about the traveling wave has a normal-mode solution of the form

$$w = e^{\lambda\tau} g(\eta;\lambda), \quad \phi = e^{\lambda\tau}, \tag{1.13}$$

where w and ϕ represent the perturbations about u_0 and f_0, respectively. Substituting them into the linearized problem produces an eigenvalue problem in λ and $g(\eta;\lambda)$.

The discrete spectrum values are zero and

$$\lambda = \frac{1 - 3\nu \pm \sqrt{(3\nu-1)^2 - 4\nu^3}}{2\nu^2}. \tag{1.14}$$

The eigenfunction corresponding to the eigenvalue λ is

$$g(\eta;\lambda,\nu) = (1+\nu\lambda)\exp\left(-\left(1+\sqrt{1+4\lambda}\right)\frac{\eta}{2}\right) - \exp(-\eta). \tag{1.15}$$

Linearly unstable behavior occurs for this system only when $\Re\lambda$ is positive.

The basic solution (1.12) is neutrally stable under a small perturbation of the form (1.13) if $\Re\lambda = 0$. Setting $\Re\lambda = 0$ in (1.14) gives the critical value ν_c of ν, namely,

$$\nu_c = \frac{1}{3}. \tag{1.16}$$

The corresponding neutrally stable eigenvalues from (1.14) are $\pm i\omega$, where

$$\omega = \sqrt{3}. \tag{1.17}$$

If $\nu < 1/3$, then $\Re\lambda > 0$, and the basic solution is linearly unstable. (See, for example, [12, 21].)

2. Weakly nonlinear analysis. Let ϵ^2 be a small deviation from the neutrally stable value of ν, namely,

$$\epsilon^2 = \nu_c - \nu = \frac{1}{3} - \nu. \tag{2.1}$$

We consider the time scales

$$t_0 = \tau, \quad t_1 = \epsilon\tau, \quad t_2 = \epsilon^2\tau$$

as independent variables, so that $\partial/\partial\tau = \partial/\partial t_0 + \epsilon\,\partial/\partial t_1 + \epsilon^2\,\partial/\partial t_2$. We then seek a solution of the form

$$\begin{aligned} u(\eta,t_0,t_1,t_2) &= e^{-\eta} + \epsilon A(t_1,t_2)e^{i\sqrt{3}t_0} g\left(\eta; i\sqrt{3}, \frac{1}{3}\right) \\ &\quad + \epsilon^2 w_2(\eta,t_0,t_1,t_2) + \cdots + \text{CC}, \\ f(t_0,t_1,t_2) &= t_0 + \epsilon\left\{A(t_1,t_2)e^{i\sqrt{3}t_0} + \frac{1}{2}B(t_1,t_2)\right\} \\ &\quad + \epsilon^2\phi_2(t_0,t_1,t_2) + \cdots + \text{CC}, \end{aligned} \tag{2.2}$$

where $A(t_1,t_2)$ is complex, and "CC" stands for complex-conjugate terms. The real-valued function $B(t_1,t_2)$ modulates the constant-velocity solution to the linearized problem.

Notice that in $O(\epsilon)$, the weakly nonlinear solution (2.2) has only one Fourier term in t_0. We will show below in (2.19)–(2.20) that the $O(\epsilon^2)$ term contains the second harmonic. We refer to the expansion (2.2) as a "single-mode approximation" because the leading-order perturbation contains only one mode in fast time.

Making the substitutions (2.2) and equating like powers of ϵ results in subproblems for the terms in the perturbation expansions above, subject to solvability conditions on the amplitudes A and B. The $O(1)$ problem is satisfied identically because in (2.2) we took the temperature-interface pair (u, f) perturbed about $(e^{-\eta}, t_0)$, a solution to the nonlinear problem (1.8)–(1.11). The $O(\epsilon)$ problem is just the linearized problem with $\nu = \nu_c = 1/3$, which is satisfied identically by the $O(\epsilon)$ terms in the expansions (2.2).

The problems of order ϵ^j, $j = 2, 3$, are

$$\frac{\partial w_j}{\partial t_0} - \frac{\partial^2 w_j}{\partial \eta^2} - \frac{\partial w_j}{\partial \eta} + e^{-\eta}\frac{\partial \phi_j}{\partial t_0} = Q_j(\eta, \mathbf{t}), \tag{2.3}$$

$$w_j|_{\eta=0} - \frac{1}{3}\frac{\partial \phi_j}{\partial t_0} = \alpha_j(\mathbf{t}), \tag{2.4}$$

$$\left.\frac{\partial w_j}{\partial \eta}\right|_{\eta=0} + \frac{\partial \phi_j}{\partial t_0} = \beta_j(\mathbf{t}), \tag{2.5}$$

$$\lim_{\eta\to\infty} w_j = 0, \tag{2.6}$$

where $\mathbf{t} = (t_0, t_1, t_2)$. For brevity, we have named the right-hand sides above as Q_j, α_j, and β_j. The PDEs (2.3) can be represented as

$$\mathcal{L}_1 w_j + \mathcal{L}_2 \phi_j = \mathcal{P}(w_1, \phi_1, \ldots, w_{j-1}, \phi_{j-1}). \tag{2.7}$$

$\mathcal{L}_1$ and $\mathcal{L}_2$ are linear operators on bounded functions in $L^2(\Omega)$, where $\Omega = \{(\eta, \tau) | 0 \le \eta < \infty,\ 0 \le \tau < \infty\}$.

According to Fredholm's alternative, equation (2.7) has a nonsecular (bounded-in-time) solution if the right-hand side is orthogonal to the null space of the adjoint operator $\mathcal{L}^*$. That is,

$$(\mathcal{L}_1 w_j + \mathcal{L}_2 \phi_j,\ v) = 0 \tag{2.8}$$

for $v \in \ker \mathcal{L}^*$ and the inner product defined such that

$$(f_1, f_2) = \lim_{T\to\infty} \frac{1}{T} \int_0^T \int_0^\infty f_1(\eta, \tau)\overline{f_2(\eta, \tau)}\, d\eta\, d\tau. \tag{2.9}$$

The quantity v in (2.8) satisfies

$$\left(-\frac{\partial}{\partial t_0} - \frac{\partial^2}{\partial \eta^2} + \frac{\partial}{\partial \eta}\right) v = 0, \tag{2.10}$$

$$2i\frac{\sqrt{3}}{3}\ v|_{\eta=0} - \left(1 - i\frac{1}{3}\right)\left.\frac{\partial v}{\partial \eta}\right|_{\eta=0} = 0. \tag{2.11}$$

Nonzero solutions are

$$u_1(\eta, t_0; i\sqrt{3}) = \exp(i\sqrt{3}t_0)h(\eta; i\sqrt{3}) \ \text{ and } \ u_0(\eta, t_0; 0) = 1, \tag{2.12}$$

where

$$h(\eta; i\sqrt{3}) = \exp\left(\left(1 - \sqrt{1 - 4i\sqrt{3}}\right)\frac{\eta}{2}\right). \tag{2.13}$$

Substituting $v = u_1$ into the solvability condition (2.8) when $j = 2$ produces the differential equation

$$\frac{\partial A}{\partial t_1} = 0. \tag{2.14}$$

Substituting $v = u_0$ into the solvability condition (2.8) when $j = 2$ produces the differential equation

$$\frac{\partial B}{\partial t_1} = A\bar{A}r_0, \quad r_0 = -3(2 + K''(1)). \tag{2.15}$$

Substituting (2.14) and (2.15) into the $O(\epsilon^2)$ problem in (2.3)–(2.6) yields the problem

$$\frac{\partial w_2}{\partial t_0} - \frac{\partial^2 w_2}{\partial \eta^2} - \frac{\partial w_2}{\partial \eta} + e^{-\eta}\frac{\partial \phi_2}{\partial t_0} = A^2 e^{2i\sqrt{3}t_0}\mathcal{R}_2(\eta) + A\bar{A}\mathcal{R}_0(\eta) + \text{CC}, \tag{2.16}$$

$$w_2|_{\eta=0} - \frac{1}{3}\frac{\partial \phi_2}{\partial t_0} = A^2 e^{2i\sqrt{3}t_0}F_2 + A\bar{A}F_0 + \text{CC}, \tag{2.17}$$

$$\left.\frac{\partial w_2}{\partial \eta}\right|_{\eta=0} + \frac{\partial \phi_2}{\partial t_0} = A^2 e^{2i\sqrt{3}t_0}G_2 + A\bar{A}G_0 + \text{CC}. \tag{2.18}$$

The solution (w_2, ϕ_2) consists of a homogeneous and a particular solution. Because only the inhomogeneous terms will contribute to the solvability condition at the next order, we present the nonsecular solution as

$$w_2 = A^2 e^{2i\sqrt{3}t_0}g_2(\eta) + A\bar{A}g_0(\eta) + \text{CC}, \tag{2.19}$$

$$\phi_2 = A^2 e^{2i\sqrt{3}t_0}C_2 + A\bar{A} + \text{CC}, \tag{2.20}$$

where $g_j(\eta)$, $j = 0, 2$, satisfy the initial-value problems

$$g_j'' + g_j' - ji\sqrt{3}g_j = ji\sqrt{3}C_2 e^{-\eta} - \mathcal{R}_j(\eta), \tag{2.21}$$

$$g_j(0) = ji\frac{\sqrt{3}}{3}C_2 + F_j, \tag{2.22}$$

$$g_j'(0) = -ji\sqrt{3}C_2 + G_j, \tag{2.23}$$

$$g_2(\eta) \to 0 \text{ as } \eta \to \infty, \tag{2.24}$$

where

$$\mathcal{R}_j(\eta) = -(-1)^{j/2}g'(\eta)i\sqrt{3} - \frac{2-j}{4}r_0 e^{-\eta}, \tag{2.25}$$

$$F_j = \frac{1}{2}(-1)^{j/2}K''(1) + \frac{2-j}{12}r_0, \tag{2.26}$$

$$G_j = -\frac{2-j}{4}r_0. \tag{2.27}$$

Recall that r_0 is given in (2.15). Also, $g(\eta) = g(\eta; i\sqrt{3})$ is defined via (1.15).

Substituting $v = u_1$ into the solvability condition (2.8) when $j = 3$ produces the Landau–Stuart equation

$$\frac{dA}{dt_2} = \chi A + \beta A^2 \bar{A}, \tag{2.28}$$

where

$$\chi \equiv -\frac{\partial \lambda_0}{\partial \nu}\bigg|_{\nu=\nu_c} = \frac{3}{2}(9 + \sqrt{3}i). \tag{2.29}$$

The coefficient β is defined as

$$\beta = \frac{\int_0^\infty R(\eta)\bar{h}(\eta)\, d\eta + F\mathcal{U}}{\int_0^\infty (g(\eta) + e^{-\eta})\bar{h}(\eta)\, d\eta - \frac{1}{3}\mathcal{U} - 1}, \tag{2.30}$$

where

$$R(\eta) = r_0 g'(\eta) + i\sqrt{3}\,[2C_2\bar{g}'(\eta) - g_2'(\eta) + 2\mathrm{Re}\,(g_0'(\eta))]; \tag{2.31}$$

$$F = \frac{1}{3}(6C_2 + ir_0\sqrt{3})K''(1) + i\frac{\sqrt{3}}{2}K'''(1); \tag{2.32}$$

$$\mathcal{U} = -\frac{1}{2}(3 + i\sqrt{3}). \tag{2.33}$$

Once we solve the Landau–Stuart equation (2.28) subject to an initial condition, the full asymptotic expansion (2.2) is known with w_2 and ϕ_2 given in (2.19)–(2.20) and B given in (2.15). In what follows, we compare the asymptotic solution with a numerical solution over the range $0 < \sigma < 1$ with ν fixed at a small deviation ϵ^2 from the neutrally stable value $1/3$, as given in (2.1).

The amplitude equation (2.28) determines the dynamics of the unstable mode $A(t_2)e^{i\sqrt{3}t_0}$, subject to self-interaction. The dynamics of the mode depend on the relationships between the coefficients χ and β and are affected by the kinetics function $K(V)$, introduced in (1.7). Recall that the function $K(V)$ is normalized such that $K(1) = 0$ and $K'(1) = 1$. The form of the kinetics function comes into play via $K''(1)$ and $K'''(1)$, which appear explicitly in r_0, F_j, and F of (2.15), (2.26), and (2.32), respectively. Note from (2.17) that F_j, $j = 2, 0$, are the coefficients of $A^2e^{2i\sqrt{3}t_0}$ and $A\bar{A}$, respectively, on the right-hand side of an $O(\epsilon^2)$ boundary condition. Also, F is the coefficient of $A^2\bar{A}e^{i\sqrt{3}t_0}$ on the right-hand side $\alpha_3(\mathbf{t})$ of the $O(\epsilon^3)$ boundary condition (2.4). In particular, $\alpha_3(\mathbf{t})$ has the form

$$\alpha_3(\mathbf{t}) = \left\{ \left(\frac{\partial A}{\partial t_2} - \chi A\right)\frac{1}{3}e^{i\sqrt{3}t_0} + A^3e^{3i\sqrt{3}t_0}F_3 + A^2\bar{A}e^{i\sqrt{3}t_0}F + \mathrm{CC}\right\} + \frac{1}{3}\frac{\partial B}{\partial t_2}. \tag{2.34}$$

(F_3 does not pertain to this discussion.)

To examine the behavior of the front in the different parameter regimes, let us consider the real equation in $|A|$ corresponding to the complex equation (2.28), namely,

$$\frac{d|A|}{dt_2} = |A|(\mathrm{Re}(\chi) + |A|^2\mathrm{Re}(\beta)). \tag{2.35}$$

Perturbation amplitude $|A| = 0$ is a stationary solution for (2.35). Because $\mathrm{Re}(\chi) = 27/2$ is greater than zero, trajectories with an initial point near the origin in the complex A plane tend away from the origin. That is, in the absence of other equilibria, the amplitude blows up in (slow) time for ν slightly below the critical value $1/3$.

Equation (2.35) has a second equilibrium $|A| = \sqrt{-\mathrm{Re}(\chi)/\mathrm{Re}(\beta)}$ (a circle in the complex-A plane) if $\mathrm{Re}(\beta)$ is negative. A simple stability analysis of (2.35) shows that $d|A|/dt_2 < 0$ outside of the circle, and $d|A|/dt_2 > 0$ inside it. As a result, the limit cycle in the complex-A plane is asymptotically stable in this setting. A supercritical Hopf bifurcation occurs at $\nu = 1/3$. The nonlinear solution develops oscillations of magnitude $O(\epsilon)$ on the time scale $O(\epsilon^{-2})$. (See the expansion in (2.2).)

The quantity $\mathrm{Re}(\beta)$ is a quadratic function in σ with no roots at physical values of σ. For all $0 < \sigma < 1$, $\mathrm{Re}(\beta)$ is negative. The amplitude of the flat mode $A(t_2)e^{i\sqrt{3}t_0}$ approaches the limit cycle $|A| = \sqrt{-\mathrm{Re}(\chi)/\mathrm{Re}(\beta)}$. The nonlinear problem (1.8)–(1.11) develops oscillations, as detailed below.

3. Numerical method. We integrate numerically the exact problem as given by (1.8)–(1.11). In section 4, we compare the numerical solution with the asymptotics derived above. As was pointed out in [7], numerical solutions of (1.8)–(1.11) are very sensitive to the boundary condition (1.10). In order to obtain an alternative condition, we integrate (1.8) with respect to η from 0 to ∞. Subsequently applying conditions (1.9)–(1.11) results in the equation

$$\frac{d}{dt}\int_0^\infty u\,d\eta = -\nu f_t K(V). \tag{3.1}$$

We use (3.1) to replace (1.10) and adopt the Crank–Nicolson method for the numerical solution. The computation domain for η is $[0, 10]$ with $\delta t = \delta\eta = 0.025$. This produces a nonlinear system of m $(= 401)$ equations. In particular, in reference to (2.2), we introduce perturbation variables u^* and f^* defined by

$$u = e^{-\eta} + \epsilon u^*; \qquad f = t + \epsilon f^*. \tag{3.2}$$

Our discretization of condition (3.1) is

$$\begin{aligned}\Big(\int_0^\infty u^*\,d\eta\Big)\Big|_{t_k}^{t_{k+1}} = -\frac{\nu}{2\epsilon}[&(1+\epsilon f_t^*(t_{k+1}))K(1+\epsilon f_t^*(t_{k+1}))\\ &+(1+\epsilon f_t^*(t_k))K(1+\epsilon f_t^*(t_k))](t_{k+1}-t_k),\end{aligned} \tag{3.3}$$

where the integral on the left-hand side of (3.3) can be approximated by a composite trapezoidal rule.

We solve the nonlinear system of equations using Newton's method. The Jacobian matrix has the following sparse structure:

$$\begin{pmatrix} \# & \# & 0 & 0 & 0 & 0 & \dots & 0 \\ \# & \# & \# & \# & \# & \# & \dots & \# \\ \# & \# & \# & \# & 0 & 0 & \dots & 0 \\ \# & 0 & \# & \# & \# & 0 & \dots & 0 \\ \vdots & & & \ddots & \ddots & \ddots & & \vdots \\ \# & 0 & \dots & 0 & \# & \# & \# & 0 \\ \# & 0 & \dots & 0 & 0 & \# & \# & \# \\ \# & 0 & \dots & 0 & 0 & 0 & \# & \# \end{pmatrix}, \tag{3.4}$$

where # denotes a nonzero element. The matrix can be efficiently inverted by using Gaussian elimination with backward substitution.

The next section contains comparisons between numerical and asymptotic solutions. For the asymptotic solution, we integrate the ordinary differential equation (2.28) using a fourth-order Runge–Kutta method. As was pointed out in section 2, the Landau–Stuart equation (2.28) has circular limit cycles in the complex-A plane for all values of the kinetic parameter σ in the interval $0 < \sigma < 1$.

4. Comparison between asymptotics and numerics. To fix the idea, we first consider $\epsilon = 0.1$. The value of ν remains at the marginally unstable value $\nu_c - \epsilon^2$, as introduced in (2.1), so $\nu \approx 0.32\bar{3}$. We show in this section that this choice of ϵ corresponds to a mix of dynamics as σ varies. Subsequently, we both decrease and increase ϵ and discuss the impact on the front behavior. For the remainder of this paper, we take the initial condition $A(0) = 0.1$.

To start, take $\sigma = 0.48$ in the kinetic function (1.7). Figure 4.1 shows the numerical (solid line) and asymptotic (dashed line) values of front speed perturbation as a function of time t in the interval $0 \le t \le 60$. Specifically, for the numerical and asymptotic solutions we have graphed the quantities

$$v_n = f_t^* \quad \text{and} \quad v_a = A(t_2)e^{i\sqrt{3}t_0} + \frac{1}{2}B(t_1, t_2) + \epsilon\phi_2(t_0, t_1, t_2), \tag{4.1}$$

respectively, where f^* is defined in (3.2), and v_a contains the first three terms in the perturbation in (2.2).

Figure 4.1 shows that from $t = 0$ to about $t = 30$, the small front speed perturbation is linearly unstable, and its amplitude grows exponentially in time. As this amplitude becomes large, nonlinearity takes effect. At around $t = 30$, the front speed

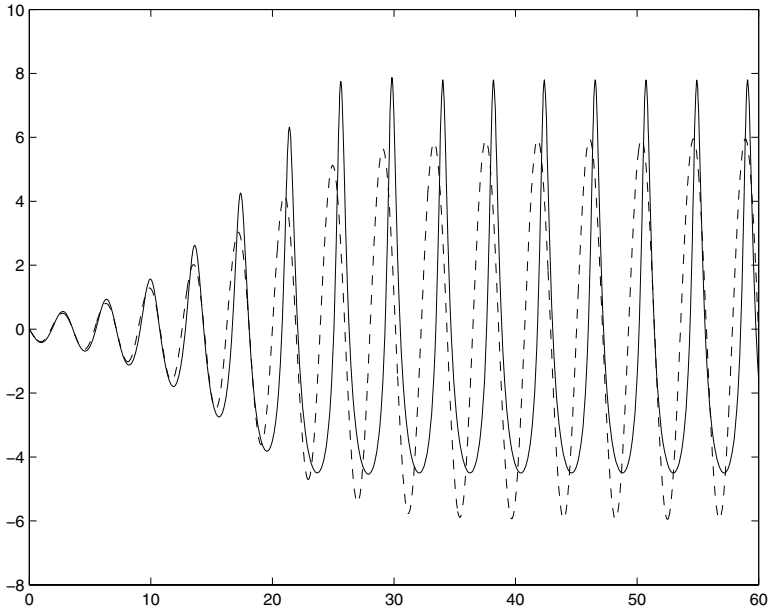


FIG. 4.1. *Velocity perturbation versus time: Comparison between numerical (solid line) and asymptotic (dashed line);* $\sigma = 0.48$, $\epsilon = 0.1$, $A(0) = 0.1$ *(*$\nu \approx \nu_c - \epsilon^2 = 1/3 - (0.1)^2 = 0.32\bar{3}$*).*

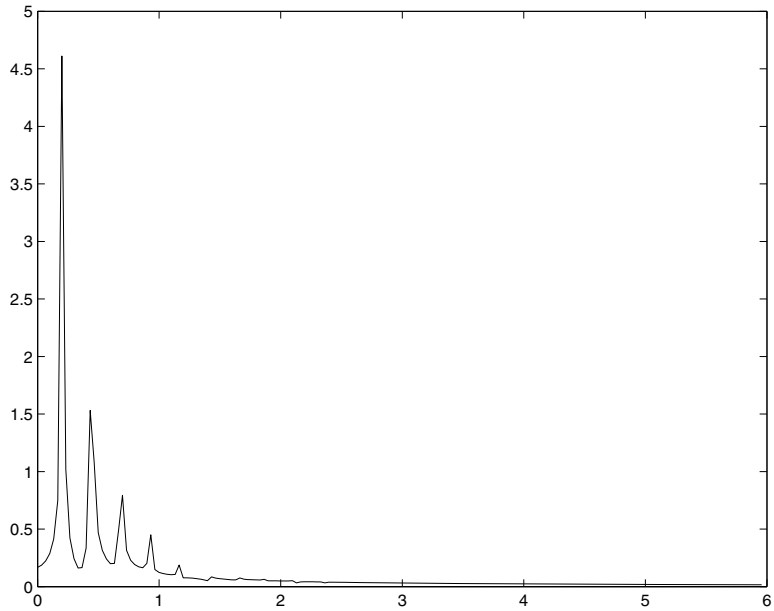


FIG. 4.2. *Fourier amplitude of the numerical steady-state velocity perturbation;* $\sigma = 0.48$, $\epsilon = 0.1$, $A(0) = 0.1$, $50 < t < 100$ *(*$\nu \approx \nu_c - \epsilon^2 = 1/3 - (0.1)^2 = 0.32\bar{3}$*).*

perturbation has reached steady oscillation. The asymptotic solution accurately captures the period in both transient behavior for $t = 0$ to 30 and the long-time behavior after $t = 30$. The amplitude and phase differ somewhat.

The weakly nonlinear approach describes well, by definition, marginally unstable large-time behaviors when a single modulated temporal mode of frequency $\sqrt{3}$ captures the dynamics. (See (2.2).) We have illustrated such a case in Figure 4.1. We then numerically calculated the velocity perturbation data $\{f_t^*(t_i)\}$ on the time interval $50 < t < 100$, using the parameter values as in Figure 4.1. The discrete Fourier transform of the data reveals the dominance of one mode. (See Figure 4.2.)

However, the subsequent modes do contribute to the solution as well. The second spike in Figure 4.2 is about 3/5 the height of the first, and the third is fully 1/2 the height of the second. Contributions of higher-order modes may explain some quantitative discrepancies between the numerical and asymptotic solutions in Figure 4.1.

Figure 4.3 summarizes the Fourier transformed velocity data for all physical values of σ $(0 < \sigma < 1)$. For each σ value and each frequency, the color indicates the corresponding amplitude, with the red end of the spectrum standing for larger numbers than the violet end, as the legend to the right of the figure illustrates. For roughly $0.3 < \sigma < 0.6$, the figure shows the dominance of the lowest-order mode, suggesting the appropriateness of the weakly nonlinear analysis in this range.

Notice, however, that at least four additional modes appear significant as well. Nevertheless, for σ in the interval approximately $(0.3, 0.6)$, the weakly nonlinear solution captures the gross features of the oscillation. (See, for example, Figure 4.1.)

With our choice of $\epsilon = 0.1$, Figure 4.3 shows that for σ greater than approximately 0.6, a single mode cannot be expected to capture the full dynamics of the solution. For example, for $\sigma = 0.85$, the asymptotic solution certainly will not exhibit a velocity perturbation with the very sharp peaks seen in the numerical solution in Figure 4.4.

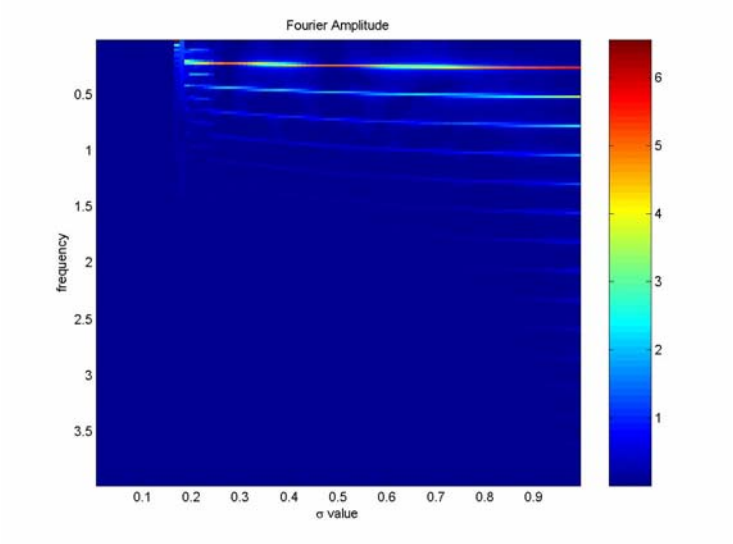


FIG. 4.3. *Amplitudes corresponding to each frequency of the Fourier transformed velocity perturbation data for the Arrhenius kinetics parameter* σ *in the interval* $(0,1)$*;* $\epsilon = 0.1$*,* $A(0) = 0.1$*,* $35 < t < 85$ *(*$\nu \approx \nu_c - \epsilon^2 = 1/3 - (0.1)^2 = 0.32\bar{3}$*).*

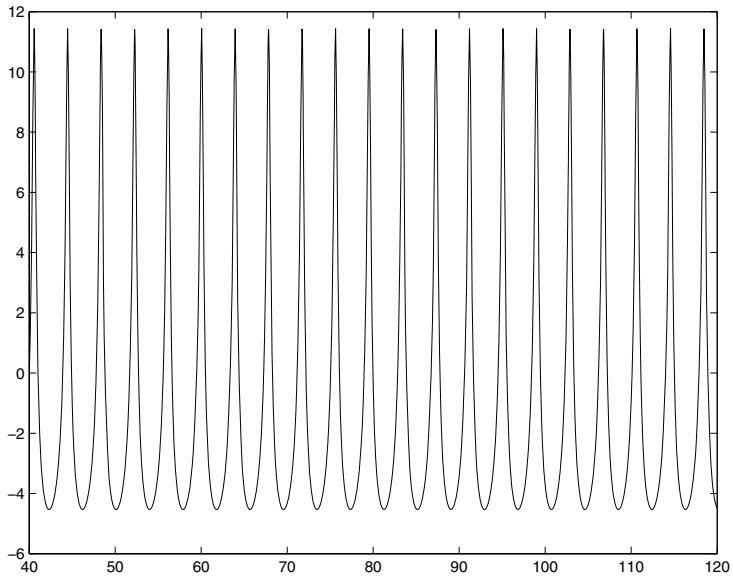


FIG. 4.4. *Velocity perturbation versus time: Numerical solution for* $\sigma = 0.85$*,* $\epsilon = 0.1$*,* $A(0) = 0.1$ *(*$\nu \approx \nu_c - \epsilon^2 = 1/3 - (0.1)^2 = 0.32\bar{3}$*).*

Further, Figure 4.3 shows that the Fourier spectrum has a complicated character for σ sufficiently small, starting with the emergence of a period-doubling solution for $\sigma \approx 0.25$. Naturally, the asymptotic solution captures neither the period-doubling solution nor the period-quadrupling computed for $\sigma = 0.22$ and $\sigma = 0.21$, respectively. (Numerical solutions in Figures 4.5 and 4.6 illustrate the dynamics.) Figure 4.3 reflects the breakdown of the numerical solution for σ less than approximately 0.15.

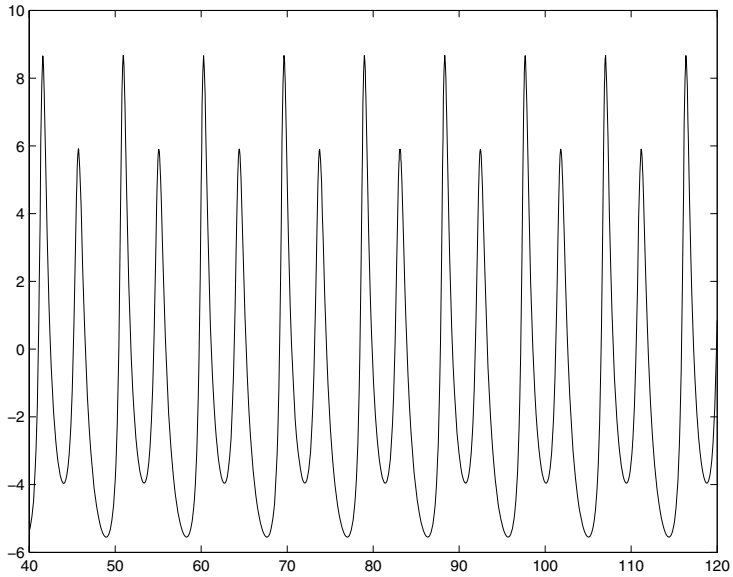


FIG. 4.5. *Velocity perturbation versus time: Period-doubling numerical solution for* $\sigma = 0.22$, $\epsilon = 0.1$, $A(0) = 0.1$ *(*$\nu \approx \nu_c - \epsilon^2 = 1/3 - (0.1)^2 = 0.32\bar{3}$*).*

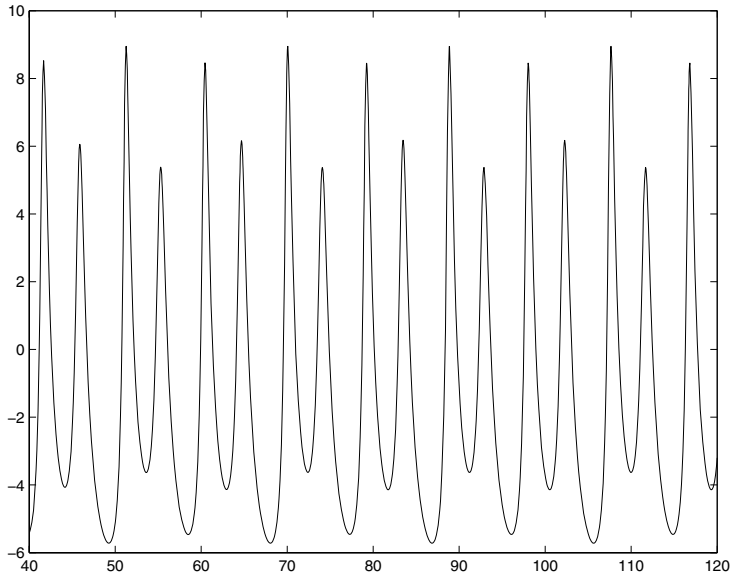


FIG. 4.6. *Velocity perturbation versus time: Period-quadrupling numerical solution for* $\sigma = 0.21$, $\epsilon = 0.1$, $A(0) = 0.1$ *(*$\nu \approx \nu_c - \epsilon^2 = 1/3 - (0.1)^2 = 0.32\bar{3}$*).*

The weakly nonlinear analysis of section 2 predicts periodic single-mode-dominant solutions for all physical values of σ $(0 < \sigma < 1)$ when $\nu = 1/3 - \epsilon^2$ is sufficiently close to the neutrally stable value $\nu_c = 1/3$. From numerical simulation with $\epsilon = 0.1$, the interval in which a single mode dominates has been identified via Figure 4.3 as a subinterval of $(0, 1)$, namely, $(0.3, 0.6)$. The corresponding asymptotic solution

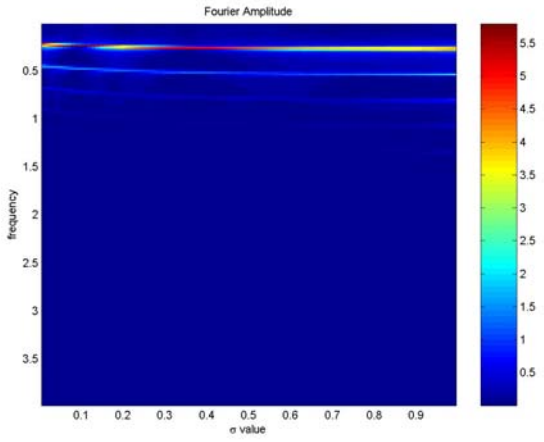


FIG. 4.7. *Amplitudes corresponding to each frequency of the Fourier transformed velocity perturbation data for the Arrhenius kinetics parameter σ in the interval $(0,1)$; $\epsilon = 0.06$, $A(0) = 0.1$, $35 < t < 85$ ($\nu \approx \nu_c - \epsilon^2 = 1/3 - (0.06)^2$).*

captures the numerical solution accurately even for ϵ as large as 0.1 when the dynamics associated with varying the parameter σ are not too complex. In particular, the interval $0.3 < \sigma < 0.6$ corresponds to good agreement. In what follows, we discuss the effects of decreasing and increasing ϵ.

We have seen that the approximate interval of $\sigma \geq 0.6$ in Figure 4.3 corresponds to sharply spiking solutions with many Fourier modes contributing. This far right interval moves farther and farther to the right as ϵ decreases. When ϵ drops to 0.06, the interval of spiking solutions disappears.

As we decrease ϵ, graphs analogous to Figure 4.3 also show the period-doubling region pushed farther and farther to the left along the σ axis. Similarly, the far left code-failure region moves farther to the left. When ϵ drops to 0.07, the period-doubling interval essentially disappears, and solutions can be computed even for extremely small σ values.

Figure 4.7 shows that for $\epsilon = 0.06$, one mode dominates strongly throughout the entire interval $0 < \sigma < 1$. The asymptotic and numerical solutions are consistent for *all* physical values of σ when $\nu = 1/3 - \epsilon^2$ if ϵ lies in the relatively small interval $0 < \epsilon < 0.06$.

Figure 4.7 also shows that only three higher-order modes appear to make slight additional contributions, fewer than for any value of σ illustrated in Figure 4.3 for $\epsilon = 0.1$. Therefore, as expected, the weakly nonlinear and numerical solutions agree more closely with ϵ reduced from 0.1 to 0.06. Figure 4.8 ($\epsilon = 0.06$) shows good agreement in period—as does Figure 4.1 when $\epsilon = 0.1$. Also, the phase, amplitude, and centerline agreement has improved considerably in Figure 4.8 for the decreased ϵ. In Figure 4.8, the asymptotic solution oscillates between about -6 and 6, while the numerical extends from -5 to 7, and their difference at the quasi-steady-state peaks is about 1. In Figure 4.1 for $\epsilon = 0.1$, the asymptotic solution also lies between -6 and 6, but the numerical solution varies between -4.5 to 7.9. The difference at the peaks is about 2.

We have discussed the impact of reducing ϵ from the value 0.1 used in Figures 4.1–4.6, which all pertain to the dynamics when $\nu = 1/3 - (0.1)^2$. If ϵ *increases* beyond 0.1,

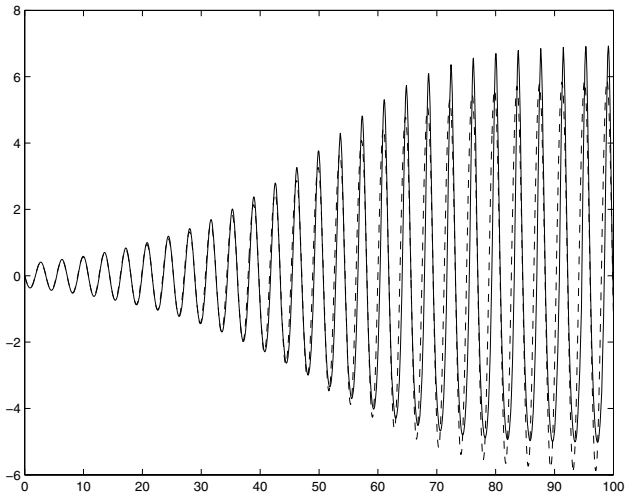


FIG. 4.8. *Velocity perturbation versus time: Comparison between numerical (solid line) and asymptotic (dashed line),* $\sigma = 0.48$, $\epsilon = 0.06$, $A(0) = 0.1$ *(*$\nu \approx \nu_c - \epsilon^2 = 1/3 - (0.06)^2 = 0.3297\overline{3}$*).*

then the far right σ interval in Figure 4.3, corresponding to sharply spiking solutions, has a left-hand endpoint that moves farther and farther to the *left*: The σ interval of solutions with sharp peaks expands.

Also, as we increase ϵ, graphs analogous to Figure 4.3 show the period-doubling region pushed farther and farther to the *right* along the σ axis, when compared with Figure 4.3. In addition, the far left code-failure region has a right-hand endpoint that moves farther and farther to the *right*: The σ interval on which the code breaks down expands.

When ϵ grows to 0.12, the σ zone in which one mode dominates strongly becomes extremely narrow. The asymptotic and numerical solutions agree well for σ in the approximate interval $(0.41, 0.42)$.

5. Conclusions and discussion. We have quantitatively compared a weakly nonlinear analysis and direct numerical integration for a solid combustion model. By definition, the weakly nonlinear approach is well suited to the study of marginally unstable large-time behaviors when the modulated most-unstable mode captures the dynamics.

For both the asymptotic and numerical methods, we examined nonuniform solutions corresponding to ν fixed within ϵ^2 of the stability boundary. When $\epsilon = 0.1$, for values of σ in the approximate interval $(0.3, 0.6)$, the weakly nonlinear analysis predicted accurately the transient and steady-state behaviors and particularly the period of oscillation. Beyond these special values of σ, the steady-state front propagation exhibited complicated nonlinear behaviors. We took the Fourier transforms of computational solutions to illustrate that higher-order modes play a significant role on σ intervals outside of $(0.3, 0.6)$ when $\epsilon = 0.1$.

For larger values of ϵ, the σ interval of applicability of the weakly nonlinear analysis shrinks. By contrast, when ϵ drops to approximately 0.06, the asymptotic and numerical solutions agree well for all physical values of σ.

Specifically, as ϵ increases from 0.06 (Figure 4.7) to $\epsilon = 0.1$ (Figure 4.3) and beyond, a period-doubling sequence develops in σ. As ϵ gets larger (thereby pushing

ν somewhat deeper into the instability region), the period-doubling bifurcation occurs at larger and larger values of σ, and the subsequent bifurcations occur as σ decreases. We note that this result, when viewed for a fixed value of σ, shows a period-doubling sequence in ν (as ϵ increases), which concurs with the dynamical scenarios described in the literature for experiments, as well as for simulations on reaction-diffusion and free-interface models.

In future work, we will suggest a hybrid expansion-perturbation technique for capturing more complex dynamics than those that have single-mode dominance. A more flexible general expansion as in [22] will assume that the temperature and interface position can be represented as a Fourier-like series that includes multiple temporal modes varying in fast time.

Acknowledgments. The authors would like to express their gratitude to the editor Stephen B. Margolis and to the referees for very insightful comments and suggestions. We also thank Yi Yang of the Institute of Applied Physics and Computational Mathematics in Beijing for helpful recommendations.

REFERENCES

[1] A. Bayliss and B. J. Matkowsky, *Two routes to chaos in condensed phase combustion*, SIAM J. Appl. Math., 50 (1990), pp. 437–459.
[2] A. F. Belyaev and L. D. Komkova, *Dependence of burning velocity of thermites on pressure*, Zh. Fiz. Khim., 24 (1950), pp. 1302–1311.
[3] M. R. Booty, S. B. Margolis, and B. J. Matkowsky, *Interaction of pulsating and spinning waves in condensed phase combustion*, SIAM J. Appl. Math., 46 (1986), pp. 801–843.
[4] I. Brailovsky and G. Sivashinsky, *Chaotic dynamics in solid fuel combustion*, Phys. D, 65 (1993), pp. 191–198.
[5] P. Dimitriou, J. Puszinski, and V. Hlavacek, *On the dynamics of equations describing gasless combustion*, Combust. Sci. Tech., 68 (1989), pp. 101–111.
[6] M. L. Frankel, L. K. Gross, and V. Roytburd, *Thermo-kinetically controlled pattern selection*, Interfaces Free Bound., 2 (2000), pp. 313–330.
[7] M. L. Frankel and V. Roytburd, *Dynamical portrait of a model of thermal instability: Cascades, chaos, reversed cascades, and infinite period bifurcations*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 4 (1994), pp. 579–593.
[8] M. Frankel, V. Roytburd, and G. Sivashinsky, *A sequence of period doublings and chaotic pulsations in a free boundary problem modeling thermal instabilities*, SIAM J. Appl. Math, 54 (1994), pp. 1101–1112.
[9] M. L. Frankel, V. Roytburd, and G. Sivashinsky, *Complex dynamics generated by a sharp interface model of self-propagating high-temperature synthesis*, Combust. Theory Model., 2 (1998), pp. 1–18.
[10] M. Garbey, H. G. Kaper, G. K. Leaf, and B. J. Matkowsky, *Quasi-periodic waves and the transfer of stability in condensed-phase surface combustion*, SIAM J. Appl. Math., 52 (1992), pp. 384–395.
[11] L. K. Gross, *Weakly Nonlinear Dynamics of Interface Propagation*, Ph.D. thesis, Rensselaer Polytechnic Institute, Troy, NY, 1997.
[12] L. K. Gross, *Weakly nonlinear dynamics of interface propagation*, Stud. Appl. Math., 108 (2002), pp. 323–350.
[13] S. B. Margolis, *Transition to nonsteady deflagration in gasless combustion*, Progr. Energy Combust. Sci., 17 (1991), pp. 135–162.
[14] B. J. Matkowsky and G. I. Sivashinsky, *Propagation of a pulsating reaction front in solid fuel combustion*, SIAM J. Appl. Math., 35 (1978), pp. 465–478.
[15] A. G. Merzhanov, *SHS processes: Combustion theory and practice*, Arch. Combust., 1 (1981), pp. 23–48.
[16] A. G. Merzhanov, A. K. Filonenko, and I. P. Borovinskaya, *New phenomena in combustion of condensed systems*, Soviet Phys. Dokl., 208 (1973), pp. 892–894.
[17] Z. A. Munir and U. Anselmi-Tamburini, *Self-propagating exothermic reactions: The synthesis of high-temperature materials by combustion*, Mat. Sci. Rep., 3 (1989), pp. 277–365.

[18] K. G. Shkadinsky, B. I. Khaikin, and A. G. Merzhanov, *Propagation of a pulsating exothermic reaction front in the condensed phase*, Combust. Expl. Shock Waves, 7 (1971), pp. 15–22.
[19] G. I. Sivashinsky, *The structure of Bunsen flames*, J. Chem. Phys., 62 (1975), pp. 638–643.
[20] A. Varma, A. S. Rogachev, A. S. Mukasyan, and S. Huang, *Combustion synthesis of advanced materials: Principles and applications*, Adv. Chem. Engrg., 24 (1998), pp. 79–226.
[21] J. Yu and L. K. Gross, *The onset of linear instabilities in a solid combustion model*, Stud. Appl. Math., 107 (2001), pp. 81–101.
[22] J. Yu and Y. Yang, *Evolution of small periodic disturbances into roll waves in channel flow with internal dissipation*, Stud. Appl. Math., 111 (2003), pp. 1–27.

SIAM J. APPL. MATH.
Vol. 65, No. 5, pp. 1726–1747

NÉEL WALLS IN LOW ANISOTROPY SYMMETRIC DOUBLE LAYERS*

CARLOS J. GARCIA-CERVERA†

Abstract. A new model for the study of one-dimensional walls in magnetic multilayers is presented. We obtain the optimal scaling of this energy functional for low anisotropy double layers with magnetic layers of equal thickness. We prove that the optimal scaling may be attained by opposing Néel walls. We obtain the core length of the Néel wall and a detailed description of its structure. We illustrate our findings numerically.

Key words. micromagnetics, Landau–Lifshitz, Γ-limit, Néel walls

AMS subject classifications. 34B15, 34B60, 34E05, 49S05, 49K30, 65D99

DOI. 10.1137/S003613990343776X

1. Introduction. A magnetic multilayer consists of two or more magnetic films, separated by a layer of nonmagnetic material (see Figure 1). Each layer may be of a different thickness. Multilayers seem to have good permanent magnet properties, in particular a high coercive field and approximately rectangular hysteresis loop [17]. For that reason multilayers are an integral part of magnetic memories (MRAMs) and have been one of the most important applications of ferromagnetic thin films in the past few years.

The magnetization distribution in a ferromagnetic material is described by the micromagnetics model, introduced by Landau and Lifshitz [12]. In nondimensional variables, the Landau–Lifshitz energy functional for a sample occupying a volume V is

$$F[\mathbf{m}] = \frac{q}{2}\int_V \Phi(\mathbf{m})\,dx + \frac{1}{2}\int_V |\nabla \mathbf{m}|^2\,dx + \frac{1}{2}\int_{\mathbb{R}^3} |\nabla \eta|^2\,dx. \tag{1.1}$$

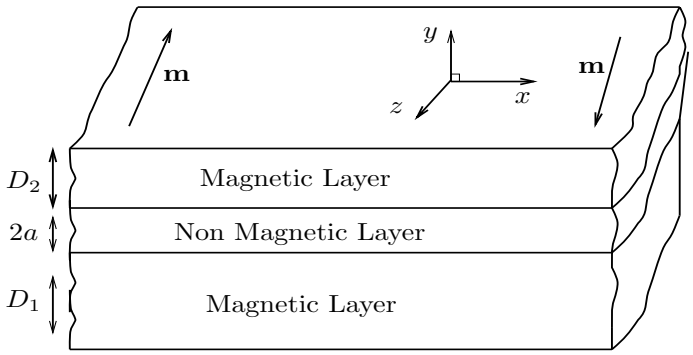


FIG. 1. *One-dimensional wall setting in a multilayer.*

*Received by the editors November 18, 2003; accepted for publication (in revised form) January 3, 2005; published electronically July 26, 2005.
http://www.siam.org/journals/siap/65-5/43776.html
†Mathematics Department, University of California, Santa Barbara, CA 93106 (cgarcia@math.ucsb.edu, http://www.math.ucsb.edu/∼cgarcia).

In (1.1), $|\mathbf{m}| = 1$ in V, and $\mathbf{m} = 0$ outside V. The three terms in (1.1) are anisotropy, exchange, and stray field energy, respectively. The parameter q is the quality factor, defined as $q = K_u/(\mu_0 M_s^2)$, where K_u is the crystalline anisotropy constant, M_s is the saturation magnetization, and μ_0 is the permeability of vacuum ($\mu_0 = 4\pi \times 10^{-7} N/A^2$). In (1.1), lengths are measured in units of the exchange length, $l = \sqrt{C_{ex}/(\mu_0 M_s^2)}$, where C_{ex} is the exchange constant. The energy is measured in units of $e = \sqrt{\mu_0 M_s^2 C_{ex}}$.

The stray field is $\mathbf{h}_s = -\nabla\eta$, where η is obtained by solving the equation

$$\text{div } (-\nabla\eta + \mathbf{m}) = 0 \quad \text{in } \mathbb{R}^3 \tag{1.2}$$

in the sense of distributions. The solution has the explicit form

$$\eta = \nabla N * \mathbf{m}, \tag{1.3}$$

where $N(\mathbf{x}) = -\frac{1}{4\pi}\frac{1}{|\mathbf{x}|}$ is the Newtonian potential.

For physical parameters typical of Permalloy ($C_{ex} = 1.3 \times 10^{-11} J/m$, $K_u = 5 \times 10^2 J/m^3$, $M_s = 8 \times 10^5 A/m$), the quality factor is $q \approx 6.21 \times 10^{-3}$. Thus it is physically relevant to consider the low anisotropy limit $q \to 0$, which is the situation considered in this article.

Functional (1.1) has been the focus of recent attention in the mathematical community, and the energy landscape for a single magnetic layer is now fairly well understood [9, 4, 5, 6, 8, 7, 18].

Due to the nonlocal nature of the magnetostatic interactions, the behavior of the magnetization distribution in double layers is very different from the single layer case. The magnetization patterns correspond to local minimizers of the Landau–Lifshitz energy. In a double layer, a pattern that would otherwise be energetically unfavorable for a single layer can be permitted by producing a pattern in the other layer which will cause the necessary field cancellations. With this compensating mechanism, new phenomena occur that are intrinsic to double layers.

The domain structure in magnetic films can be rather complicated [10, 6, 18]. In order to understand the structure in double films, we start by analyzing one-dimensional profiles, which will be the building blocks of more complicated structures. We are interested in both the structure and the energy of the minimizers. We are mainly interested in the scaling of the energy in terms of q as $q \to 0$, since all other parameters are kept fixed. To determine the energy of the minimizers, we consider an appropriate function in the admissible class, which provides us with an upper bound for the energy in terms of q. Subsequently we find a lower bound for the energy with the same scaling in q. The upper bound and the matching lower bound ensure that the energy is optimal, at least in terms of scaling.

In this article, we focus on the study of Néel walls in multilayers formed by two layers of equal thickness. Throughout this article we will refer to these multilayers as symmetric double layers. A description of the Néel wall in a single ferromagnetic layer was presented in [7], where the following energy functional, due originally to Aharoni [1], was analyzed:

$$\begin{aligned} F^A_{q,\delta}[\mathbf{m}] = {} & \frac{q}{2}\int_{\mathbb{R}} \left(m_1^2 + m_2^2\right) + \frac{1}{2}\int_{\mathbb{R}} |\mathbf{m}'|^2 + \frac{1}{2}\int_{\mathbb{R}} \left(m_1^2 - m_1\,(\Gamma_\delta * m_1)\right) \\ & + \frac{1}{2}\int_{\mathbb{R}} m_2\,(\Gamma_\delta * m_2). \end{aligned} \tag{1.4}$$

The set of admissible functions is

$$\mathcal{A} = \Big\{ \mathbf{m} = (m_1, m_2, m_3) | m_1, m_2 \in H^1(\mathbb{R}),\ m_3' \in L^2(\mathbb{R}),\ |\mathbf{m}| = 1 \text{ a.e.},\ \mathbf{m} \to \pm \mathbf{e}_3 \text{ as } x \to \pm\infty \Big\}. \tag{1.5}$$

In (1.4), δ represents the (rescaled) thickness of the sample, and

$$\Gamma_\delta(x) = \frac{1}{4\pi\delta} \log\left(1 + \frac{4\delta^2}{x^2}\right). \tag{1.6}$$

Functional (1.4) is derived directly from (1.1) by considering magnetization profiles that depend only on the x-variable. Functional (1.4) provides an accurate description of the minimizers for thin films ($\delta \ll 1$), since the dependence on the thickness variable becomes negligible as $\delta \to 0$ [9, 6, 8, 7, 18, 11, 3]. For thicker films, lower energy can be achieved with higher-dimensional structures [10, 18, 16].

For the study of Néel walls, we consider magnetization profiles such that $m_2 = 0$. The optimal energy scaling for a Néel wall in a single layer was obtained in [7]. In particular, it was proved that for a given $\delta > 0$ there exist positive constants c_0 and C_0 such that

$$\frac{c_0}{\log \frac{1}{q}} \le \inf_{\mathbf{m} \in \mathcal{A}, m_2 = 0} F^A_{q,\delta}[\mathbf{m}] \le \frac{C_0}{\log \frac{1}{q}} \tag{1.7}$$

as $q \to 0$. Moreover, it was shown that the Néel wall has a long logarithmic tail, which extends the stray field interactions to great distances.

In double layers the structure of Néel walls can be very different [14, 19, 20]. In this article we prove that in low anisotropy symmetric double layers, the logarithmic tail of the Néel wall disappears. The stray field becomes an exchange-type energy, and the wall becomes more localized and similar to the Landau–Lifshitz wall [12].

Considering a double layer as depicted in Figure 1, we have derived the following one-dimensional model for the study of magnetic walls in double layers:

$$\begin{aligned} G_{q,\alpha,\delta_1,\delta_2}[\mathbf{m}_1, \mathbf{m}_2] = {} & F_{q,\delta_1}[\mathbf{m}_1] + \frac{\delta_2}{\delta_1} F_{q,\delta_2}[\mathbf{m}_2] \\ & + \frac{\delta_2}{2} \int_{\mathbb{R}} u_1 \left(u_2 * \Theta_{\alpha,\delta_1,\delta_2}\right) - v_1 \left(v_2 * \Theta_{\alpha,\delta_1,\delta_2}\right) dx \\ & + \frac{\delta_2}{2} \int_{\mathbb{R}} v_1 \left(u_2 * \Psi_{\alpha,\delta_1,\delta_2}\right) - u_1 \left(v_2 * \Psi_{\alpha,\delta_1,\delta_2}\right) dx, \end{aligned} \tag{1.8}$$

where $\mathbf{m}_1 = (u_1, v_1, w_1)$ and $\mathbf{m}_2 = (u_2, v_2, w_2)$ represent the magnetization inside each layer, F_{q,δ_1} and F_{q,δ_2} are as in (1.4), and

$$\begin{aligned} \Theta_{\alpha,\delta_1,\delta_2}(x) &= \frac{1}{2\delta_1\delta_2\pi} \left(\log\left(\frac{x^2 + (2\alpha + \delta_1)^2}{x^2 + (2\alpha + \delta_1 + \delta_2)^2}\right) - \log\left(\frac{x^2 + 4\alpha^2}{x^2 + (2\alpha + \delta_2)^2}\right)\right), \\ \Psi_{\alpha,\delta_1,\delta_2}(x) &= \frac{1}{\delta_1\delta_2\pi} \Bigg(\arctan\left(\frac{2\alpha + \delta_1}{x - s}\right) - \arctan\left(\frac{2\alpha}{x - s}\right) \\ &\qquad - \arctan\left(\frac{2\alpha + \delta_1 + \delta_2}{x - s}\right) + \arctan\left(\frac{2\alpha + \delta_2}{x - s}\right)\Bigg). \end{aligned} \tag{1.9}$$

We have not been able to find this model in the literature, and therefore a complete derivation of this model is given in Appendix A. For the study of Néel walls, we assume $v_1 = v_2 = 0$. For a symmetric double layer, $\delta_1 = \delta_2 = \delta$.

This article is organized as follows. In section 2, we obtain the optimal energy scaling for Néel walls in symmetric double layers. In particular, we show that for fixed $\delta > 0$ and $\alpha > 0$ there exists a constant $C_0 > 0$ such that

$$4\sqrt{q} \le \inf_{\mathbf{m}\in\mathcal{A}} G_{q,\alpha,\delta,\delta}[\mathbf{m}] \le C_0\sqrt{q}. \tag{1.10}$$

The upper bound is obtained considering Néel walls in the double layer.

A more detailed analysis is carried out in section 3, where we prove that for a family of minimizers of the Néel wall functional (2.7), $\{\mathbf{m}_q\}_{\{q>0\}}$,

$$\lim_{q\to 0} \frac{1}{\sqrt{q}} G_{q,\alpha,\delta,\delta}[\mathbf{m}_q] = \min_{\mathbf{m}\in\mathcal{A}, m_2=0} \int_{\mathbb{R}} m_1^2\,dx + \frac{\delta\,(\delta+3\alpha)}{3}\int_{\mathbb{R}} (m_1')^2\,dx + \int_{\mathbb{R}} |\mathbf{m}'|^2\,dx. \tag{1.11}$$

We also prove that, given a family of minimizers $\{\mathbf{m}_q\}_{\{q>0\}}$, we can extract a sequence (not relabeled) such that the rescaled family $\{\mathbf{m}_q(x/\sqrt{q})\}_{\{q>0\}}$ converges strongly in $H^1(\mathbb{R})$. We interpret this in the Γ-limit sense of an appropriately scaled family of functionals. The limiting profile is studied in section 4 using a formal asymptotic expansion.

In section 5 we illustrate all of our findings numerically. To this end, we have implemented a modified Newton method for energy minimization. Finally, a detailed derivation of the model used in this article is presented in Appendix A.

2. Optimal scaling: Opposing Néel walls. Since the stray field energy is nonnegative, the Landau–Lifshitz wall profile always provides us with a lower bound for the energy:

$$G_{q,\alpha,\delta}[\mathbf{m}_1,\mathbf{m}_2] \ge \min_{\mathbf{m}_1,\mathbf{m}_2\in\mathcal{A}} \widetilde{F}_q[\mathbf{m}_1,\mathbf{m}_2], \tag{2.1}$$

where

$$\widetilde{F}_q[\mathbf{m}_1,\mathbf{m}_2] = \frac{1}{2}\sum_{j=1}^{2}\left(q\int_{\mathbb{R}}(u_j^2+v_j^2)\,dx + \int_{\mathbb{R}}|\mathbf{m}_j'|^2\,dx\right). \tag{2.2}$$

Since $\mathbf{m}_1$ and $\mathbf{m}_2$ are decoupled in (2.2), the minimum will be achieved for $\mathbf{m}_1 = \mathbf{m}_2 = (u,v,w)$. Moreover, we can assume that either $u = 0$ or $v = 0$. Otherwise, following Lemma 4 in [7], consider $\widetilde{\mathbf{m}} = (\sqrt{u^2+v^2}, 0, w)$. Then

$$\begin{aligned}\widetilde{F}_q[\widetilde{\mathbf{m}},\widetilde{\mathbf{m}}] &= q\int_{\mathbb{R}}(u^2+v^2) + \int_{\mathbb{R}}\left(\frac{(uu'+vv')^2}{u^2+v^2} + (w')^2\right)\\ &= q\int_{\mathbb{R}}(u^2+v^2) + \int_{\mathbb{R}}|\mathbf{m}'|^2 - \int_{\mathbb{R}}\frac{(uv'-vu')^2}{u^2+v^2} \le \widetilde{F}_q[\mathbf{m},\mathbf{m}].\end{aligned} \tag{2.3}$$

Therefore, a lower bound for (1.8) in a symmetric double layer is obtained by minimizing

$$\min_{\mathbf{m}\in\mathcal{A}, m_2=0} \widetilde{F}_q[\mathbf{m},\mathbf{m}]. \tag{2.4}$$

This is the minimization problem studied by Landau and Lifshitz in [12]. The minimizer is

$$\mathbf{m} = (\operatorname{sech}(\sqrt{q}x), 0, \tanh(\sqrt{q}x)), \tag{2.5}$$

and the minimum energy is

$$q\int_{\mathbb{R}} u^2\,dx + \int_{\mathbb{R}} |\mathbf{m}'|^2\,dx = 4\sqrt{q}. \tag{2.6}$$

In a single layer, this is not optimal for a Néel wall, as proved in [7]. However, we can prove that, due to stray field cancellations, this energy scaling is indeed optimal in a symmetric double layer.

To obtain a matching upper bound for the energy, we consider Néel walls in symmetric double layers and study the functional

$$\begin{aligned} G_{q,\alpha,\delta}[\mathbf{m}_1,\mathbf{m}_2] = {} & \frac{q}{2}\int_{\mathbb{R}} u_1^2 + \frac{1}{2}\int_{\mathbb{R}} |\mathbf{m}_1'|^2 + \frac{1}{2}\int_{\mathbb{R}} \left(u_1^2 - u_1\,(\Gamma_\delta * u_1)\right) + \frac{q}{2}\int_{\mathbb{R}} u_2^2 \\ & + \frac{1}{2}\int_{\mathbb{R}} |\mathbf{m}_2'|^2 + \frac{1}{2}\int_{\mathbb{R}} \left(u_2^2 - u_2\,(\Gamma_\delta * u_2)\right) + \frac{\delta}{2}\int_{\mathbb{R}} u_1\,(u_2 * \Theta_{\alpha,\delta}). \end{aligned} \tag{2.7}$$

In (2.13), we have renamed the energy functional $G_{q,\alpha,\delta} \equiv G_{q,\alpha,\delta,\delta}$, and the convolution kernel $\Theta_{\alpha,\delta} \equiv \Theta_{\alpha,\delta,\delta}$, in view of definitions (1.8) and (1.9), respectively. In Fourier space,

$$\begin{aligned} & G_{q,\alpha,\delta}[\mathbf{m}_1,\mathbf{m}_2] = \frac{q}{2}\int_{\mathbb{R}} |\widehat{u}_1|^2\,d\xi + \frac{1}{2}\int_{\mathbb{R}} 4\pi^2\xi^2|\widehat{\mathbf{m}}_1|^2\,d\xi + \frac{1}{2}\int_{\mathbb{R}} |\widehat{u}_1|^2\left(1-\widehat{\Gamma}_\delta(\xi)\right)d\xi \\ & + \frac{q}{2}\int_{\mathbb{R}} |\widehat{u}_2|^2\,d\xi + \frac{1}{2}\int_{\mathbb{R}} 4\pi^2\xi^2|\widehat{\mathbf{m}}_2|^2\,d\xi + \frac{1}{2}\int_{\mathbb{R}} |\widehat{u}_2|^2\left(1-\widehat{\Gamma}_\delta(\xi)\right)d\xi + \frac{\delta}{2}\Re\int_{\mathbb{R}} \widehat{u}_1\overline{\widehat{u}}_2\widehat{\Theta}_{a,\delta}(\xi)\,d\xi. \end{aligned} \tag{2.8}$$

In the following lemma, we prove that in a symmetric double layer the minimum of (2.7) is achieved by opposing Néel walls; i.e., $u_1 = -u_2$.

LEMMA 2.1. *Consider* $\mathbf{m}_1, \mathbf{m}_2 \in \mathcal{A}$, *where* $\mathbf{m}_1 = (u_1, 0, w_1)$ *and* $\mathbf{m}_2 = (u_2, 0, w_2)$. *Define* $\widetilde{\mathbf{m}}_1 = (-u_1, 0, w_1)$ *and* $\widetilde{\mathbf{m}}_2 = (-u_2, 0, w_2)$. *Then, either*

$$G_{q,\alpha,\delta}[\mathbf{m}_1, \widetilde{\mathbf{m}}_1] \le G_{q,\alpha,\delta}[\mathbf{m}_1, \mathbf{m}_2] \tag{2.9}$$

or

$$G_{q,\alpha,\delta}[\mathbf{m}_2, \widetilde{\mathbf{m}}_2] \le G_{q,\alpha,\delta}[\mathbf{m}_1, \mathbf{m}_2]. \tag{2.10}$$

Proof. We can rewrite the Fourier representation (2.8) as

$$\begin{aligned} G_{q,\alpha,\delta}[\mathbf{m}_1,\mathbf{m}_2] = {} & \frac{q}{2}\int_{\mathbb{R}} |\widehat{u}_1|^2\,d\xi + \frac{1}{2}\int_{\mathbb{R}} 4\pi^2\xi^2|\widehat{\mathbf{m}}_1|^2\,d\xi \\ & + \frac{1}{2}\int_{\mathbb{R}} |\widehat{u}_1|^2\left(1-\widehat{\Gamma}_\delta(\xi) - \frac{\delta}{2}\widehat{\Theta}_{\alpha,\delta}(\xi)\right)d\xi + \frac{q}{2}\int_{\mathbb{R}} |\widehat{u}_2|^2\,d\xi + \frac{1}{2}\int_{\mathbb{R}} 4\pi^2\xi^2|\widehat{\mathbf{m}}_2|^2\,d\xi \\ & + \frac{1}{2}\int_{\mathbb{R}} |\widehat{u}_2|^2\left(1-\widehat{\Gamma}_\delta(\xi) - \frac{\delta}{2}\widehat{\Theta}_{\alpha,\delta}(\xi)\right)d\xi + \frac{\delta}{4}\int_{\mathbb{R}} |\widehat{u}_1 + \widehat{u}_2|^2\widehat{\Theta}_{a,\delta}(\xi)\,d\xi. \end{aligned} \tag{2.11}$$

Note that

$$\int_{\mathbb{R}} |\widehat{u}_2|^2 \left(1 - \widehat{\Gamma}_\delta(\xi) - \frac{\delta}{2}\widehat{\Theta}_{\alpha,\delta}(\xi)\right) d\xi \tag{2.12}$$

is the stray field energy that corresponds to a symmetric double layer where $u_1 = -u_2$, and therefore it is nonnegative. Thus, given $(\mathbf{m}_1, \mathbf{m}_2)$, since all the terms in (2.11) are nonnegative, we can always lower the total energy by selecting $(\mathbf{m}_1, \widetilde{\mathbf{m}}_1)$ or $(\mathbf{m}_2, \widetilde{\mathbf{m}}_2)$. □

In view of this lemma, we need to consider only opposing Néel walls. Thus, we need to study functional

$$\begin{aligned}\widetilde{G}_{q,\alpha,\delta}[\mathbf{m}] = q\int_{\mathbb{R}} u^2\,dx + \int_{\mathbb{R}} |\mathbf{m}'|^2\,dx + \int_{\mathbb{R}} u^2\,dx - \int_{\mathbb{R}} u\,(\Gamma_\delta * u)\,dx \\ -\frac{\delta}{2}\int_{\mathbb{R}} u\,(u * \Theta_{\alpha,\delta})\,dx,\end{aligned} \tag{2.13}$$

which can be written in Fourier space as

$$\widetilde{G}_{q,\alpha,\delta}[\mathbf{m}] = q\int_{\mathbb{R}} |\widehat{u}|^2\,d\xi + \int_{\mathbb{R}} 4\pi^2\xi^2|\widehat{\mathbf{m}}|^2\,d\xi + \int_{\mathbb{R}} |\widehat{u}|^2 \left(1 - \widehat{\Gamma}_\delta(\xi) - \frac{\delta}{2}\widehat{\Theta}_{\alpha,\delta}(\xi)\right) d\xi. \tag{2.14}$$

The lower semicontinuity and existence of minimizers of functional (2.13) follow from Lemmas 1, 2, and 3 in [7]. The main difficulty in establishing the existence of minimizers lies in the fact that functional (2.13) is translation invariant. This problem is resolved in [7] by considering a translation of $\mathbf{m}$ such that $\mathbf{m}(0) = (1, 0, 0)$. The result then follows from the Sobolev embedding and the Rellich compactness theorem [21].

As a consequence of Lemma 2.1,

$$\inf_{\mathbf{m}_1,\mathbf{m}_2\in\mathcal{A}} G_{q,\alpha,\delta}[\mathbf{m}_1, \mathbf{m}_2] = \inf_{\mathbf{m}\in\mathcal{A}} \widetilde{G}_{q,\alpha,\delta}[\mathbf{m}]. \tag{2.15}$$

Thus, the existence of minimizers for functional (2.7) is established. To simplify notation, in what follows we will drop the tilde from functional (2.13).

The matching upper bound that we need can be obtained by considering the test function $\mathbf{m} = (\operatorname{sech}(\sqrt{q}x), 0, \tanh(\sqrt{q}x))$, which is the Landau–Lifshitz wall. The Fourier transform of $u(x) = \operatorname{sech}(\sqrt{q}x)$ can be obtained by residues [2]:

$$\widehat{u}(\xi) = \frac{\pi}{\sqrt{q}}\operatorname{sech}\left(\frac{\pi^2\xi}{\sqrt{q}}\right). \tag{2.16}$$

The stray field energy of this profile is

$$\begin{aligned} E_s &= \int_{\mathbb{R}} \widehat{u}^2 \left(1 - \frac{1 - e^{-2\pi\delta|\xi|}}{2\pi\delta|\xi|} - \frac{1}{2}e^{-4\pi a|\xi|}\frac{(1 - e^{-2\pi\delta|\xi|})^2}{2\pi\delta|\xi|}\right) d\xi \\ &= \frac{\pi^2}{q}\int_{\mathbb{R}} \operatorname{sech}^2\left(\frac{\pi^2\xi}{\sqrt{q}}\right)\left(1 - \frac{1 - e^{-2\pi\delta|\xi|}}{2\pi\delta|\xi|} - \frac{1}{2}e^{-4\pi\alpha|\xi|}\frac{(1 - e^{-2\pi\delta|\xi|})^2}{2\pi\delta|\xi|}\right) d\xi \\ &= \frac{1}{\sqrt{q}}\int_{\mathbb{R}} \operatorname{sech}^2(\xi)\left(1 - \frac{1 - e^{-2\delta|\xi|\sqrt{q}/\pi}}{2\delta|\xi|\sqrt{q}/\pi} - \frac{1}{2}e^{-4\alpha|\xi|\sqrt{q}/\pi}\frac{(1 - e^{-2\delta|\xi|\sqrt{q}/\pi})^2}{2\delta|\xi|\sqrt{q}/\pi}\right) d\xi \\ &= \frac{4\delta\sqrt{q}}{\pi^2}\left(\frac{\delta}{3} + \alpha\right)\int_{\mathbb{R}} |\xi|^2\operatorname{sech}^2(\xi)\,d\xi + O(q) = \frac{2\delta\sqrt{q}}{3}\left(\frac{\delta}{3} + \alpha\right) + O(q). \end{aligned} \tag{2.17}$$

Therefore $\exists q_0 > 0$ such that

$$E_s \leq \frac{4\delta\sqrt{q}}{3}\left(\frac{\delta}{3}+\alpha\right) \quad \forall q \leq q_0. \tag{2.18}$$

The total energy is therefore

$$\widetilde{G}_{q,\alpha,\delta}[\mathbf{m}] = \left\{4+\frac{2\delta}{3}\left(\frac{\delta}{3}+\alpha\right)\right\}\sqrt{q}+O(q) \leq 4\left\{1+\frac{\delta}{3}\left(\frac{\delta}{3}+\alpha\right)\right\}\sqrt{q} \quad \forall q \leq q_0. \tag{2.19}$$

We collect all this in the following theorem.

THEOREM 2.2. *Consider the one-dimensional wall energy functional for a symmetric double layer* (1.8). *Given* $\alpha > 0$ *and* $\delta > 0$ *fixed,* $\exists\ q_0 > 0$ *such that*

$$4\sqrt{q} \leq \min_{\mathbf{m}_1,\mathbf{m}_2\in\mathcal{A}} G_{q,\alpha,\delta,\delta}[\mathbf{m}_1,\mathbf{m}_2] \leq 4\left\{1+\frac{\delta}{3}\left(\frac{\delta}{3}+\alpha\right)\right\}\sqrt{q} \quad \forall q \leq q_0. \tag{2.20}$$

An estimate of the value of q_0 is presented in Appendix B.

3. Néel walls: Limiting behavior. In this section we study the structure of Néel walls in symmetric double layers and obtain the limiting behavior of any sequence of minimizers of the Néel wall functional (2.13). This is the content of the following theorem.

THEOREM 3.1. *Given* $\alpha > 0$ *and* $\delta > 0$, *consider* $\{\mathbf{m}_q\}_{\{q>0\}} \subset \mathcal{A}_N = \{\mathbf{m} = (m_1, 0, m_3) \in \mathcal{A}\}$ *such that*

$$G_{q,\alpha,\delta}[\mathbf{m}_q] \leq C\sqrt{q}, \tag{3.1}$$

and define $\widetilde{\mathbf{m}}_q(x) = \mathbf{m}_q(\frac{x}{\sqrt{q}})$. *There exists a subsequence of* $\{\mathbf{m}_q\}_{\{q>0\}}$ *(not relabeled) such that the following two statements hold:*

(i)

$$\lim_{q\to 0}\frac{1}{\sqrt{q}}G_{q,\alpha,\delta}[\mathbf{m}_q] = \min_{\mathbf{m}\in\mathcal{A}_N} F_{\alpha,\delta}[\mathbf{m}], \tag{3.2}$$

where

$$F_{\alpha,\delta}[\mathbf{m}] = \int_{\mathbb{R}} m_1^2\,dx + \left(1+\delta\left(\frac{1}{3}\delta+\alpha\right)\right)\int_{\mathbb{R}}(m_1')^2\,dx + \int_{\mathbb{R}}(m_3')^2\,dx. \tag{3.3}$$

(ii) *The subsequence converges strongly in* $\mathcal{A}_N$ *to* $\mathbf{n} \in \mathcal{A}_N$ *such that*

$$F_{\alpha,\delta}[\mathbf{n}] = \min_{\mathbf{m}\in\mathcal{A}_N} F_{\alpha,\delta}[\mathbf{m}]. \tag{3.4}$$

Proof. Given the sequence of minimizers, consider the new rescaled sequence $\widetilde{\mathbf{m}}_q(x) = \mathbf{m}_q(\frac{x}{\sqrt{q}})$. This sequence is bounded in the following sense:

$$\begin{aligned}\int_{\mathbb{R}}\widetilde{u}_q^2(x)\,dx + \int_{\mathbb{R}}|\widetilde{\mathbf{m}}_q'|^2\,dx &= \sqrt{q}\int_{\mathbb{R}}u_q^2(x)\,dx + \frac{1}{\sqrt{q}}\int_{\mathbb{R}}|\mathbf{m}_q'|^2\,dx \\ &\leq \frac{1}{\sqrt{q}}G_{q,\alpha,\delta}[\mathbf{m}_q] \leq C.\end{aligned} \tag{3.5}$$

Thus there is a subsequence (not relabeled) that converges weakly in $\mathcal{A}_N$ to $\mathbf{n} \in \mathcal{A}_N$.

Consider now $\mathbf{n}_q(x) = \mathbf{n}(\sqrt{q}x)$. Then,

$$\widehat{\mathbf{n}}_q(\xi) = \frac{1}{\sqrt{q}}\widehat{\mathbf{n}}\left(\frac{\xi}{\sqrt{q}}\right) \tag{3.6}$$

and

$$\begin{aligned} G_{q,\alpha,\delta}[\mathbf{n}_q] &= \sqrt{q}\int_{\mathbb{R}} u^2(\xi)\,d\xi + \sqrt{q}\int_{\mathbb{R}} 4\pi^2\xi^2|\widehat{\mathbf{n}}(\xi)|^2\,d\xi \\ &\quad + \frac{1}{\sqrt{q}}\int_{\mathbb{R}} \widehat{u}^2(\xi)\left(1 - \frac{1-e^{-2\pi\delta|\xi|\sqrt{q}}}{2\pi\delta|\xi|\sqrt{q}} - \frac{1}{2}e^{-4\pi a|\xi|\sqrt{q}}\frac{(1-e^{-2\pi\delta|\xi|\sqrt{q}})^2}{2\pi\delta|\xi|\sqrt{q}}\right)d\xi. \end{aligned} \tag{3.7}$$

We need to take the limit of the stray field energy. Since $\alpha\sqrt{q} \ll 1$ and $\delta\sqrt{q} \ll 1$,

$$\lim_{q\to 0}\frac{1}{q}\left(1 - \frac{1-e^{-2\pi\delta|\xi|\sqrt{q}}}{2\pi\delta|\xi|\sqrt{q}} - \frac{1}{2}e^{-4\pi a|\xi|\sqrt{q}}\frac{(1-e^{-2\pi\delta|\xi|\sqrt{q}})^2}{2\pi\delta|\xi|\sqrt{q}}\right) = \left(\frac{1}{3}\delta + a\right)4\pi^2|\xi|^2\delta. \tag{3.8}$$

The stray field energy can be written as

$$\frac{1}{\sqrt{q}}E_s = \delta\left(\frac{1}{3}\delta + a\right)\int_{\mathbb{R}} 4\pi^2\xi^2\widehat{u}^2(\xi)\widehat{\varphi}(\sqrt{q}\xi)\,d\xi, \tag{3.9}$$

where

$$\widehat{\varphi}(\xi) = \frac{1}{\left(\frac{1}{3}\delta + a\right)}\frac{1}{4\pi^2\xi^2}\left(1 - \frac{1-e^{-2\pi\delta|\xi|}}{2\pi\delta|\xi|} - \frac{1}{2}e^{-4\pi a|\xi|}\frac{(1-e^{-2\pi\delta|\xi|})^2}{2\pi\delta|\xi|}\right). \tag{3.10}$$

Note that $\widehat{\varphi}(0) = 1$ and $\widehat{\varphi} \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$, so $\varphi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. Therefore, the stray field can be written, in real space, as

$$\frac{1}{\sqrt{q}}E_s = \delta\left(\frac{1}{3}\delta + a\right)\int_{\mathbb{R}} u\left(u * \varphi_{\sqrt{q}}\right)dx, \tag{3.11}$$

where $\varphi_{\sqrt{q}}(x) = \frac{1}{\sqrt{q}}\phi(\frac{x}{\sqrt{q}})$, which is an approximation to the identity. Therefore we can take the limit in (3.7), and we obtain

$$\begin{aligned} \lim_{q\to 0}\frac{1}{\sqrt{q}}G_{q,\alpha,\delta}[\mathbf{n}_q] &= \int_{\mathbb{R}}\widehat{u}^2(\xi)\,d\xi + \int_{\mathbb{R}} 4\pi^2\xi^2\left(1 + \delta\left(\frac{1}{3}\delta + a\right)\right)|\widehat{u}(\xi)|^2\,d\xi \\ + \frac{1}{2}\int_{\mathbb{R}} 4\pi^2\xi^2|\widehat{w}(\xi)|^2\,d\xi &= \int_{\mathbb{R}} u^2\,dx + \left(1 + \delta\left(\frac{1}{3}\delta + a\right)\right)\int_{\mathbb{R}}(u')^2\,dx + \int_{\mathbb{R}}(w')^2\,dx \\ &= F_{\alpha,\delta}[\mathbf{n}]. \end{aligned} \tag{3.12}$$

Since $\mathbf{m}_q$ was a minimizer,

$$\frac{1}{\sqrt{q}}G_{q,\alpha,\delta}[\mathbf{m}_q] \le \frac{1}{\sqrt{q}}G_{q,\alpha,\delta}[\mathbf{n}_q], \tag{3.13}$$

and thus

$$\limsup_{q\to 0}\frac{1}{\sqrt{q}}G_{q,\alpha,\delta}[\mathbf{m}_q] \le \lim_{q\to 0}\frac{1}{\sqrt{q}}G_{q,\alpha,\delta}[\mathbf{n}_q] = F_{\alpha,\delta}[\mathbf{n}]. \tag{3.14}$$

Observe now that

$$\frac{1}{\sqrt{q}} G_{q,\alpha,\delta}[\mathbf{m}_q] = H_{q,\alpha,\delta}[\widetilde{\mathbf{m}}_q], \tag{3.15}$$

where

$$H_{q,\alpha,\delta}[\mathbf{m}] = \int_{\mathbb{R}} u^2\,dx + \int_{\mathbb{R}} |\mathbf{m}'|^2\,dx + \delta\left(\frac{1}{3}\delta + \alpha\right)\int_{\mathbb{R}} u\left(u * \varphi_{\sqrt{q}}\right)dx. \tag{3.16}$$

Since $\widetilde{\mathbf{m}}_q$ converges to $\mathbf{n}$ weakly in $\mathcal{A}_N$, by the lower semicontinuity of the functional,

$$F_{\alpha,\delta}[\mathbf{n}] \le \liminf_{q\to 0} \frac{1}{\sqrt{q}} G_{q,\alpha,\delta}[\mathbf{m}_q]. \tag{3.17}$$

Combining (3.14) and (3.17), we conclude that

$$\lim_{q\to 0} \frac{1}{\sqrt{q}} G_{q,\alpha,\delta}[\mathbf{m}_q] = F_{\alpha,\delta}[\mathbf{n}]. \tag{3.18}$$

It is easy to see that $\mathbf{n}$ must be a minimizer of $F_{\alpha,\delta}$: Given any $\mathbf{m}_0 \in \mathcal{A}$, consider $\widetilde{\mathbf{n}}_q(x) = \mathbf{m}_0\left(\sqrt{q}x\right)$. Then

$$F_{\alpha,\delta}[\mathbf{m}_0] = \lim_{q\to 0} \frac{1}{\sqrt{q}} G_{q,\alpha,\delta}[\widetilde{\mathbf{n}}_q] \ge \lim_{q\to 0} \frac{1}{\sqrt{q}} G_{q,\alpha,\delta}[\mathbf{m}_q] = F_{\alpha,\delta}[\mathbf{n}]. \tag{3.19}$$

This proves (i). Since the sequence $\mathbf{n}_q$ converges weakly to $\mathbf{n}$ and the energies converge, the limit is strong, which proves (ii). □

From the previous proof, it is easy to see that $H_{q,\alpha,\delta} \to F_{\alpha,\delta}$ in $\mathcal{A}_N$ as $q \to 0$, in the Γ-limit sense [13].

4. Asymptotic analysis of the limiting profile. We perform a formal asymptotic expansion of the minimizers of functional $F_{\alpha,\delta}$, defined in (3.3), for $\delta \ll 1$. Since $|\mathbf{m}| = 1$, it is customary to write $\mathbf{m} = (\cos\theta, 0, \sin\theta)$, where $\theta \to \pm\frac{\pi}{2}$ as $x \to \pm\infty$. We consider the functional

$$F_\beta[\mathbf{m}] = \int_{\mathbb{R}} m_1^2\,dx + (1+\beta)\int_{\mathbb{R}} (m_1')^2\,dx + \int_{\mathbb{R}} (m_3')^2\,dx, \tag{4.1}$$

where $\beta = \delta(3\alpha + \delta)/3 \ll 1$. We do the change of variables to θ, and get

$$F_\beta[\theta] = \int_{\mathbb{R}} \cos^2\theta\,dx + \int_{\mathbb{R}} (\theta')^2\,dx + \beta\int_{\mathbb{R}} (\theta')^2\sin^2\theta\,dx. \tag{4.2}$$

The Euler–Lagrange equation is

$$(1+\beta\sin^2\theta)\theta'' + (1+\beta(\theta')^2)\sin\theta\cos\theta = 0. \tag{4.3}$$

For $\beta = 0$, we get

$$\theta'' + \sin\theta\cos\theta = 0, \tag{4.4}$$

which has as solution $\cos\theta = \tanh x$. This is the Landau–Lifshitz profile [12]. The asymptotic analysis can be carried out more easily if we consider profiles of the form

$\mathbf{m} = (\operatorname{sech}\varphi, 0, \tanh\varphi)$ instead, and write the equation for φ. Since $\cos\theta = \operatorname{sech}\varphi$ and $\sin\theta = \tanh\varphi$, we get $\theta' = \varphi'\operatorname{sech}\varphi$ and

$$\theta'' = \operatorname{sech}\varphi\left(\varphi'' - (\varphi')^2\tanh\varphi\right). \tag{4.5}$$

The equation becomes then

$$\left(1+\beta\tanh^2\varphi\right)\left(\varphi'' - (\varphi')^2\tanh\varphi\right) + \left(1+\beta(\varphi')^2\operatorname{sech}^2\varphi\right)\tanh\varphi = 0. \tag{4.6}$$

Assume that $\varphi \sim \varphi_0 + \beta\varphi_1 + O(\beta^2)$. Then,

$$\begin{aligned}\left(1+\beta\tanh^2(\varphi_0+\beta\varphi_1)\right)&\left(\varphi_0'' + \beta\varphi_1'' - (\varphi_0'+\beta\varphi_1')^2\tanh(\varphi_0+\beta\varphi_1)\right)\\ &+\left(1+\beta(\varphi_0'+\beta\varphi_1')^2\operatorname{sech}^2(\varphi_0+\beta\varphi_1)\right)\tanh(\varphi_0+\beta\varphi_1) = 0.\end{aligned} \tag{4.7}$$

Collecting terms, we get

$$\varphi_0'' + (1-(\varphi_0')^2)\tanh\varphi_0 = 0. \tag{4.8}$$

The solution is $\varphi_0 = x$. The next term in (4.7) is

$$\begin{aligned}\varphi_1'' - 2\varphi_1'\varphi_0'\tanh\varphi_0 - (\varphi_0')^2\operatorname{sech}^2\varphi_0\varphi_1 + \tanh^2(\varphi_0)&\left(\varphi_0'' - (\varphi_0')^2\tanh\varphi_0\right)\\ +(\varphi_0')^2\operatorname{sech}^2\varphi_0\tanh\varphi_0 + \operatorname{sech}^2\varphi_0\varphi_1 &= 0.\end{aligned} \tag{4.9}$$

Since $\varphi_0 = x$ and $\varphi_0' = 1$, the equation simplifies to

$$\varphi_1'' - 2\varphi_1'\tanh x - \tanh^3 x + \operatorname{sech}^2 x\tanh x = 0. \tag{4.10}$$

Then

$$\left(\frac{\varphi_1'}{\cosh^2 x}\right)' = \frac{\varphi_1'' - 2\varphi_1'\tanh x}{\cosh^2 x} = \operatorname{sech}^2 x\tanh^3 x - \operatorname{sech}^4 x\tanh x. \tag{4.11}$$

We can integrate the right-hand side:

$$\begin{aligned}\int \operatorname{sech}^2 x\tanh^3 x\,dx &= \int \frac{\sinh x\left(\cosh^2 x - 1\right)}{\cosh^5 x}\,dx = \frac{1}{4}\frac{1}{\cosh^4 x} - \frac{1}{2}\frac{1}{\cosh^2 x},\\ \int \operatorname{sech}^4 x\tanh x\,dx &= \int \frac{\sinh x}{\cosh^5 x}\,dx = -\frac{1}{4}\frac{1}{\cosh^4 x}.\end{aligned} \tag{4.12}$$

Therefore,

$$\varphi_1' = C\cosh^2 x + \frac{1}{2}\frac{1}{\cosh^2 x} - \frac{1}{2} \tag{4.13}$$

and, integrating,

$$\varphi_1 = C\frac{\sinh 2x + 2x}{4} + \frac{1}{2}\tanh x - \frac{x}{2} + D. \tag{4.14}$$

We want φ to be odd, so $D = 0$. Unless $C = 0$, φ_1 will dominate over φ_0, so we take $C = 0$, and finally,

$$\varphi_1 = \frac{1}{2}\tanh x - \frac{x}{2}. \tag{4.15}$$

Therefore,

$$\mathbf{m} = \left(\operatorname{sech}\left(x + \frac{\beta}{2}\left(\tanh x - x\right)\right), 0, \tanh\left(x + \frac{\beta}{2}\left(\tanh x - x\right)\right)\right) + O(\beta^2). \tag{4.16}$$

5. Numerical experiments. We have implemented a truncated Newton method with an inexact line search for the minimization of

$$\begin{aligned} G_{q,\alpha,\delta}[\mathbf{m}_1,\mathbf{m}_2] &= \frac{1}{2}q\int_{\mathbb{R}} u_1^2\,dx + \frac{1}{2}\int_{\mathbb{R}} |\mathbf{m}_1'|^2\,dx + \frac{1}{2}\int_{\mathbb{R}} u_1^2\,dx - \frac{1}{2}\int_{\mathbb{R}} u_1\,(\Gamma_\delta * u_1)\,dx \\ &\quad + \frac{1}{2}q\int_{\mathbb{R}} u_2^2\,dx + \frac{1}{2}\int_{\mathbb{R}} |\mathbf{m}_2'|^2\,dx + \frac{1}{2}\int_{\mathbb{R}} u_2^2\,dx - \frac{1}{2}\int_{\mathbb{R}} u_2\,(\Gamma_\delta * u_2)\,dx \\ &\quad - \frac{\delta}{2}\int_{\mathbb{R}} u_1\,(u_2 * \Theta_{\alpha,\delta})\,dx. \end{aligned} \tag{5.1}$$

The method is well known, and the details can be found in the literature [15], so we will simply describe some of the details particular to our implementation.

We consider a finite interval $I = [-M, M]$, and restrict the functional to I. We have performed simulations in several intervals of increasing size until no change was found in the characteristics of the wall. For the results presented here we used $I = [-200, 200]$. We define the grid points $x_i = -M + i\Delta x$, for $i = 0, 1, \ldots, n+1$, where $\Delta x = \frac{2M}{n+1}$. The magnetization is approximated by a linear interpolant in the subinterval $I_i = [x_i, x_{i+1}]$, for $i = 0, 1, \ldots, n$. We impose the boundary conditions $u_0 = u_{n+1} = 0$. For the simulations presented here we fixed the parameters $\delta = 1$ and $a = 10^{-1}$. The parameter q varied in the range $q \in [10^{-3}, 1]$.

To evaluate the stray field, we need to approximate convolution integrals of the form

$$v(x_j) = \int_{-M}^{M} u(s)K(x_j - s)\,ds. \tag{5.2}$$

Substituting the piecewise linear interpolant,

$$v(x_j) \approx \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} \left(u_i + \frac{u_{i+1} - u_i}{\Delta x}(s - x_i)\right) K(x_j - s)\,ds. \tag{5.3}$$

Grouping terms,

$$v_j = \sum_{i=1}^{n} u_i \left[\int_{x_i}^{x_{i+1}} \left(1 - \frac{(s - x_i)}{\Delta x}\right) K(x_j - s)\,ds + \int_{x_{i-1}}^{x_i} \frac{(s - x_{i-1})}{\Delta x} K(x_j - s)\,ds\right]. \tag{5.4}$$

This can be written in the form

$$v_j = \sum_{i=1}^{n} K_{j-i} u_i, \tag{5.5}$$

where

$$K_\lambda = \Delta x \int_0^1 (1-t)\,K\left(\Delta x(\lambda - t)\right) + tK\left(\Delta x(\lambda + 1 - t)\right)\,dt. \tag{5.6}$$

The sum (5.5) has the shape of a discrete convolution, and it can therefore be efficiently evaluated using the fast Fourier transform (FFT) in $O(n \log n)$ operations.

The unit length constraint in the magnetization is taken into account by considering a line search on the function

$$h(\epsilon) = G_{q,\alpha,\delta}\left[\frac{\mathbf{m}_1 + \epsilon\mathbf{p}_1}{|\mathbf{m}_1 + \epsilon\mathbf{p}_1|}, \frac{\mathbf{m}_2 + \epsilon\mathbf{p}_2}{|\mathbf{m}_2 + \epsilon\mathbf{p}_2|}\right], \tag{5.7}$$

where $(\mathbf{p}_1, \mathbf{p}_2)$ is a descent direction, i.e., $h'(0) < 0$.

In Figure 2 we show the profiles of several minimizers as a function of q. All the numerically computed minimizers were opposed Néel walls, consistent with Lemma 2.1. These same profiles are presented in Figure 3, but this time the abscissa is rescaled by $\sqrt{q}$. As can be seen, all the plots collapse into one, illustrating that the core length is $1/\sqrt{q}$, as proved in section 4. This limiting profile is plotted in Figure 4, where it is compared to the profile of the minimizer of (5.1) computed numerically. The profiles are almost indistinguishable.

The computed values for the minimum energy as a function of q are presented in Table 1. From the results in the third column it is clear that the energy scales like $\sqrt{q}$. We plot the results in a logarithmic scale in Figure 5. The energy of the asymptotic approximation (4.16) was computed to machine precision using adaptive Gaussian quadrature. The computed energy coincides to a remarkable degree with the energy obtained using the asymptotic analysis described in the previous section.

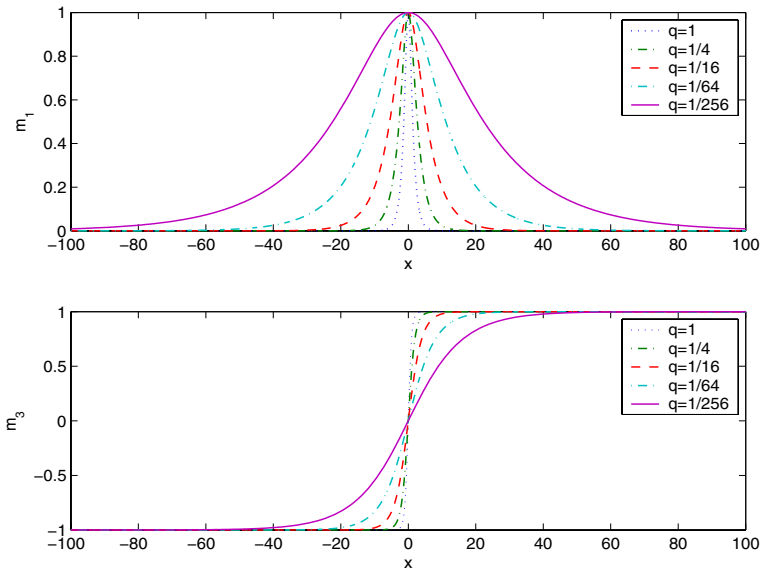


FIG. 2. *Néel wall profiles for several values of q.*

6. Conclusion. We have presented a new model for the analysis of one-dimensional walls in double layers. Using this new model, we have studied the structure of Néel walls and obtained the core length of the wall, the optimal energy scaling, and the structure of the minimizers. The main observation is that in a symmetric double layer the Néel wall no longer has a long logarithmic tail. The wall profile becomes local and similar to the classic Landau–Lifshitz wall. Thus, the range of nonlocal interactions is considerably reduced. We have implemented a truncated Newton method for energy minimization, and illustrated all the results numerically. In our simulations

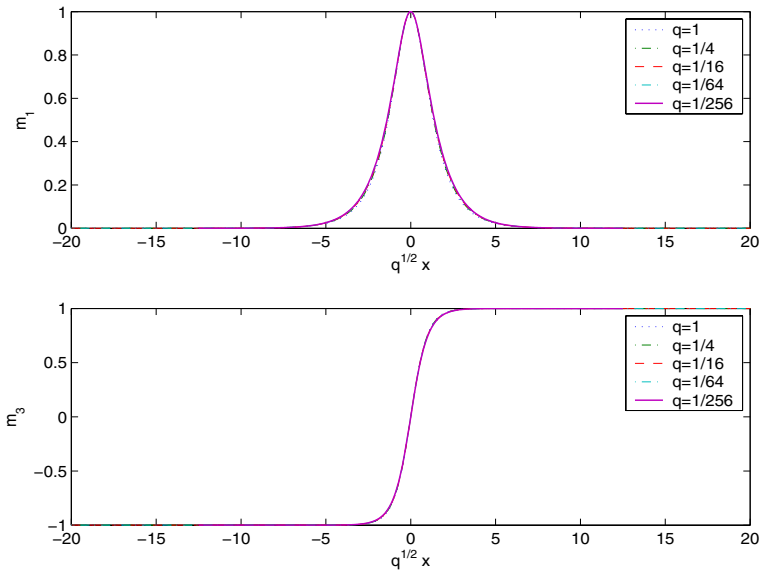


FIG. 3. *Néel wall profiles for several values of q. The abscissa has been rescaled to illustrate that the core length scales like $\sqrt{q}$.*

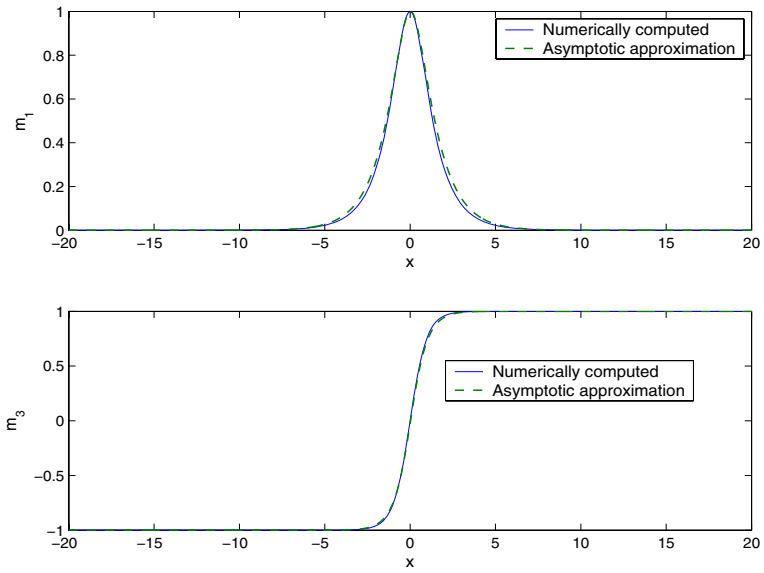


FIG. 4. *Comparison between the computed limiting profile and the asymptotic approximation obtained in section* 2.

we have managed to accurately capture the energy scaling, the core length, and the structure of the wall.

Appendix A. One-dimensional model for double layers. In this section we derive the model used in this article, starting from the Landau–Lifshitz energy functional.

TABLE 1

Minimum energy as a function of q. The energy is computed with the energy minimization algorithm described in section 5, and the energy scaling in q is obtained. We fixed the parameters $\delta = 1$ and $a = 10^{-1}$.

Energy scaling as $q \to 0$			
n	$q_n = 2^{-n}$	E_n	$\log \frac{E_{n-1}}{E_n} / \log 2$
0	1.000000E+00	0.413832547E+01	
1	5.000000E-01	0.294556828E+01	0.490501
2	2.500000E-01	0.209484869E+01	0.491699
3	1.250000E-01	0.148838645E+01	0.493096
4	6.250000E-02	0.105647543E+01	0.494489
5	3.125000E-02	0.749249198E+00	0.495741
6	1.562500E-02	0.530979818E+00	0.496788
7	7.812500E-03	0.376078916E+00	0.497621
8	3.906250E-03	0.266248579E+00	0.498261
9	1.953125E-04	0.188430720E+00	0.498739

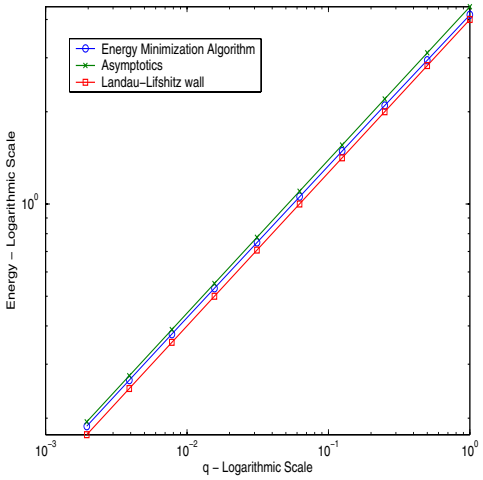


FIG. 5. *Comparison between the numerically computed energy, the optimal energy obtained in section 2, and the energy of the Landau–Lifshitz wall. The asymptotics give us an upper bound, while the Landau–Lifshitz wall provides us with a lower bound.*

We consider a double layer, infinite in both the x and y directions, and solve the magnetostatic equation in two dimensions. The stray field is $\mathbf{h}_s(x,z) = -\nabla\eta$, where

$$\eta = \nabla N * \mathbf{m} \tag{A.1}$$

is the magnetostatic potential, and $N(\mathbf{x}) = \frac{1}{2\pi}\log(|\mathbf{x}|)$, $\mathbf{x} = (x,z) \in \mathbb{R}^2$.

For the study on one-dimensional walls, we assume that $\mathbf{m}$ depends only on x. The double layer will be identified with the domain $\Omega = \mathbb{R}\times[-a-D_1,-a]\cup[a,a+D_2]$. In the bottom layer, we have $\mathbf{m} = (u_1, v_1, w_1)$, and in the top layer, $\mathbf{m} = (u_2, v_2, w_2)$.

Then,

$$\eta(x,z) = \int_\Omega \partial_x N(x-s,z-t)u(s)\,ds + \int_\Omega \partial_z N(x-s,z-t)v(s)\,ds \tag{A.2}$$
$$= \int_\Omega \partial_x N(x-s,z-t)u(s)\,ds + \int_{\mathbb{R}} \left(N(x-s,z+a+D_1) - N(x-s,z+a)\right) v_1(s)\,ds$$
$$+ \int_{\mathbb{R}} \left(N(x-s,z-a) - N(x-s,z-a-D_2)\right) v_2(s)\,ds.$$

The stray field energy is

$$\frac{2}{\mu_0 M_s^2} E = \int_\Omega \nabla\eta\cdot\mathbf{m}\,dx\,dz = \int_\Omega u\partial_x\eta + v\partial_z\eta\,dx\,dz \tag{A.3}$$
$$= \int_{\mathbb{R}} u_1(x) \int_{-a-D_1}^{-a} \partial_x\eta(x,z)\,dz\,dx + \int_{\mathbb{R}} u_2(x) \int_a^{a+D_2} \partial_x\eta(x,z)\,dz\,dx$$
$$+ \int_{\mathbb{R}} v_1(x)\left(\eta(x,-a) - \eta(x,-a-D_1)\right)dx + \int_{\mathbb{R}} v_2(x)\left(\eta(x,a+D_2) - \eta(x,a)\right)dx.$$

We can easily compute the derivative of η w.r.t. x:

$$\partial_x\eta(x,z) = \int_\Omega \partial_{xx} N(x-s,z-t)u(s)\,ds \tag{A.4}$$
$$+ \int_{\mathbb{R}} \left(\partial_x N(x-s,z+a+D_1) - \partial_x N(x-s,z+a)\right) v_1(s)\,ds$$
$$+ \int_{\mathbb{R}} \left(\partial_x N(x-s,z-a) - \partial_x N(x-s,z-a-D_2)\right) v_2(s)\,ds.$$

Since $\Delta N = \delta$, we get that $\partial_{xx}N(x-s,z-t) = \delta_{(x,z)} - \partial_{zz}N(x-s,z-t)$. Substituting this,

$$\partial_x\eta(x,z) = u(x) - \int_\Omega \partial_{zz} N(x-s,z-t)u(s)\,ds$$
$$+ \int_{\mathbb{R}} \left(\partial_x N(x-s,z+a+D_1) - \partial_x N(x-s,z+a)\right) v_1(s)\,ds \tag{A.5}$$
$$+ \int_{\mathbb{R}} \left(\partial_x N(x-s,z-a) - \partial_x N(x-s,z-a-D_2)\right) v_2(s)\,ds$$
$$= u(x) - \int_{\mathbb{R}} \left(\partial_z N(x-s,z+a+D_1) - \partial_z N(x-s,z+a)\right) u_1(s)\,ds$$
$$- \int_{\mathbb{R}} \left(\partial_z N(x-s,z-a) - \partial_z N(x-s,z-a-D_2)\right) u_2(s)\,ds$$
$$+ \int_{\mathbb{R}} \left(\partial_x N(x-s,z+a+D_1) - \partial_x N(x-s,z+a)\right) v_1(s)\,ds$$
$$+ \int_{\mathbb{R}} \left(\partial_x N(x-s,z-a) - \partial_x N(x-s,z-a-D_2)\right) v_2(s)\,ds.$$

Now we compute the energy step by step:

(A.6)
$$\int_{\mathbb{R}} u_1(x) \int_{-a-D_1}^{-a} \partial_x \eta(x,z)\,dz\,dx = D_1 \int_{\mathbb{R}} u_1^2(x)\,dx - \int_{\mathbb{R}} u_1(x) \int_{\mathbb{R}} u_1(s)(N(x-s,D_1)$$
$$-2N(x-s,0)+N(x-s,-D_1))ds\,dx$$
$$-\int_{\mathbb{R}} u_1(x) \int_{\mathbb{R}} u_2(s)\big(N(x-s,-2a)-N(x-s,-2a-D_1)$$
$$-N(x-s,-2a-D_2)+N(x-s,-2a-D_1-D_2)\big)ds\,dx$$
$$+\frac{1}{2\pi}\int_{\mathbb{R}} u_1(x) \int_{\mathbb{R}} v_2(s)\bigg(\arctan\left(\frac{2a+D_1}{x-s}\right)-\arctan\left(\frac{2a}{x-s}\right)$$
$$-\arctan\left(\frac{2a+D_1+D_2}{x-s}\right)+\arctan\left(\frac{2a+D_2}{x-s}\right)\bigg)ds\,dx$$
$$= D_1\int_{\mathbb{R}} u_1^2\,dx - D_1\int_{\mathbb{R}} u_1\,(u_1 * \Gamma_{D_1})\,dx + \frac{D_1D_2}{2}\int_{\mathbb{R}} u_1\,(u_2 * \Theta_{a,D_1,D_2})\,dx$$
$$+\frac{D_1D_2}{2}\int_{\mathbb{R}} u_1\,(v_2 * \Psi_{a,D_1,D_2})\,dx,$$

where we have defined

(A.7)
$$\Gamma_{D_i}(x) = \frac{1}{2\pi D_i}\log\left(1+\frac{D_i^2}{x^2}\right), \quad i=1,2,$$
$$\Theta_{a,D_1,D_2}(x) = \frac{1}{2D_1D_2\pi}\left(\log\left(\frac{x^2+(2a+D_1)^2}{x^2+(2a+D_1+D_2)^2}\right)-\log\left(\frac{x^2+4a^2}{x^2+(2a+D_2)^2}\right)\right),$$
$$\Psi_{a,D_1,D_2}(x) = \frac{1}{D_1D_2\pi}\bigg(\arctan\left(\frac{2a+D_1}{x-s}\right)-\arctan\left(\frac{2a}{x-s}\right)$$
(A.8)
$$-\arctan\left(\frac{2a+D_1+D_2}{x-s}\right)+\arctan\left(\frac{2a+D_2}{x-s}\right)\bigg),$$

(A.9)
$$\int_{\mathbb{R}} u_2(x) \int_a^{a+D_2} \partial_x \eta(x,z)\,dz\,dx = D_2\int_{\mathbb{R}} u_2^2\,dx - D_2\int_{\mathbb{R}} u_2\,(u_2 * \Gamma_{D_2})\,dx$$
$$+\frac{D_1D_2}{2}\int_{\mathbb{R}} u_2\,(u_1 * \Theta_{a,D_1,D_2})\,dx - \frac{D_1D_2}{2}\int_{\mathbb{R}} u_2\,(v_1 * \Psi_{a,D_1,D_2})\,dx.$$

Now

$$\begin{aligned}
&\eta(x,-a)-\eta(x,-a-D_1)\\
&\quad=\int_{\mathbb{R}} u_1(s)\int_{-a-D_1}^{-a}\left(\partial_x N(x-s,-a-t)-\partial_x N(x-s,-a-D_1-t)\right)dt\,ds\\
&\qquad+\int_{\mathbb{R}} u_2(s)\int_{a}^{a+D_2}\left(\partial_x N(x-s,-a-t)-\partial_x N(x-s,-a-D_1-t)\right)dt\,ds\\
&\qquad+\int_{\mathbb{R}}\left(N(x-s,D_1)-2N(x-s,0)+N(x-s,-D_1)\right)v_1(s)\,ds\\
&\qquad+\int_{\mathbb{R}}\big(N(x-s,-2a)-N(x-s,-2a-D_2)\\
&\qquad\qquad-N(x-s,-2a-D_1)+N(x-s,-2a-D_1-D_2)\big)v_2(s)\,ds\\
&\quad=\frac{D_1D_2}{2}u_2*\Psi_{a,D_1,D_2}+D_1v_1*\Gamma_{D_1}-\frac{D_1D_2}{2}v_2*\Theta_{a,D_1,D_2}
\end{aligned}\tag{A.10}$$

and

$$\begin{aligned}
&\eta(x,a+D_2)-\eta(x,a)\\
&\quad=\int_{\mathbb{R}} u_2(s)\int_{a}^{a+D_2}\left(\partial_x N(x-s,a+D_2-t)-\partial_x N(x-s,a-t)\right)dt\,ds\\
&\qquad+\int_{\mathbb{R}} u_1(s)\int_{-a-D_1}^{-a}\left(\partial_x N(x-s,a+D_2-t)-\partial_x N(x-s,a-t)\right)dt\,ds\\
&\qquad+\int_{\mathbb{R}}\big(N(x-s,2a+D_1+D_2)-N(x-s,2a+D_2)\\
&\qquad\qquad-N(x-s,2a+D_1)+N(x-s,2a)\big)v_1(s)\,ds\\
&\qquad+\int_{\mathbb{R}}\big(N(x-s,D_2)-2N(x-s,0)+N(x-s,-D_2)\big)v_2(s)\,ds\\
&\quad=-\frac{D_1D_2}{2}u_1*\Psi_{a,D_1,D_2}-\frac{D_1D_2}{2}v_1*\Theta_{a,D_1,D_2}+D_2v_2*\Gamma_{D_2}.
\end{aligned}\tag{A.11}$$

Assembling all this, we get the stray field energy:

$$\begin{aligned}
\frac{2}{\mu_0 M_s^2}E_s&=D_1\int_{\mathbb{R}}u_1^2\,dx-D_1\int_{\mathbb{R}}u_1\left(u_1*\Gamma_{D_1}\right)dx+D_1\int_{\mathbb{R}}v_1\left(v_1*\Gamma_{D_1}\right)dx\\
&\quad+D_2\int_{\mathbb{R}}u_2^2\,dx-D_2\int_{\mathbb{R}}u_2\left(u_2*\Gamma_{D_2}\right)dx+D_2\int_{\mathbb{R}}v_2\left(v_2*\Gamma_{D_2}\right)dx\\
&\quad+D_1D_2\int_{\mathbb{R}}u_1\left(u_2*\Theta_{a,D_1,D_2}\right)-v_1\left(v_2*\Theta_{a,D_1,D_2}\right)dx\\
&\quad+D_1D_2\int_{\mathbb{R}}u_1\left(v_2*\Psi_{a,D_1,D_2}\right)-v_1\left(u_2*\Psi_{a,D_1,D_2}\right)dx.
\end{aligned}\tag{A.12}$$

In order to write this in Fourier space, we need the Fourier transform of the convolution kernels. We start with the following:

$$\begin{aligned}\int_{\mathbb{R}} \log\left(\frac{x^2+\alpha^2}{x^2+\beta^2}\right) e^{-2\pi i\xi x}\,dx &= \int_{\mathbb{R}} \log\left(1+\frac{\alpha^2-\beta^2}{x^2+\beta^2}\right) e^{-2\pi i\xi x}\,dx \\ &= \frac{\beta^2-\alpha^2}{2\pi i\xi}\int_{\mathbb{R}} \frac{1}{1+\frac{\alpha^2-\beta^2}{x^2+\beta^2}}\frac{2x}{(x^2+\beta^2)^2} e^{-2\pi i\xi x}\,dx \\ &= \frac{\beta^2-\alpha^2}{2\pi i\xi}\int_{\mathbb{R}} \frac{2x}{(x^2+\alpha^2)(x^2+\beta^2)} e^{-2\pi i\xi x}\,dx.\end{aligned} \tag{A.13}$$

Using residue theory, we get that

$$\begin{aligned}\int_{\mathbb{R}} \frac{2x}{(x^2+\alpha^2)(x^2+\beta^2)} e^{-2\pi i\xi x}\,dx &= 2\pi i(\text{Res }(f,i\alpha)+\text{Res }(f,i\beta)) \\ &= 2\pi i\left(\frac{2i\beta}{2i\beta(\alpha^2-\beta^2)}e^{2\pi\beta\xi}+\frac{2i\alpha}{2i\alpha(\beta^2-\alpha^2)}e^{2\pi\alpha\xi}\right)\end{aligned} \tag{A.14}$$

for $\xi<0$. When we put it all together, we get

$$\int_{\mathbb{R}} \log\left(\frac{x^2+\alpha^2}{x^2+\beta^2}\right) e^{-2\pi i\xi x}\,dx = \frac{e^{-2\pi\beta|\xi|}-e^{-2\pi\alpha|\xi|}}{|\xi|}. \tag{A.15}$$

Therefore

$$\frac{1}{2\pi}\int_{\mathbb{R}} \log\left(1+\frac{D_j^2}{x^2}\right) e^{-2\pi i\xi x}\,dx = \frac{1-e^{-2\pi D_j|\xi|}}{2\pi|\xi|}, \quad j=1,2, \tag{A.16}$$

and

$$\begin{aligned}\frac{1}{2\pi}\int_{\mathbb{R}} \left(\log\left(\frac{x^2+4a^2}{x^2+(2a+D_2)^2}\right)-\log\left(\frac{x^2+(2a+D_1)^2}{x^2+(2a+D_1+D_2)^2}\right)\right) e^{-2\pi i\xi x}\,dx \\ = -e^{-4\pi a|\xi|}\frac{(1-e^{-2\pi D_1|\xi|})(1-e^{-2\pi D_2|\xi|})}{2\pi|\xi|}.\end{aligned} \tag{A.17}$$

Finally,

$$\begin{aligned}&\frac{1}{\pi}\int_{\mathbb{R}} e^{-2\pi i\xi x}\arctan\left(\frac{\alpha}{x}\right)dx = -\frac{2i}{\pi}\int_0^\infty \sin(2\pi\xi x)\arctan\left(\frac{\alpha}{x}\right)dx \\ &\quad = -\frac{2i}{\pi}\left(-\frac{1}{2\pi\xi}\cos(2\pi\xi x)\arctan\left(\frac{\alpha}{x}\right)\Big|_0^\infty - \frac{\alpha}{2\pi\xi}\int_0^\infty \cos(2\pi\xi x)\frac{1}{\alpha^2+x^2}\,dx\right) \\ &\quad = -i\frac{1-e^{-2\pi|\xi|\alpha}}{2\pi\xi}.\end{aligned} \tag{A.18}$$

Therefore

$$\begin{aligned}&\frac{1}{\pi}\int_{\mathbb{R}} e^{-2\pi i\xi x}\Big(\arctan\left(\frac{2a+D_1}{x}\right)-\arctan\left(\frac{2a}{x}\right) \\ &\qquad -\arctan\left(\frac{2a+D_1+D_2}{x}\right)+\arctan\left(\frac{2a+D_2}{x}\right)\Big)dx \\ &\quad = -ie^{-4\pi a|\xi|}\frac{(1-e^{-2\pi|\xi|D_1})(1-e^{-2\pi|\xi|D_2})}{2\pi\xi}.\end{aligned} \tag{A.19}$$

Thus, the stray field energy can be written in Fourier space as

(A.20)
$$\begin{aligned}\frac{2}{\mu_0 M_s^2}E_s &= D_1\int_{\mathbb{R}}\widehat{u_1}^2(\xi)\left(1-\frac{1-e^{-2\pi D_1|\xi|}}{2\pi D_1|\xi|}\right)d\xi + D_1\int_{\mathbb{R}}\widehat{v_1}^2(\xi)\frac{1-e^{-2\pi D_1|\xi|}}{2\pi D_1|\xi|}\,d\xi\\ &+ D_2\int_{\mathbb{R}}\widehat{u_2}^2(\xi)\left(1-\frac{1-e^{-2\pi D_2|\xi|}}{2\pi D_2|\xi|}\right)d\xi + D_2\int_{\mathbb{R}}\widehat{v_2}^2(\xi)\frac{1-e^{-2\pi D_2|\xi|}}{2\pi D_2|\xi|}\,d\xi\\ &+ \int_{\mathbb{R}}(\widehat{u_1}(\xi)\overline{\widehat{u_2}}(\xi)-\widehat{v_1}(\xi)\overline{\widehat{v_2}}(\xi))e^{-4\pi a|\xi|}\frac{(1-e^{-2\pi D_1|\xi|})(1-e^{-2\pi D_2|\xi|})}{2\pi|\xi|}\,d\xi\\ &- i\int_{\mathbb{R}}\Big(\widehat{v_1}(\xi)\overline{\widehat{u_2}}(\xi)-\widehat{u_1}(\xi)\overline{\widehat{v_2}}(\xi)\Big)e^{-4a\pi|\xi|}\frac{(1-e^{-2\pi D_1|\xi|})(1-e^{-2\pi D_2|\xi|})}{2\pi\xi}\,d\xi.\end{aligned}$$

Note that the interaction between the layers decays exponentially with the spacer distance.

In this one-dimensional setting, the Landau–Lifshitz energy for a double layer reduces to

(A.21)
$$\begin{aligned}F[\mathbf{m}_1,\mathbf{m}_2] &= \frac{K_u D_1}{2}\int_{\mathbb{R}}\left(u_1^2+v_1^2\right)dx + \frac{AD_1}{2}\int_{\mathbb{R}}|\mathbf{m}_1'|^2\,dx\\ &+ \frac{D_1\mu_0 M_s^2}{2}\int_{\mathbb{R}}u_1^2\,dx - \frac{D_1\mu_0 M_s^2}{2}\int_{\mathbb{R}}u_1\left(\Gamma_{D_1}*u_1\right)dx + \frac{D_1\mu_0 M_s^2}{2}\int_{\mathbb{R}}v_1\left(\Gamma_{D_1}*v_1\right)dx\\ &+ \frac{K_u D_2}{2}\int_{\mathbb{R}}\left(u_2^2+v_2^2\right)dx + \frac{AD_2}{2}\int_{\mathbb{R}}|\mathbf{m}_2'|^2\,dx\\ &+ \frac{D_2\mu_0 M_s^2}{2}\int_{\mathbb{R}}u_2^2\,dx - \frac{D_2\mu_0 M_s^2}{2}\int_{\mathbb{R}}u_2\left(\Gamma_{D_2}*u_2\right)dx + \frac{D_2\mu_0 M_s^2}{2}\int_{\mathbb{R}}v_2\left(\Gamma_{D_2}*v_2\right)dx\\ &+ \frac{D_1 D_2\mu_0 M_s^2}{2}\int_{\mathbb{R}}u_1\left(u_2*\Theta_{a,D_1,D_2}\right)-v_1\left(v_2*\Theta_{a,D_1,D_2}\right)dx\\ &+ \frac{D_1 D_2\mu_0 M_s^2}{2}\int_{\mathbb{R}}v_1\left(u_2*\Psi_{a,D_1,D_2}\right)-u_1\left(v_2*\Psi_{a,D_1,D_2}\right)dx,\end{aligned}$$

where

$$\Gamma_i(x) = \frac{1}{2\pi D_i}\log\left(1+\frac{D_i^2}{x^2}\right),\quad i=1,2,$$

$$\Theta_{a,D_1,D_2}(x) = \frac{1}{2D_1D_2\pi}\left(\log\left(\frac{x^2+(2a+D_1)^2}{x^2+(2a+D_1+D_2)^2}\right)-\log\left(\frac{x^2+4a^2}{x^2+(2a+D_2)^2}\right)\right),$$

$$\begin{aligned}\Psi_{a,D_1,D_2}(x) &= \frac{1}{D_1D_2\pi}\Bigg(\arctan\left(\frac{2a+D_1}{x-s}\right)-\arctan\left(\frac{2a}{x-s}\right)\\ &\qquad -\arctan\left(\frac{2a+D_1+D_2}{x-s}\right)+\arctan\left(\frac{2a+D_2}{x-s}\right)\Bigg)\end{aligned}\tag{A.22}$$

and in Fourier space

$$\widehat{\Gamma}_{D_j}(\xi) = \frac{1 - e^{-2\pi D_j|\xi|}}{2\pi D_j|\xi|}, \quad j = 1, 2,$$

$$\widehat{\Theta}_{a,D_1,D_2}(\xi) = e^{-4\pi a|\xi|}\frac{\left(1 - e^{-2\pi D_1|\xi|}\right)\left(1 - e^{-2\pi D_2|\xi|}\right)}{2\pi D_1 D_2|\xi|},$$

$$\widehat{\Psi}_{a,D_1,D_2}(\xi) = ie^{-4\pi a|\xi|}\frac{\left(1 - e^{-2\pi D_1|\xi|}\right)\left(1 - e^{-2\pi D_2|\xi|}\right)}{2\pi D_1 D_2\xi}. \tag{A.23}$$

To write the energy in dimensionless variables, we define $x = lx'$, $a = l\alpha$, $D_1 = l\delta_1$, and $D_2 = l\delta_2$, where $l = \sqrt{A/(\mu_0 M_s^2)}$. Define also $q = K_u/(\mu_0 M_s^2)$. Performing this change of variables and dropping the prime in x', we obtain

$$\begin{aligned}\frac{1}{D_1\sqrt{\mu_0 M_s^2 A}}F[\mathbf{m}_1, \mathbf{m}_2] = {} & \frac{q}{2}\int_{\mathbb{R}}\left(u_1^2 + v_1^2\right)dx + \frac{1}{2}\int_{\mathbb{R}}|\mathbf{m}_1'|^2\,dx \\ & + \frac{1}{2}\int_{\mathbb{R}}u_1^2\,dx - \frac{1}{2}\int_{\mathbb{R}}u_1\left(\Gamma_{\delta_1} * u_1\right)dx \\ & + \frac{1}{2}\int_{\mathbb{R}}v_1\left(\Gamma_{\delta_1} * v_1\right)dx + \frac{q\delta_2}{2\delta_1}\int_{\mathbb{R}}\left(u_2^2 + v_2^2\right)dx \\ & + \frac{\delta_2}{2\delta_1}\int_{\mathbb{R}}|\mathbf{m}_2'|^2\,dx + \frac{\delta_2}{2\delta_1}\int_{\mathbb{R}}u_2^2\,dx \\ & - \frac{\delta_2}{2\delta_1}\int_{\mathbb{R}}u_2\left(\Gamma_{\delta_2} * u_2\right)dx + \frac{\delta_2}{2\delta_1}\int_{\mathbb{R}}v_2\left(\Gamma_{\delta_2} * v_2\right)dx \\ & + \frac{\delta_2}{2}\int_{\mathbb{R}}u_1\left(u_2 * \Theta_{\alpha,\delta_1,\delta_2}\right) - v_1\left(v_2 * \Theta_{\alpha,\delta_1,\delta_2}\right)dx \\ & + \frac{\delta_2}{2}\int_{\mathbb{R}}v_1\left(u_2 * \Psi_{\alpha,\delta_1,\delta_2}\right) - u_1\left(v_2 * \Psi_{\alpha,\delta_1,\delta_2}\right)dx.\end{aligned} \tag{A.24}$$

Appendix B. Validity range for the upper bound. In order to estimate the value of q_0 in (2.18), we need to study the stray field (2.17). Define

$$\eta = \frac{2\delta|\xi|\sqrt{q}}{\pi} \tag{B.1}$$

and

$$\beta = \frac{\alpha}{\delta}. \tag{B.2}$$

We need to study the Taylor series of

$$\phi(\eta) = 1 - \frac{1 - e^{-\eta}}{\eta} - \frac{1}{2}e^{-2\beta\eta}\frac{(1 - e^{-\eta})^2}{\eta}. \tag{B.3}$$

Using the Taylor polynomials of $e^{-\eta}$ and $e^{-2\beta\eta}$, we obtain

$$\begin{aligned}\phi(\eta) = \left(\frac{1}{3} + \beta\right)\eta^2 + \eta^3\Bigg(\frac{e^{-\eta_1}}{4!} - \beta^2 e^{-2\beta\eta_3} - \beta e^{-2\beta\eta_4} \\ - \frac{e^{-2\beta\eta}}{2}\left(\frac{1}{4} + \frac{1}{3}e^{-\eta_2} - \frac{\eta}{3!}e^{-\eta_2} + \frac{\eta^2}{36}e^{-2\eta_2}\right)\Bigg),\end{aligned} \tag{B.4}$$

where $\eta_k \in [0,\eta]$ for $k = 1,2,3,4$. Using (B.4), (B.1), and (B.2) in (2.17), we obtain

$$E_s \le \frac{2\delta\sqrt{q}}{3}\left(\frac{\delta}{3}+\alpha\right)+\frac{8\delta^3 q}{\pi^3}\left(\frac{1}{3}+\frac{\alpha^2}{\delta^2}+\frac{\alpha}{\delta}\right)\int |\xi|^3\mathrm{sech}^2(\xi)\,d\xi + \frac{4\delta^4 q^{3/2}}{3\pi^4}\int |\xi|^4\mathrm{sech}^2(\xi)\,d\xi + \frac{4\delta^5 q^2}{9\pi^5}\int |\xi|^5\mathrm{sech}^2(\xi)\,d\xi. \tag{B.5}$$

Estimate (2.20) holds for

$$q_0 \le \min\left\{M, \frac{4}{9}\left(\frac{\delta}{3}+\alpha\right)^2\left[\frac{8}{\pi^3}\left(\frac{\delta^2}{3}+\alpha^2+\delta\alpha\right)\int |\xi|^3\mathrm{sech}^2(\xi)\,d\xi + \frac{4\delta^3 M^{1/2}}{3\pi^4}\int |\xi|^4\mathrm{sech}^2(\xi)\,d\xi + \frac{4\delta^4 M^{3/2}}{9\pi^5}\int |\xi|^5\mathrm{sech}^2(\xi)\,d\xi\right]^{-2}\right\}, \tag{B.6}$$

where $M > 0$ is arbitrary.

Note that as $\alpha \to \infty$, $q_0 \to 0$. This is consistent with the fact that as $\alpha \to \infty$, the layers are decoupled, and we recover the Néel wall energy for a single layer, for which the upper bound is not (2.20) but (1.7).

Acknowledgment. The author wishes to thank the anonymous reviewer. His/her comments helped improve considerably the presentation of the results in this article.

REFERENCES

[1] A. Aharoni, *Energy of one dimensional domain walls in ferromagnetic films*, J. Appl. Phys., 37 (1966), pp. 3271–3279.
[2] L. Ahlfors, *Complex Analysis*, 3rd ed., McGraw-Hill, New York, 1979.
[3] G. Carbou, *Thin layers in micromagnetism*, Math. Models Methods Appl. Sci., 11 (2001), pp. 1529–1546.
[4] R. Choksi and R. V. Kohn, *Bounds on the micromagnetic energy of a uniaxial ferromagnet*, Comm. Pure Appl. Math., 51 (1998), pp. 259–289.
[5] R. Choksi, R. V. Kohn, and F. Otto, *Domain branching in uniaxial ferromagnets: A scaling law for the minimum energy*, Comm. Math. Phys., 201 (1999), pp. 61–79.
[6] C. J. García-Cervera, *Magnetic Domains and Magnetic Domain Walls*, Ph.D. thesis, Courant Institute of Mathematical Sciences, New York University, New York, 1999.
[7] C. J. García-Cervera, *One-dimensional magnetic domain walls*, European J. Appl. Math., 15 (2004), pp. 451–486.
[8] C. J. García-Cervera and W. E, *Effective dynamics in thin ferromagnetic films*, J. Appl. Phys., 90 (2000), pp. 370–374.
[9] G. Gioia and R. D. James, *Micromagnetics of very thin films*, Proc. Roy. Soc. London Ser. A, 453 (1997), pp. 213–223.
[10] A. Hubert and R. Schäfer, *Magnetic Domains: The Analysis of Magnetic Microstructures*, Springer-Verlag, Berlin, Heidelberg, New York, 1998.
[11] R. V. Kohn and V. V. Slastikov, *Effective dynamics for ferromagnetic thin films: A rigorous justification*, Proc. Roy. Soc. London Ser. A, 461 (2005), pp. 143–154.
[12] L. Landau and E. Lifshitz, *On the theory of the dispersion of magnetic permeability in ferromagnetic bodies*, Physikalische Zeitschrift der Sowjetunion, 8 (1935), pp. 153–169.
[13] G. Dal Maso, *An introduction to* Γ*-Convergence*, Progr. Nonlinear Differential Equations Appl. 8, Birkhäuser Boston, Cambridge, MA, 1993.
[14] S. Middelhoek, *Domain wall structures in magnetic double films*, J. Appl. Phys., 37 (1966), pp. 1276–1282.
[15] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Ser. Oper. Res., Springer-Verlag, New York, 1999.
[16] F. Otto, *Cross-over in scaling laws: A simple example from micromagnetics*, in Proceedings of the International Congress of Mathematicians, Vol. 3, Higher Education Press, Beijing, 2002, pp. 829–838.

[17] I. Puchalska and H. Niedoba, *Magnetization process in permalloy multilayer films*, IEEE Trans. Magn., 27 (1991), pp. 3579–3587.
[18] A. De Simone, R. V. Kohn, S. Müller, and F. Otto, *Magnetic microstructures—A paradigm of multiscale problems*, in Proceedings of the ICIAM, Edinburgh, 1999, Oxford University Press, Oxford, UK, 2000, pp. 175–190.
[19] J. C. Slonczewski, *Theory of domain-wall motion in magnetic films and platelets*, J. Appl. Phys., 44 (1973), pp. 1759–1770.
[20] J. C. Slonczewski, *Theory of bloch-line and bloch-wall motion*, J. Appl. Phys., 45 (1974), pp. 2705–2715.
[21] W. P. Ziemer, *Weakly Differentiable Functions. Sobolev Spaces and Functions of Bounded Variation*, Grad. Texts in Math. 120, Springer-Verlag, New York, 1989.

SIAM J. APPL. MATH.
Vol. 65, No. 5, pp. 1748–1771

EFFECTS OF SOG ON DPP-RECEPTOR BINDING*

YUAN LOU†, QING NIE‡, AND FREDERIC Y. M. WAN†

Abstract. Concentration gradients of morphogens are known to be instrumental in cell signaling and tissue patterning. Of interest here is how the presence of a competitor of BMP ligands affects cell signaling. The effects of Sog on the binding of Dpp with cell receptors are analyzed for dorsal-ventral morphogen gradient formation in vertebrate and *Drosophila* embryos. This prototype system includes diffusing ligands, degradation of morphogens, and cleavage of Dpp-Sog complexes by Tolloid to free up Dpp. Simple and biologically meaningful necessary and sufficient conditions for the existence of a steady state gradient configuration are established, and existence theorems are proved. For high Sog production rates (relative to the Dpp production rate), it is found that the steady state configuration exhibits a more intense Dpp-receptor concentration near the dorsal midline. Numerical simulations of the evolution of the system show that, beyond some threshold Sog production rate, the transient Dpp-receptor concentration at the dorsal midline would become more intense than that of the steady state, before subsiding and approaching a nonuniform steady state of lower magnitude. The magnitude of the transient concentration has been found to increase by several fold with increasing Sog production rate. The highly intense Dpp activity at and around the dorsal midline is consistent with available experimental observations and other analytical studies.

Key words. morphogen gradients, reaction-diffusion, pattern formation, mathematical modeling

AMS subject classifications. 92C15, 92C37, 35K57

DOI. 10.1137/S0036139903433219

1. Introduction.

The proper functioning of tissues and organs requires that each cell differentiate appropriately for its position. In many cases, the positional information that instructs cells about their prospective fate is conveyed by concentration gradients of morphogens bound to cell receptors. Morphogens are signaling molecules that, when bound to cell receptors, assign different cell fates at different concentrations [1], [2]. Morphogen action is of special importance in understanding development because it is a highly efficient way for a population of uncommitted cells in an embryo to create complex patterns of gene expression in space. This role of morphogens has been the prevailing thought in tissue patterning for over half a century, but only recently have there been sufficient experimental data and adequate analytical studies for us to begin to understand how various useful morphogen concentration gradients are formed [3], [4].

Dorsal-ventral (back-to-belly) patterning in vertebrate and *Drosophila* (fruit fly) embryos is now known to be regulated by bone morphogenetic proteins (BMP). The BMP activity is controlled mainly by several secreted factors including the antagonists *chordin* and *short gastrulation* (Sog). In *Drosophila*, seven zygotic genes have been proposed to regulate dorsal-ventral patterning. Among them, *decapentaplegic* (Dpp) encodes BMP homologues that promote dorsal cell fates such as *amnioserosa* and inhibits development of the ventral central nervous system. The chordin homologue

*Received by the editors August 13, 2003; accepted for publication (in revised form) December 28, 2004; published electronically July 26, 2005. The work was partially supported by NIH grants R01GM67247, P20GM066051, and NSF SCREM grant DMS0112416.
http://www.siam.org/journals/siap/65-5/43321.html
†Department of Mathematics, The Ohio State University, Columbus, OH 43210 (lou@math.ohio-state.edu).
‡Department of Mathematics, Center for Complex Biology Systems, University of California, Irvine, CA 92697 (qnie@math.uci.edu, fwan@math.uci.edu).

Sog is expressed ventrally and promotes central nervous system development. The phenotype of Sog loss-of-function mutants is intriguing; as expected for a Dpp antagonist, ventral structures are lost but, in addition, the amnioserosa is reduced. This result is paradoxical, as the amnioserosa is the dorsal-most tissue, and thus apparently a BMP antagonist is required for maximal BMP signaling [5], [6], [7], [8].

In principle, morphogen concentration gradients can be generated through the production of morphogens at particular sources, followed by their diffusion and degradation in appropriate regions [4], [9], [10], [11], [12]. In the above Dpp/Sog system, the production of Dpp is pretty much uniform in the dorsal region and absent in the ventral region, while the opposite is true for Sog. However, the Dpp activity has a sharp peak around the midline of the dorsal region in the presence of its "inhibitor" Sog. Mutation of Sog results in a reduction and a broadening of Dpp activity around the midline of the dorsal region. As the system contains many variables, the question of what leads to a sharp concentration peak is difficult to tackle by traditional experimental means.

Recently, Eldar et al. [13] studied a more complex morphogen system that includes the effects of Sog (and other morphogens) on Dpp activities. By performing massive computer calculations to search for molecular networks that support robustness, they found that the presence of the BMP inhibitor Sog stimulates intense Dpp activity at the dorsal midline resulting in highly nonuniform Dpp-receptor concentration in space for the the patterning process. They also showed that the Dpp concentration gradient itself is robust to changes in gene dosage. Two conditions were stipulated in their model to produce agreement with experimentally observed gradient formation. First, the steady state of the system is achieved by shutting off the production of Dpp through setting the production rate to zero 10 minutes after the initiation of the system [14], and there is no degradation of Dpp-receptor complex in the model. Second, the model requires immobility of free Dpp molecules; i.e., Dpp does not diffuse, but diffusion of the Dpp-Sog complexes and other ligands can occur.

For formation of morphogen gradients in a wing imaginal disc (a structure in the larva that will become the wing of the adult fly), Lander, Nie, and Wan [4] and Lou, Nie, and Wan [9] have demonstrated the important biological roles of diffusion for Dpp, and degradation for the Dpp-receptor complex. Without degradation, the steady state of the system is not achievable unless ligand production is shut off after a while, as in [14]. Eldar et al., in a recent paper [15], have also studied how degradation of ligand affects robustness of morphogen gradients. Most recently, the diffusion coefficient of Dpp has been measured *in vivo* using FCS (fluorescence correlation spectroscopy) techniques [16], and it was found that the magnitude of diffusion coefficient for Dpp is close to the magnitude of the diffusion coefficient for the Dpp-Sog complex used in [14] and hence not negligible.

Given the rather special restrictions on the Dpp/Sog system in [13] and [14], it is desirable to investigate the possibility of an alternative and simpler known biological mechanism for the generation of the intense Dpp activities around the dorsal midline. In this paper, we will extend the dynamic Dpp/Sog system formulated in [17] for morphogen activities in dorsal-ventral patterning by allowing for diffusion of ligands, degradation of the morphogens, and the cleavage of Dpp-Sog complexes by the enzyme Tolloid to free up a fraction τ of Dpp and to degrade part of Sog.

In this study, we will establish a biologically meaningful necessary and sufficient condition for the existence of a steady state. This condition requires a balance of the production of ligands, strength of degradation, and rate of cleavage of Dpp-Sog complex by Tolloid, with *no* restrictions on the diffusion coefficients of the ligands.

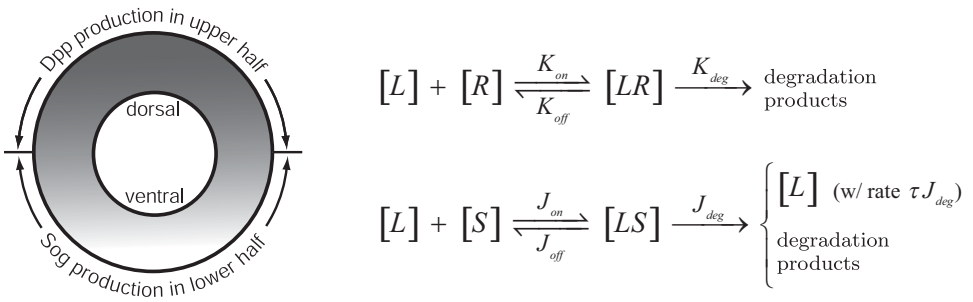


FIG. 1. *Cross section of a Drosophila embryo, and the reaction schemes with rate constants.*

To gain insight into the dependence of the morphogen activities on various biological parameters, we will obtain a perturbation solution of the steady state gradients with a biologically relevant assumption that the Sog production rate is high compared to that of Dpp [13], [14]. The solution indicates that the requirement for complete immobility of Dpp is not necessary for a biologically realistic Dpp-receptor gradient that is intense in Dpp activity at the dorsal midline. Finally, we will perform numerical simulations for the dynamics of the system. It is found that the cleavage of Sog-Dpp complex by Tolloid produces a transient peak of the Dpp-receptor concentration around the dorsal midline that is significantly stronger than the corresponding concentration at the steady state. The dependence of the peak on various biological parameters, including Sog production rate and diffusion coefficients, is also investigated. The overall features of the various concentrations of the model are consistent with experimental observations [5], [6], [7], [8]. A more complete model including more biological components and its comparison with new experiments on robustness of morphogen gradients will be presented in a separate paper [18].

2. Mathematical formulation. For an analytical and computational study of the biological phenomenon of interest, a system of partial differential equations and auxiliary conditions is formulated to capture the essential features of the dynamics of the two interacting morphogens. This approach was first applied to study the development of the *Drosophila* wing imaginal disc [19], [20], [4]. The three basic biological processes involving Dpp in the wing disc are diffusion for free Dpp molecules, their reversible binding with receptors, and degradation of the bound Dpp. The main purpose was to investigate the role of diffusion in the formation of a Dpp-receptor concentration gradient in the wing disc. That system was extended to include the effect of Sog on Dpp activities in a dorsal-ventral configuration [17] in an embryo, with the cleavage of Dpp-Sog complexes by Tolloid implicitly incorporated into the system through the complete recovery of Dpp after cleavage (while the Sog components degrade completely). The cleavage-recovery phenomenon has been suggested by previous experimental studies [21], [22]. Here we consider an even more general system than that in [17] by allowing fractional recovery through the fraction parameter τ, $0 \le \tau \le 1$, with $\tau = 1$ corresponding to complete recovery.

The setting for dorsal-ventral patterning in a *Drosophila* embryo during development is different and more complex than that considered in [4]. As shown in the sketch of the dorsal-ventral cross section of the embryo in Figure 1, Dpp is produced only in the dorsal region (with the rate $v_L(x)$), while Sog is produced only in the ventral region (with the rate $v_S(x)$). For a one-dimensional study of the dynamics of Sog and Dpp in the presence of cell receptors, we have idealized the dorsal-ventral an-

nular cross-section of the embryo as a ring and introduced an artificial cut of the ring at the ventral midline to map the cut ring onto the line segment $[-X_{\max}, X_{\max}]$, with $X = 0$ corresponding to the dorsal midline. Let [L], [S], [LS], [LR] denote the concentrations of Dpp, Sog, Dpp-Sog complexes, and Dpp bound to receptors, respectively. The first three diffuse with coefficients of diffusion D_L, D_S, and D_{LS}, respectively, while the concentration for the immobile and undegradable receptor is fixed at R_0 and uniformly distributed in $[-X_{\max}, X_{\max}]$. The system of equations governing the morphogen dynamics of such a system consists of the following four coupled second order differential equations, three of them being nonlinear partial differential equations (PDE) of the reaction-diffusion type:

$$\frac{\partial[\mathrm{L}]}{\partial T} = D_{\mathrm{L}}\frac{\partial^2[\mathrm{L}]}{\partial X^2} - K_{\mathrm{on}}[\mathrm{L}]\,(R_0 - [\mathrm{LR}]) - J_{\mathrm{on}}[\mathrm{L}][\mathrm{S}] + K_{\mathrm{off}}[\mathrm{LR}] + \big(J_{\mathrm{off}} + \tau J_{\mathrm{deg}}\big)[\mathrm{LS}] + v_L(X), \tag{1}$$

$$\frac{\partial[\mathrm{LR}]}{\partial T} = K_{\mathrm{on}}[\mathrm{L}]\,(R_0 - [\mathrm{LR}]) - \big(K_{\mathrm{off}} + K_{\mathrm{deg}}\big)[\mathrm{LR}], \tag{2}$$

$$\frac{\partial[\mathrm{LS}]}{\partial T} = D_{\mathrm{LS}}\frac{\partial^2[\mathrm{LS}]}{\partial X^2} + J_{\mathrm{on}}[\mathrm{L}][\mathrm{S}] - \big(J_{\mathrm{off}} + J_{\mathrm{deg}}\big)[\mathrm{LS}], \tag{3}$$

$$\frac{\partial[\mathrm{S}]}{\partial T} = D_{\mathrm{S}}\frac{\partial^2[\mathrm{S}]}{\partial X^2} - J_{\mathrm{on}}[\mathrm{L}][\mathrm{S}] + J_{\mathrm{off}}[\mathrm{LS}] + v_S(X) \tag{4}$$

for $-X_{\max} < X < X_{\max}$ and $T > 0$. The coefficients $\{K_{on}, J_{on}\}$, $\{K_{off}, J_{off}\}$, $\{K_{deg}, J_{deg}\}$ are the binding rate constants, the off rate constants, and the degradation rate constants of Dpp and Sog, respectively. With the morphogen activities symmetric about the ventral (as well as dorsal) midline, we must have the following symmetry (no flux) conditions at the two ends of the solution domain:

$$X = \pm X_{\max} : \frac{\partial[L]}{\partial X} = \frac{\partial[LS]}{\partial X} = \frac{\partial[S]}{\partial X} = 0.$$

The number of independent parameters may be reduced by suitable normalization. Let

$$x = \frac{X}{X_{max}}, \qquad t = \frac{D_0 T}{X_{max}^2}, \tag{5}$$

$$\{h_L, h_{LS}\} = \frac{X_{max}^2 R_0}{D_0}\{K_{on}, J_{on}\}, \tag{6}$$

$$\{f_L, f_{LS}, g_L, g_{LS}\} = \frac{X_{max}^2}{D_0}\{K_{off}, J_{off}, K_{deg}, J_{deg}\}, \tag{7}$$

$$\{V_L(x), V_S(x)\} = \frac{X_{max}^2}{R_0 D_0}\{v_L(X), v_S(X)\}, \tag{8}$$

$$\{\rho_L, \rho_S, \rho_{LS}\} = \left\{\frac{D_L}{D_0}, \frac{D_{LS}}{D_0}, \frac{D_S}{D_0}\right\}, \tag{9}$$

$$\{A, B, C, D\} = \left\{\frac{[\mathrm{L}]}{R_0}, \frac{[\mathrm{LR}]}{R_0}, \frac{[\mathrm{LS}]}{R_0}, \frac{[\mathrm{S}]}{R_0}\right\}. \tag{10}$$

In these relations, it would seem natural to choose the normalizing diffusion coefficient D_0 to be the maximum of the three diffusion coefficients. However, it turns out to be more appropriate to choose $D_0 = D_S$ to facilitate an appreciation of the implication of the solution. At this time, we will leave D_0 unspecified, but will see in section 4 why it is expeditious to specify it as D_S. In terms of these normalized quantities, (1)–(4) may be written as

$$A_{,t} = \rho_L A_{,xx} - h_L A(1-B) - h_{LS} AD + f_L B + (f_{LS} + \tau g_{LS})C + V_L(x), \tag{11}$$

$$B_{,t} = h_L A(1-B) - (f_L + g_L)B, \tag{12}$$

$$C_{,t} = \rho_{LS} C_{,xx} + h_{LS} AD - (f_{LS} + g_{LS})C, \tag{13}$$

$$D_{,t} = \rho_S D_{,xx} - h_{LS} AD + f_{LS} C + V_S(x) \tag{14}$$

for $-1 < x < 1$ and $t > 0$, with $(\ )_{,z} = \partial(\ )/\partial z$ for the temporal and spatial derivatives of the dependent variables A, B, C, D.

3. Existence of steady state solutions. In this section, we examine the existence of time-independent (or steady state) solutions of the system (11)–(14) subject to the no flux conditions at the two end points, which can now be written in terms of the normalized unknowns as

$$x = \pm 1: \quad A_{,x} = C_{,x} = D_{,x} = 0 \qquad (t > 0). \tag{15}$$

With the steady state solution independent of time, (12) becomes an algebraic equation and can be solved for B in terms of A:

$$B = \frac{A}{\alpha_L + A}, \qquad \alpha_L = \frac{g_L + f_L}{h_L}. \tag{16}$$

The expression for B is then used to eliminate it from (11), leaving the following three simultaneous equations for the three unknowns A, C, and D:

$$\rho_L A_{,xx} - \frac{g_L h_L A}{f_L + g_L + h_L A} - h_{LS} AD + (f_{LS} + \tau g_{LS})C + V_L = 0, \tag{17}$$

$$\rho_{LS} C_{,xx} + h_{LS} AD - (f_{LS} + g_{LS})C = 0, \tag{18}$$

$$\rho_S D_{,xx} - h_{LS} AD + f_{LS} C + V_S = 0 \tag{19}$$

for $-1 < x < 1$ subject to the boundary conditions (15).

Throughout this section we assume the following:

(A1) f_L, f_{LS}, g_L, g_{LS}, h_L, and h_{LS} are continuous positive functions in $[-1, 1]$; ρ_L, ρ_{LS}, and ρ_S are positive constants; V_S, V_L are nonnegative integrable functions that satisfy $\int_{-1}^{1} V_L > 0$ and $\int_{-1}^{1} V_S > 0$; and τ is a constant satisfying $0 \le \tau \le 1$.

If $V_L(x)$ and $V_S(x)$ are continuous, we seek a classical solution of (15)–(19); i.e., A, C, and D are twice continuously differentiable in $[-1, 1]$ that satisfy (15)–(19).

THEOREM 3.1. *Suppose that* (A1) *holds and* V_L, V_S *are continuous in* $[-1,1]$. *Then* (15)–(19) *has a positive classical solution if and only if both of the following inequalities hold:*

$$\int_{-1}^{1} V_L(x)dx > (1-\tau)\int_{-1}^{1} V_S(x)dx, \tag{20}$$

$$\int_{-1}^{1} V_L(x)dx < (1-\tau)\int_{-1}^{1} V_S(x)dx + \int_{-1}^{1} g_L(x)dx. \tag{21}$$

Since $\int_{-1}^{1} V_L(x) > 0$, the first condition is trivially satisfied for $\tau = 1$ (full recovery of Dpp), and the second is a distributed version of the necessary and sufficient condition for existence in [9], [10], [11], [12] (that the Dpp production rate must be slower than the degradation rate of the Dpp-receptor complexes). For $0 \le \tau < 1$, these two conditions may be combined to give a similar condition on a nonnegative "effective" Dpp production rate $[V_L - (1-\tau)V_S]$ (see section 5).

LEMMA 3.2. *If* (15)–(19) *has a positive classical solution, then* (20) *and* (21) *must hold.*

Proof. Adding up (17) and (18) and integrating over $[-1,1]$, we obtain with the help of (15)

$$\int_{-1}^{1} V_L = \int_{-1}^{1} \frac{g_L h_L A}{f_L + g_L + h_L A} + (1-\tau)\int_{-1}^{1} g_{LS}C. \tag{22}$$

Similarly, adding up (18) and (19) and integrating over $[-1,1]$, we get

$$\int_{-1}^{1} g_{LS}C = \int_{-1}^{1} V_S. \tag{23}$$

It follows from (22) and (23) that

$$\int_{-1}^{1} V_L = \int_{-1}^{1} \frac{g_L h_L A}{f_L + g_L + h_L A} + (1-\tau)\int_{-1}^{1} V_S. \tag{24}$$

For $A > 0$ in $[-1,1]$, we have

$$0 < \int_{-1}^{1} \frac{g_L h_L A}{f_L + g_L + h_L A} < \int_{-1}^{1} g_L, \tag{25}$$

which along with (24) implies (20)–(21). □

In view of Lemma 3.2, we'll assume that (20)–(21) holds for the rest of this subsection. Our goal is to show that if V_L and V_S are continuous, then the condition (20)–(21) implies that (15)–(19) has at least a positive classical solution. The idea is to introduce some parameter λ and consider the following system of equations:

$$\rho_L \tilde{A}_{,xx} + \lambda F_1(x, \tilde{A}, \tilde{C}, \tilde{D}) = 0, \quad -1 < x < 1, \tag{26}$$

$$\rho_{LS} \tilde{C}_{,xx} + \lambda F_2(x, \tilde{A}, \tilde{C}, \tilde{D}) = 0, \quad -1 < x < 1, \tag{27}$$

$$\rho_S \tilde{D}_{,xx} + \lambda F_3(x, \tilde{A}, \tilde{C}, \tilde{D}) = 0, \quad -1 < x < 1, \tag{28}$$

$$\tilde{A}_{,x} = \tilde{C}_{,x} = \tilde{D}_{,x} = 0 \quad \text{at } x = -1, 1, \tag{29}$$

where $\lambda \in (0, 1]$ and F_i $(i = 1, 2, 3)$ is given by

$$F_1(x, \tilde{A}, \tilde{C}, \tilde{D}) = -\frac{g_L h_L \tilde{A}}{f_L + g_L + h_L \tilde{A}} - h_{LS} \tilde{A} \tilde{D} + (f_{LS} + \tau g_{LS}) C + V_L, \tag{30}$$

$$F_2(x, \tilde{A}, \tilde{C}, \tilde{D}) = h_{LS} \tilde{A} \tilde{D} - (f_{LS} + g_{LS}) \tilde{C}, \tag{31}$$

$$F_3(x, \tilde{A}, \tilde{C}, \tilde{D}) = -h_{LS} \tilde{A} \tilde{D} + f_{LS} \tilde{C} + V_S. \tag{32}$$

We establish some a priori estimates for nonnegative classical solutions of (26)–(29).

LEMMA 3.3. *Let $(\tilde{A}, \tilde{C}, \tilde{D})$ be any nonnegative classical solution of* (26)–(29). *If $\lambda > 0$, then $\tilde{A}(x) > 0$, $\tilde{C}(x) > 0$, and $\tilde{D}(x) > 0$ for every $x \in [-1, 1]$.*

Proof. Similar to (23) we have $\int_{-1}^{1} g_{LS} \tilde{C} = \int_{-1}^{1} V_S$. Hence $\tilde{C} \geq 0$, $\tilde{C} \not\equiv 0$. By (27) and (31) we have

$$-\rho_{LS} \tilde{C}_{,xx} + \lambda (f_{LS} + g_{LS}) \tilde{C} = \lambda h_{LS} \tilde{A} \tilde{D} \geq 0, \quad -1 < x < 1. \tag{33}$$

This together with $\tilde{C}_{,x}(-1) = \tilde{C}_{,x}(1) = 0$, via the maximum principle [23], implies that $\tilde{C}(x) > 0$ for every $x \in [-1, 1]$. Since $V_L \not\equiv 0$ and $V_S \not\equiv 0$, similarly by (26)–(29) and the maximum principle we can show that $\tilde{A} > 0$ and $\tilde{D} > 0$ in $[-1, 1]$. ∎

LEMMA 3.4. *There exists some constant $M > 0$, independent of λ, such that for any $0 < \lambda \leq 1$ and any positive classical solution $(\tilde{A}, \tilde{C}, \tilde{D})$ of* (26)–(29) *we have*

$$\|\tilde{A}\|_{L^\infty} + \|\tilde{C}\|_{L^\infty} + \|\tilde{D}\|_{L^\infty} \leq M. \tag{34}$$

The proof of Lemma 3.4 is postponed to the appendix. Lemmas 3.3 and 3.4 enable us to define Leray–Schauder degree (see, e.g., [24]) for a certain operator whose fixed points correspond to positive solutions of (26)–(29).

Set $E = \{C[-1, 1]\}^3$ and $C_N^2[-1, 1] = \{u \in C^2[-1, 1] : u_{,x}(-1) = u_{,x}(1) = 0\}$. For any positive constant γ, let L_γ^{-1} denote the inverse of the operator $L_\gamma := -\gamma \frac{d^2}{dx^2} + I : C_N^2[-1, 1] \to C[-1, 1]$, where I denotes the identity map from $C[-1, 1]$ to itself.

For every $\lambda \in [0, 1]$, define operator $T(\lambda) : \quad E \to E$ by

$$T(\lambda)(\tilde{A}, \tilde{C}, \tilde{D}) = \begin{pmatrix} L_{\rho_L}^{-1}[\tilde{A} + \lambda F_1^+(x, \tilde{A}, \tilde{C}, \tilde{D})] \\ L_{\rho_{LS}}^{-1}[\tilde{C} + \lambda F_2(x, \tilde{A}, \tilde{C}, \tilde{D})] \\ L_{\rho_S}^{-1}[\tilde{D} + \lambda F_3(x, \tilde{A}, \tilde{C}, \tilde{D})] \end{pmatrix}, \tag{35}$$

where

$$F_1^+(x, \tilde{A}, \tilde{C}, \tilde{D}) = \frac{-g_L h_L A}{f_L + g_L + h_L A_+} - h_{LS} A D + (f_{LS} + \tau g_{LS}) C + V_L, \tag{36}$$

$A_+ = \max(A, 0)$. By standard regularity theory and the embedding theorem, we see that $T(\lambda)$ is well defined and continuous, and the operator $\tilde{T}$: $[-1, 1] \times E \to E$,

defined by $\tilde{T}(\lambda, \tilde{A}, \tilde{C}, \tilde{D}) = T(\lambda)(\tilde{A}, \tilde{C}, \tilde{D})$, is continuous and compact. For M given in (34), define

$$\Omega = \left\{ (\tilde{A}, \tilde{C}, \tilde{D}) \in E : \ 0 < \tilde{A}(x), \tilde{C}(x), \tilde{D}(x) < M + 1 \quad \forall x \in [-1, 1] \right\}.$$

Ω is an open and bounded subset of E. By Lemmas 3.3 and 3.4, we see that for any $\lambda \in (0, 1]$, $[I - T(\lambda)]^{-1} \{(0, 0, 0)\} \notin \partial\Omega$. Hence the Leray–Schauder degree, $\deg\big(I - T(\lambda), \Omega, (0, 0, 0)\big)$, is well defined for $0 < \lambda \leq 1$. Moreover, by the homotopy invariance of the Leray–Schauder degree [24], $\deg\left(I - T(\lambda), \Omega, (0, 0, 0)\right)$ is a constant function for $0 < \lambda \leq 1$. To complete the proof of Theorem 3.1, we need the following result.

PROPOSITION 3.5. *There exists* $\delta > 0$ *such that* $\deg\left(I - T(\lambda), \Omega, (0, 0, 0)\right) = 1$ *for* $\lambda \in (0, \delta)$.

The detail of the proof of this proposition is not particularly relevant to the proof of Theorem 3.1 and will be given in an appendix of this paper. Assuming Proposition 3.5, we can now complete the proof of Theorem 3.1.

Proof of Theorem 3.1. By Lemma 3.2, it suffices to establish the sufficiency part. By Proposition 3.5, for every $0 < \lambda \leq 1$, $\deg\left(I - T(\lambda), \Omega, (0, 0, 0)\right) = 1$. In particular, $\deg\left(I - T(1), \Omega, (0, 0, 0)\right) \neq 0$. This implies that there exists $(\tilde{A}, \tilde{C}, \tilde{D}) \in \Omega$ such that $(I - T(1))(\tilde{A}, \tilde{C}, \tilde{D}) = (0, 0, 0)$. By standard regularity theory we see that $\tilde{A}, \tilde{C}, \tilde{D} \in C^2[-1, 1]$ and is thus a positive classical solution of (15)–(19). ∎

Specific morphogen systems of interest include those with morphogen production rates that are discontinuous in the spatial variable (see section 4). When V_L and V_S are bounded and measurable, we will be seeking $C^{1,1}$ solutions of (15)–(19), i.e., functions A, C, D that are differentiable in $[-1, 1]$; have derivatives $A_{,x}, C_{,x}$, and $D_{,x}$ Lipschitz continuous in $[-1, 1]$; and satisfy (15) and for every $x \in [-1, 1]$

$$\rho_L A_{,x} + \int_{-1}^{x} F_1 = \rho_{LS} C_{,x} + \int_{-1}^{x} F_2 = \rho_S D_{,x} + \int_{-1}^{x} F_3 = 0. \tag{37}$$

THEOREM 3.6. *Suppose that* (A1) *holds and that* V_L *and* V_S *are bounded measurable. Then* (15)–(19) *has a positive* $C^{1,1}$ *solution if and only if* (20)–(21) *holds.*

Proof. Suppose that (15)–(19) has a positive $C^{1,1}$ solution. Setting $x = 1$ in (37) and applying the same argument as in the proof of Lemma 3.2, we see that (20)–(21) must hold. On the other hand, if (20)–(21) holds, we can choose a uniformly bounded sequence of continuous positive functions $V_L^n(x)$ and $V_S^n(x)$ such that $V_L^n(x) \to V_L$ and $V_S^n(x) \to V_S$ a.e., and $0 < \int_{-1}^{1} \left[V_L^n - (1 - \tau) V_S^n\right] < \int_{-1}^{1} g_L$. By Theorem 3.1 (17)–(19), with V_L and V_S being replaced by V_L^n and V_S^n, respectively, there is a sequence of positive classical solutions, denoted by A^n, C^n, and D^n. As for Lemma 3.4, we can show that there exists some positive constant M, independent of n, such that $\|A^n\|_{L^\infty} + \|C^n\|_{L^\infty} + \|D^n\|_{L^\infty} \leq M$. Furthermore, $\|A^n_{,xx}\|_{L^\infty}$, $\|C^n_{,xx}\|_{L^\infty}$, and $\|C^n_{,xx}\|_{L^\infty}$ are uniformly bounded. By passing to a subsequence if necessary, (A^n, C^n, D^n) converge to some functions (A, C, D) in C^1, and A, C, D satisfy (15) and are nonnegative solutions of (37). From (37) we see that $A_{,x}, C_{,x}, D_{,x}$ are Lipschitz continuous in $[-1, 1]$. By similar argument as in Lemma 3.3 (but instead using the maximum principle for weak solutions of (15)–(19)), we see that A, C, D are all positive in $[-1, 1]$. This completes the proof of Theorem 3.6. ∎

Remark 3.7. Note that $C \in C^2[-1, 1]$. If V_L and V_S are piecewise continuous, then A and D are also piecewise twice continuously differentiable in $[-1, 1]$.

4. Approximate steady state solutions for $V_L \ll V_S$. In previous studies [13], [14], the constant (in both space and time) Dpp production rate, $\bar{v}_L$, in the dorsal

region was estimated to be significantly smaller than the constant Sog production rate, $\bar{v}_S$, in the ventral region. In [14], the ratio of the two production rates, defined as $\epsilon \equiv \bar{v}_L/\bar{v}_S$, is 0.008 for its baseline study. The robustness of the solutions with respect to variations of $\bar{v}_S$ is studied for fixed $\bar{v}_L$ [13].

For $\bar{v}_L \ll \bar{v}_S$, so that $\epsilon \ll 1$, we obtain below a perturbation solution for the steady state of (15)–(19), with $\tau = 1$ for simplicity. For $\tau < 1$, perturbation solution procedure applies only if (20)–(21) hold. Similar to [14], we assume

$$V_L(x) = \bar{V}_L H\left(\frac{1}{2} - x\right), \qquad V_S(x) = \bar{V}_S H\left(x - \frac{1}{2}\right), \tag{38}$$

where $(\bar{V}_L, \bar{V}_S) = (\bar{v}_L, \bar{v}_S)X_{max}^2/(R_0 D_0)$ and $H(z)$ is the unit step function.

With $\bar{V}_L \ll \bar{V}_S$, we expect $D(x), C(x) = O(\bar{V}_S), O(\bar{V}_S)$, although the latter may be a smaller fraction of $\bar{V}_S$. On the other hand, we have $A(x) = O(\bar{V}_L)$ at most, in fact quite a bit smaller since free Dpp should eventually be bound to Sog or receptors, given that Sog is produced at a much higher rate. For these reasons, we set

$$A(x) = \frac{\overline{V}_L}{\mu_L^2} a(x), \quad C(x) = \frac{\overline{V}_S}{f_{LS} + g_{LS}} c(x), \quad D(x) = \overline{V}_S d(x), \tag{39}$$

where $\mu_L^2 = g_L/\alpha_L$ and $\alpha_L = (f_L + g_L)/h_L$. Then (17)–(19) become

$$\frac{\overline{V}_L}{\overline{V}_S}\left[\frac{\rho_L}{\mu_L^2} a'' - \frac{a}{1+\beta_L a} + H\left(\frac{1}{2} - x\right)\right] - \mu_D^2 ad + c = 0, \tag{40a}$$

$$\rho_{LS} c'' + (f_{LS} + g_{LS})[\mu_D^2 ad - c] = 0, \tag{40b}$$

$$\rho_S d'' - [\mu_D^2 ad - c] - (1 - \sigma_{LS})c + H\left(x - \frac{1}{2}\right) = 0, \tag{40c}$$

where $()' = d()/dx$, $\beta_L = \bar{V}_L/g_L$, $\sigma_{LS} = f_{LS}/(f_{LS}+g_{LS}) < 1$, and $\mu_D^2 = h_{LS}\alpha_L \bar{V}_L/g_L$. Using symmetry about $x = 0$, we need only to consider solutions for $0 < x < 1$ with the boundary conditions at $x = 0$ being again no flux for all three unknowns a, b, and c.

The form of (40a)–(40c) suggests that we seek a perturbation solution of $\{a, c, d\}$ in ε:

$$\{a(x;\epsilon), c(x;\epsilon), d(x;\epsilon)\} = \sum_{n=0}^{\infty} \{a_n(x), c_n(x), d_n(x)\}\, \epsilon^n. \tag{41}$$

For moderate values of $\bar{V}_L$ so that μ_D^2 is *not* small compared to unity, the three leading term coefficients are determined by

$$\mu_D^2 a_0 d_0 - c_0 = 0, \tag{42a}$$

$$\rho_{LS} c_0'' + (f_{LS} + g_{LS})[\mu_D^2 a_0 d_0 - c_0] = 0, \tag{42b}$$

$$\rho_S d_0'' - [\mu_D^2 a_0 d_0 - c_0] - (1 - \sigma_{LS})c_0 + H\left(x - \frac{1}{2}\right) = 0. \tag{42c}$$

The complementary case, $\mu_D^2 \ll 1$, can also be analyzed but is not relevant for our biological system.

Upon combining (42a) and (42b) we get

$$\rho_{LS} c_0'' = 0. \tag{43}$$

The no flux boundary conditions at $x = 0, 1$ require $c_0(x) \equiv \sigma_0$ for some constant σ_0. To determine σ_0, we note that (23) is still valid and requires

$$\frac{1}{2}\overline{V}_S = \int_0^1 V_s(x)dx = g_{LS}\int_0^1 C(x)dx = \overline{V}_S \frac{g_{LS}}{f_{LS}+g_{LS}}\int_0^1 c(x)dx \tag{44}$$

so that $\sigma_0 = 1/2(1-\sigma_{LS})$, i.e.,

$$(1-\sigma_{LS})c_0(x) = \frac{1}{2}, \qquad 0 \le x \le 1. \tag{45}$$

To determine $d_0(x)$, we use (42a) and (42c) to obtain

$$\rho_S d_0'' - (1-\sigma_{LS})c_0 + H\left(x-\frac{1}{2}\right) = 0. \tag{46}$$

Upon integration and application of boundary conditions at $x = 0, 1$, as well as the continuity condition at $x = 1/2$ for d_0, we obtain

$$\rho_S d_0(x) = \begin{cases} \delta_0 + \dfrac{x^2}{4} & \left(x \le \dfrac{1}{2}\right), \\ \delta_0 - \dfrac{1}{8} + \dfrac{1}{2}\left(x - \dfrac{x^2}{2}\right) & \left(x \ge \dfrac{1}{2}\right), \end{cases} \tag{47}$$

where δ_0 is an undetermined constant. By (42a) we have also

$$\frac{1-\sigma_{LS}}{\rho_S}\mu_D^2 a_0(x) = \frac{1-\sigma_{LS}}{\rho_S}\frac{c_0(x)}{d_0(x)} = \begin{cases} \dfrac{1}{2}\dfrac{1}{(\delta_0 + \frac{1}{4}x^2)} & \left(x < \dfrac{1}{2}\right), \\ \dfrac{1}{2}\dfrac{1}{[\delta_0 - \frac{1}{8} + \frac{1}{2}x - \frac{1}{4}x^2]} & \left(x > \dfrac{1}{2}\right). \end{cases} \tag{48}$$

It is rather fortuitous to have $a_0'(0) = a_0'(1) = 0$ because d_0 and c_0 satisfy no flux conditions at the two end points so that there are no boundary layers adjacent to the two ends.

It remains to determine δ_0. We note that (24) still holds, particularly when $\tau = 1$. In that case, (24) becomes

$$G(\delta_0) \equiv \int_0^1 \frac{a_0(x)}{1+\beta_L a_0(x)}dx = \frac{1}{2}. \tag{49}$$

It is easy to see that $G(\delta_0)$ is strictly monotone decreasing in δ_0 and that $G(\delta_0) \to 0$ as $\delta_0 \to \infty$. Hence $G(\delta_0) = \frac{1}{2}$ has at most one positive root, and it has one positive root if and only if $G(0) > \frac{1}{2}$. Note that $G(0)$ can be explicitly computed, and thus $G(\delta_0) = \frac{1}{2}$ determines δ_0.

Altogether, we have as the corresponding leading terms for the concentrations

$$A(x) \sim \frac{(1-\sigma_{LS})\mu_D^2 a_0(x)/\rho_S}{R_0 J_{on,eff}/(D_S/X_{\max}^2)}, \tag{50}$$

$$B(x) \sim \frac{\Gamma_{LS}(1-\sigma_{LS})\mu_D^2 a_0(x)/\rho_S}{1+\Gamma_{LS}(1-\sigma_{LS})\mu_D^2 a_0(x)/\rho_S}, \tag{51}$$

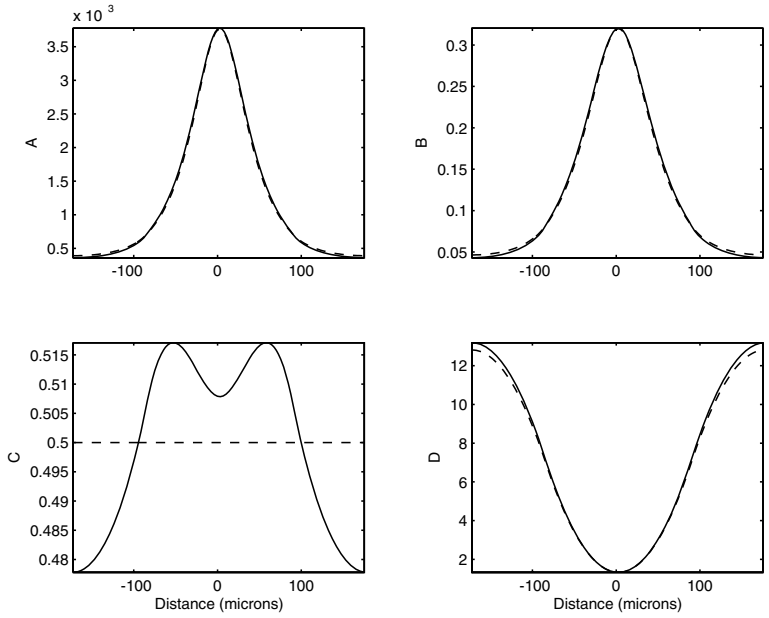


FIG. 2. *Comparisons between the numerical steady states (solid lines) and the perturbation solutions (dashed lines). The parameters are* $\bar{v}_L = 8 \times 10^{-4} s^{-1} \mu M$, $K_{on} = 0.4 s^{-1}$, $K_{off} = 4 \times 10^{-6} s^{-1}$, $K_{deg} = 3.2 \times 10^{-3} s^{-1}$, $\bar{v}_S = 6 \times 10^{-2} s^{-1} \mu M$, $J_{on} = 6 s^{-1} \mu M$, $J_{off} = 10^{-5} s^{-1}$, $J_{deg} = 6 \times 10^{-2} s^{-1}$, $\tau = 1$.

$$C(x) \sim \frac{1}{2} \frac{\bar{v}_S}{J_{deg} R_0}, \tag{52}$$

$$D(x) \sim \frac{\bar{v}_S / R_0}{D_S / X_{\max}^2} [\rho_S d_0(x)], \tag{53}$$

where

$$K_{on,eff} \equiv \frac{K_{deg} K_{on}}{K_{deg} + K_{off}}, \qquad J_{on,eff} \equiv \frac{J_{deg} J_{on}}{J_{deg} + J_{off}}, \qquad \Gamma_{LS} = \frac{K_{on,eff}}{J_{on,eff}} \frac{D_S}{K_{deg} X_{max}^2}. \tag{54}$$

In Figure 2, the perturbation solutions (50)–(53) are plotted against the numerical solutions obtained through temporal evolution (which will be discussed in the next section). The relative difference between the two solutions is 1.5% for A, 1.4% for B, 4.3% for C, and 2.9% for D for $\epsilon = \bar{v}_L / \bar{v}_S = 0.0133$ and $\mu_D^2 = 18.4$. This illustrates the approximation and accuracy of the perturbation solution for $\epsilon \ll 1$.

More interesting is the dependence of the leading term solutions (51)–(53) on the biological parameters. The simplest of the four is the uniformly distributed concentration of Dpp-Sog complexes in (53): it depends only on the production rate of Sog per receptor, which is uniform in the ventral region. Free Sog $D(x)$ is proportional to the quadratic function defined in (47) with a magnitude of $\bar{v}_S / R_0$ modified by the diffusion coefficient of Sog. That $D(x)$ is inversely proportional to D_S is not surprising, since faster diffusion of Sog would move more of it into the dorsal region for

binding with the available Dpp there. Note that $\rho_S d_0(x)$ is independent of the choice of normalizing diffusion coefficient D_0 and the effects of all biological parameters are felt by $\rho_S d_0(x)$ only implicitly through the parameter δ_0.

Less expected is the dependence of $A(x)$ and $B(x)$ on the biological parameters. From (51), we see that if $\Gamma_{LS} = O(1)$, the amplitude of $B(x)$ is determined mainly by Γ_{LS}. For $\Gamma_{LS} \gg 1$, we have $B(x) \sim 1$, except possibly for a region adjacent to the ventral midline $x = 1$. In either case, the amplitude of $B(x)$ does not depend explicitly on either of the two production rate parameters $\bar{v}_S/R_0$ or $\bar{v}_L/R_0$; the effects of these two parameters on $B(x)$ are felt only through δ_0.

The situation is similar for $A(x)$. It seems unreasonable that $A(x)$ does not tend to zero with $\bar{v}_L/R_0$ (with the same observation applied to $B(x)$ as well). However, we see from a closer examination of (40a) that $\mu_D^2 = h_{LS}\alpha_L \bar{V}_L/g_L$ tends to zero with $\bar{v}_L/R_0$. For sufficiently small $\bar{v}_L/R_0$, the first approximation relation (42a) would give $c_0(x) = 0$. In that case, $c(x)$ should be rescaled (by an additional factor μ_D^2) for a proper perturbation solution, while the solution of this section ceases to be applicable. In other words, to apply the perturbation solution $\{a_0(x), c_0(x), d_0(x)\}$ obtained above, we must have $\bar{v}_L/R_0$ sufficiently small so that $\bar{V}_L/\bar{V}_S = \bar{v}_L/\bar{v}_S \ll 1$ but not too small so that $\mu_D^2 = h_{LS}\alpha_L \bar{V}_L/g_L$ is *not* small compared to unity.

5. Numerical solutions for evolutions. The system (1)–(4) can be solved by finite difference schemes [25]. The diffusion terms are approximated by the second order central difference. The temporal evolution is approximated through the fourth order Adams–Moulton predictor-corrector method. The overall accuracy for the method is second order in space and fourth order in time.

For a typical calculation, the time step is chosen to be $\triangle t = 2 \times 10^{-4}$ seconds, and the number of points to discretize the entire dorsal and ventral region is $N = 64$. Smaller time step and larger number of points have been used to check the accuracy and convergence of the calculations.

Similar to [13], the span of both the dorsal region and the ventral region is chosen to be 175μm, i.e., $X_{max} = 175\mu$m. Unlike [13], the diffusion constants for Dpp, Sog, and Dpp-Sog are taken to be the same with $D_0 = D_L = D_{LS} = D_S = 20\mu\text{m}^2/\text{second}$ [4], so that $\rho_L = \rho_S = \rho_{LS} = 1$ (except for changes indicated in Figures 7 and 8). In this study, the synthesis rates for Dpp and Sog remain the same for all time. In particular, $v_L(X)$ is always chosen to be a nonzero constant, $\bar{v}_L$, in the dorsal region and zero in the ventral region, while $v_S(X)$ is the opposite, with $v_S(X) = \bar{v}_S$ in the ventral and zero in the dorsal region.

The dynamics of the system without Sog is very similar to that in [4], even though the ligand is produced from a localized source in [4] while the ligand is produced in the whole dorsal region for the system (1)–(4). For realistic ranges of the biological parameters of the problem, this system typically evolves quickly and monotonically to a steady state within a half hour, with the Dpp-receptor concentration almost uniform around the dorsal region. This behavior is consistent with the experimental observation of [8]. At $x = 0$ the steady state is approximately equal to $\bar{v}_L/(K_{deg}R_0)$.

Without Sog, the solution at any fixed x is found to be an increasing function of time. This feature is also observed for cases where Sog is synthesized at a slow rate or at a rate comparable to the Dpp production rate. The situation is different if the Sog production rate is significantly larger than the Dpp production rate, which is the most biologically relevant case [13]. In Figure 3, time evolution of a typical system for large $\bar{v}_S$ is plotted. It is observed that the spatial distributions of Dpp and the Dpp-receptor complex continue to have maximum concentrations at the middle of the

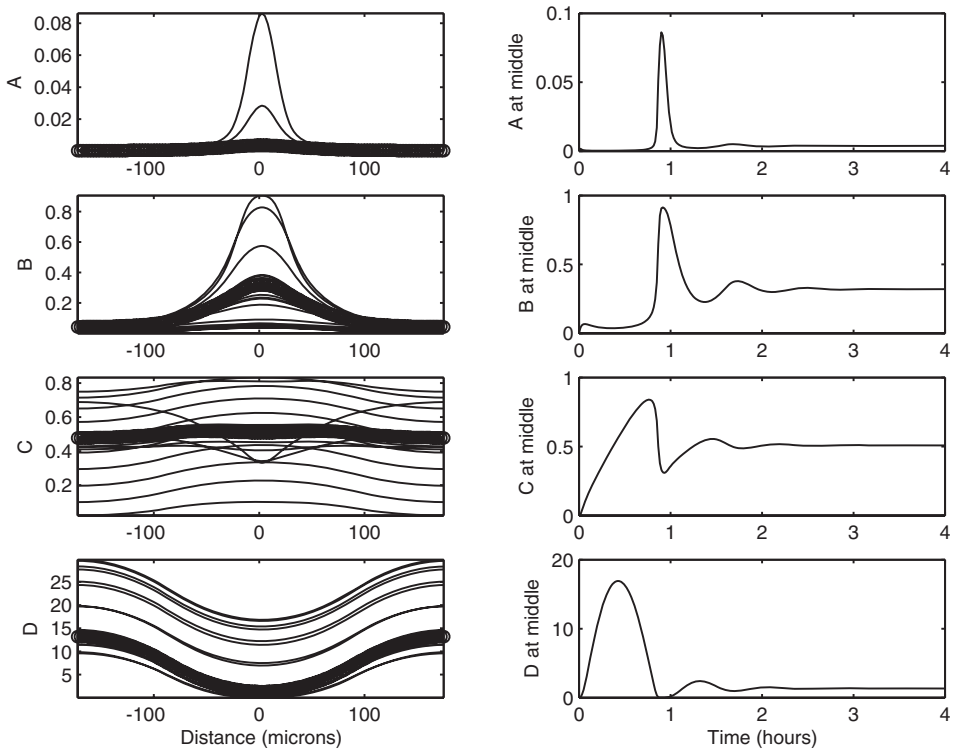


FIG. 3. *The dynamics of solutions with SOG at every* 5 *minutes; o in the left-hand panels marks the steady-state solutions. All parameters are the same as for Figure* 2.

dorsal region, $x = 0$, at any instance in time (see the left-hand panels). However, the various morphogen concentrations at $x = 0$ (the center of the dorsal region) peak at an early time, then oscillate, with the amplitude of oscillations decaying until the concentrations reach their steady state (see the right-hand panels). Therefore we record two interesting curves for Dpp-receptor concentration: the transient solution with the largest value at the dorsal midline and the steady state solution.

In Figure 4(a), the steady state for Dpp-receptor concentration of our system (from the same numerical simulations for Figure 3) are plotted. With Sog ($\bar{v}_S \neq 0$), the Dpp-receptor in the dorsal region generally has sharper gradient and larger concentration than those without Sog ($\bar{v}_S = 0$). For the transient solution at its maximal peak magnitude, the concentration with Sog is at least double that without Sog around the middle region. These are consistent with the experimental observation in [8].

In steady state, the system with or without Sog has the same total amount of Dpp-receptor complex for $\tau = 1$. This can be shown by simply adding the right-hand sides of (11)–(13) and (13)–(14), respectively, and then integrating them through the whole domain:

$$\int_{-1}^{1} g_L B dx = \int_{-1}^{1} \left(V_L(x) - (1-\tau)V_S(x)\right) dx. \tag{55}$$

This relationship is independent of the presence of Sog when $\tau = 1$. In other words, the effect of inhibitor on Dpp-receptor concentration in the steady state is a spatial

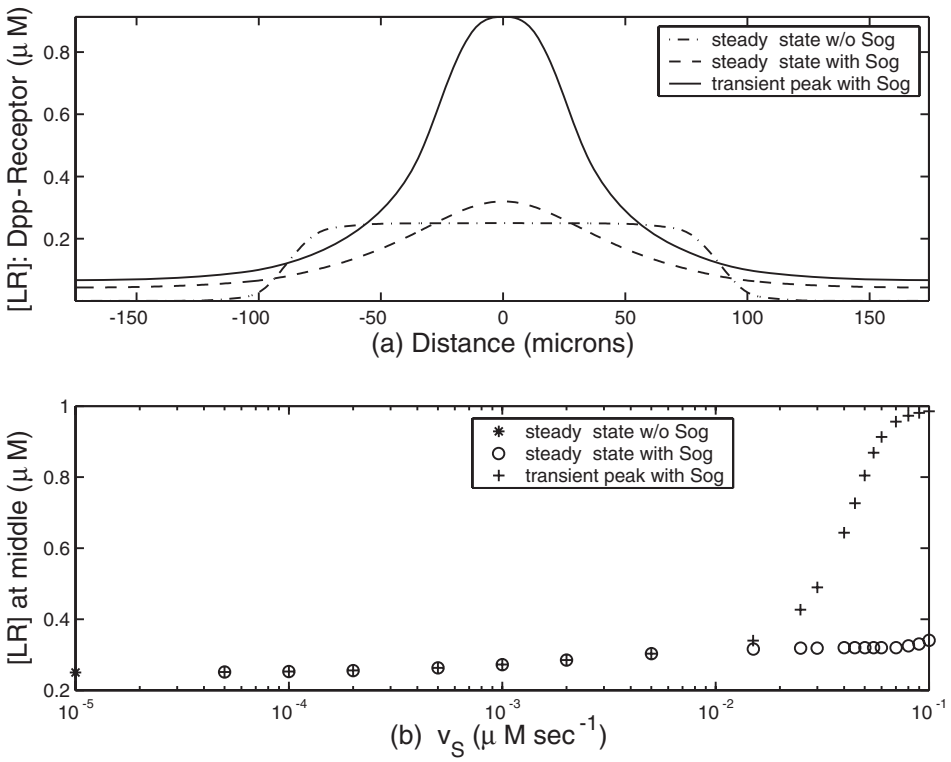


FIG. 4. *Effect of Sog on the transient and steady state solutions.* (a) $[LR]$ *as a function of space;* (b) $[LR]$ *at dorsal midline as a function of* $\bar{v}_S$. *Parameters are as in Figure* 3.

redistribution, not an increase or decrease in total concentration aggregated over the entire embryo if all degraded Dpp-Sog complexes, $[LS]$, are cleaved to free up Dpp and degrade only the Sog component.

For the transient solution, the presence of Sog clearly helps build up the Dpp-receptor complexes in terms of both gradient and concentration, as shown in Figure 3. In Figure 4(b), we study how the transient peak of Dpp-receptor and the steady state at the dorsal midline ($x = 0$) depend on $\bar{v}_S$. The steady state for B without Sog at $x = 0$ is 0.25, and its value is plotted at the y-axis in Figure 4(b). For a small amount of Sog, the transient peaks are not high, and the steady state has the largest value at $x = 0$, as shown for $\bar{v}_S/R_0 < 0.01$. Also, $B(x = 0)$ at steady state increases as $\bar{v}_S$ increases, and the transient peak begins to deviate from the steady state around $\bar{v}_S/R_0 = 0.01$. As $\bar{v}_S$ increases by one order of magnitude from 0.015 to 0.1, the transient peak increases from 0.34 to 0.99, while the steady state only from 0.32 to 0.34. Once $\bar{v}_S/R_0$ becomes large enough, the variation of the transient peak is more sensitively dependent on variation of $\bar{v}_S/R_0$ than that of the steady state at $x = 0$. The dependence of the transient peak on other parameters such as J_{on} and J_{deg} have been investigated previously in [17] for $\tau = 1$.

When $\tau < 1$, so that only a portion of the degraded Dpp-receptor complex is cleaved to free up Dpp, the dynamics of the system strongly depends on the size of τ when the steady state condition (20)–(21) holds. It is not surprising that for smaller τ, i.e., less free Dpp released from the degraded $[LS]$, the transient and steady

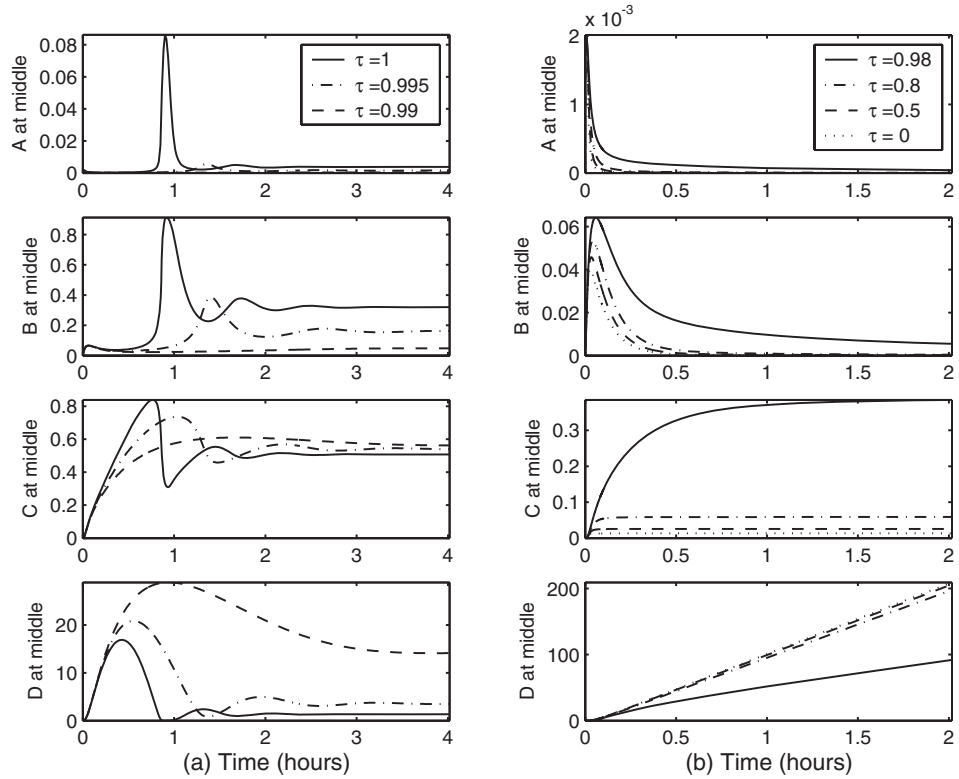


FIG. 5. *Effect of τ on the steady state solutions. Parameters are as in Figure* 3 *except for τ.* (a) *Cases with steady states;* (b) *cases without steady states.*

peaks of Dpp-receptor complex are lower, as shown in Figure 5(a) for $\tau = 0.995$ and 0.99. However, for $\tau = 0.99$, the concentration of Dpp-receptor complex around the dorsal region is much lower with Sog than without Sog, as shown in Figure 5(a). As demonstrated in (55), a small change of τ will result in a large change of $[LR]$ at steady state for a large $\bar{v}_S$, which is the case for Figure 5(a). In essence, $v_{eff} \equiv \bar{v}_L - (1-\tau)\bar{v}_S$ can be regarded as an effective production rate for Dpp.

When the effective production rate v_{eff} becomes negative, that is, the condition (20)–(21) does not hold, then the system can no longer sustain a steady state. For this situation, the concentrations of both free Dpp and the Dpp-receptor complex are typically very low, and the Dpp-receptor complex reaches the peak before Sog diffuses into the dorsal region from the ventral side and takes over the reaction with Dpp. With the availability of a large amount of Sog and its fast association rate with Dpp, Dpp-Sog reaction dominates. It is interesting to note in Figure 5(b) that as τ varies from 0.98 to 0, the time for Dpp and Dpp-receptor complex to reach their peaks barely changes. This critical time (to reach the peak) is mainly determined by the coefficient of diffusion D_S, which controls the speed of Sog movement into the dorsal zone.

In [14] (hence also in [13]), degradation for $[LR]$ is not allowed in the system ($K_{deg} = 0$); therefore the condition (20)–(21) does not hold for any $\bar{v}_L > 0$. In order to achieve steady state in [13], [14], the models there turned off production of Dpp after 10 minutes ($T^* = 10$ minutes). The effect of the choice of T^* and

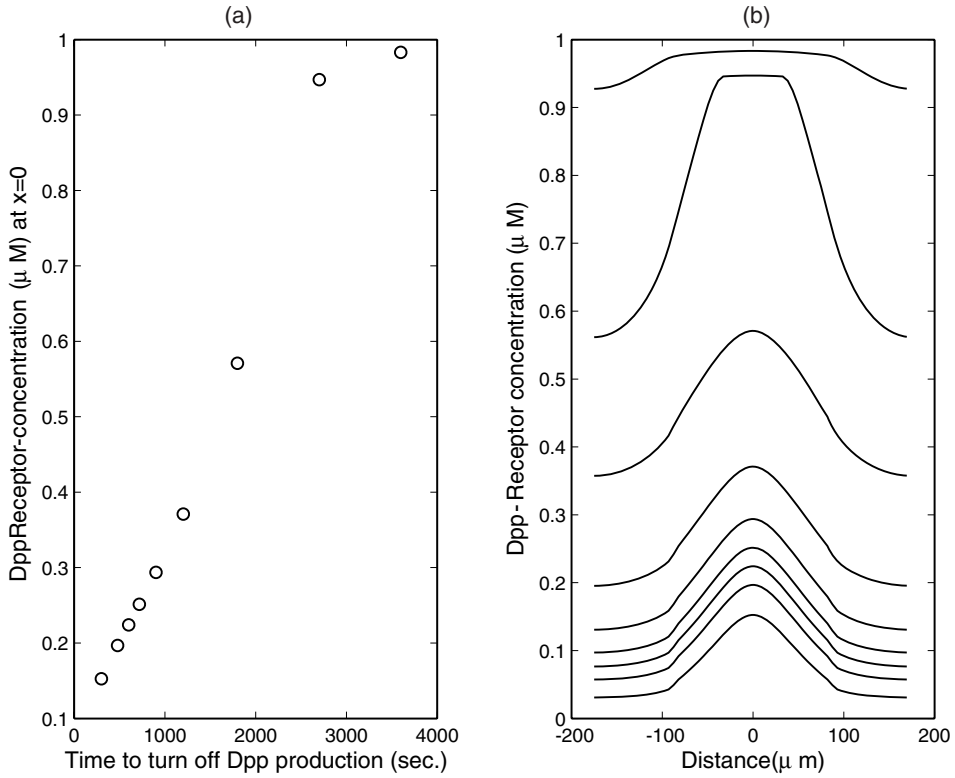


FIG. 6. *$\bar{v}_L$ is set to zero at different times.* (a) *$[LR]$ at the dorsal midline at steady state as a function of the time for turning off $\bar{v}_L$;* (b) *$[LR]$ at steady states as a function of space for different times of turning off $\bar{v}_L$, as shown in* (a). *Other parameters are as in Figure* 3 *except that $K_{deg} = 0$.*

the biological background for the choice $T^* = 10$ minutes were not discussed in [13] and [14]. In Figure 6, we study how our system reacts to the choice of T^* if $K_{deg} = 0$. It is found that the evolution of $[LR]$ at the dorsal midline becomes monotone, unlike the case in Figure 3, and as expected, the time to reach steady states strongly depends on the choice of T^*. In Figure 6, the steady states for $[LR]$ are shown for $T^* = 5, 8, 10, 12, 15, 20, 30, 45, 60$ minutes. The concentration of $[LR]$ varies almost linearly with respect to T^* until the receptors are close to being fully occupied when T^* is large.

Finally, we investigate the effect of diffusion. In Figure 7, Dpp-receptor complexes as functions of time and space are shown for five different choices of diffusion constants. Case (a) has all three diffusion constants the same as in Figure 3, cases (b)–(d) have one of the diffusion constants being 1% of the corresponding value in case (a), and the case (e) has two constants at 1% of the corresponding values in case (a). Similarly in Figure 8, some of the diffusion constants are 10-fold larger than others.

As shown in case (b) of both Figures 7 and 8, the magnitude of the diffusion coefficient for Dpp has very little effect on the broadness and intensity of Dpp activity at the dorsal midline. This is consistent with the behavior of the leading term perturbation solution. A larger diffusion for Dpp reduces the peak of transient Dpp-activity at the midline slightly and broadens it slightly. On the other hand, a decrease in diffusion constant for Dpp-Sog complexes, as in cases (d) and (e), significantly broad-

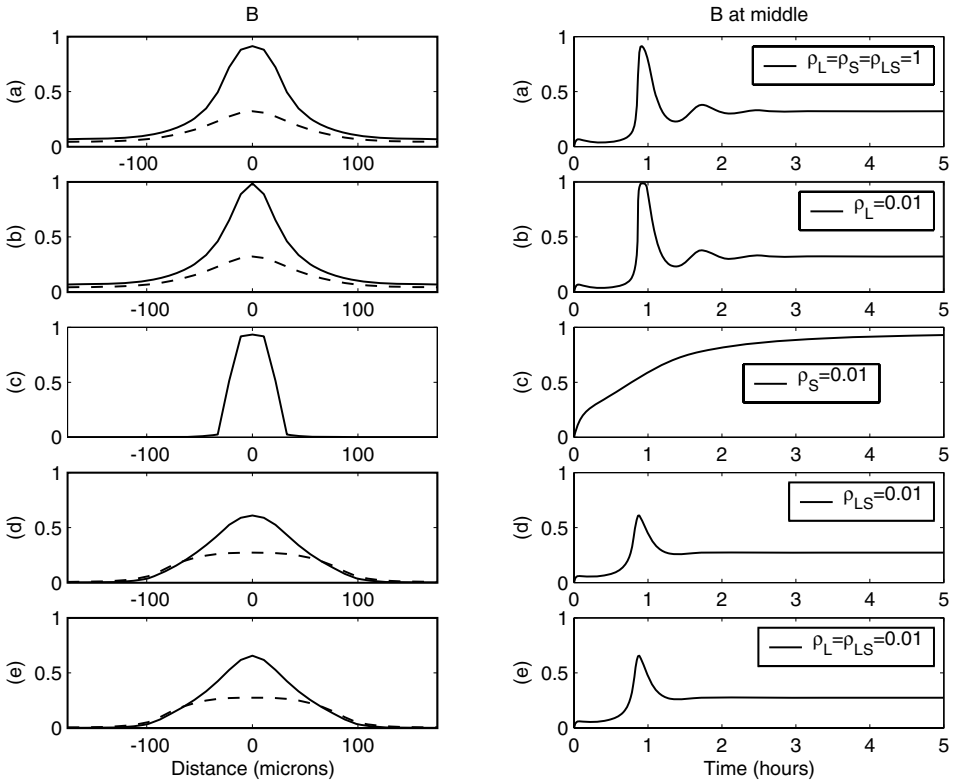


FIG. 7. *Effect of smaller diffusion constants on the transient peak and steady state. For the left-hand panels, solid line: transient peak; dotted line: steady state. Parameters are as in Figure* 3 *except for diffusion constants.*

ens the Dpp activity around the midline for both peak transient and steady state distributions, with the height of only the transient peak reduced significantly but with almost the same steady state at $x = 0$. The time to steady state and transient peaks seems to be insensitive to the change of the diffusion constants for Dpp or the Dpp-Sog complex.

As predicted by the perturbation solutions, varying the diffusion coefficient for Sog changes the Dpp activity around the dorsal midline significantly. As shown in Figure 7(c), a smaller diffusion of Sog relative to the diffusion of Dpp leads to more concentrated transient Dpp activity around the dorsal midline, but it takes much longer to reach the steady state, with a monotone increase of Dpp activity around the dorsal midline (i.e., there is no transient peak). On the other hand, larger diffusion of Sog relative to the diffusion of Dpp weakens and broadens the Dpp activity, as in Figure 8(c).

6. Conclusions. The dynamics of Dpp activities in the presence of the inhibitor Sog is analyzed herein to initiate a study of dorsal-ventral morphogen gradient formation in vertebrates and *Drosophila* embryos. Here we investigate a prototype morphogen system with typical ligand diffusion and degradation, but now with the additional feature of cleavage of Dpp-Sog complexes by Tolloid to free up Dpp.Among

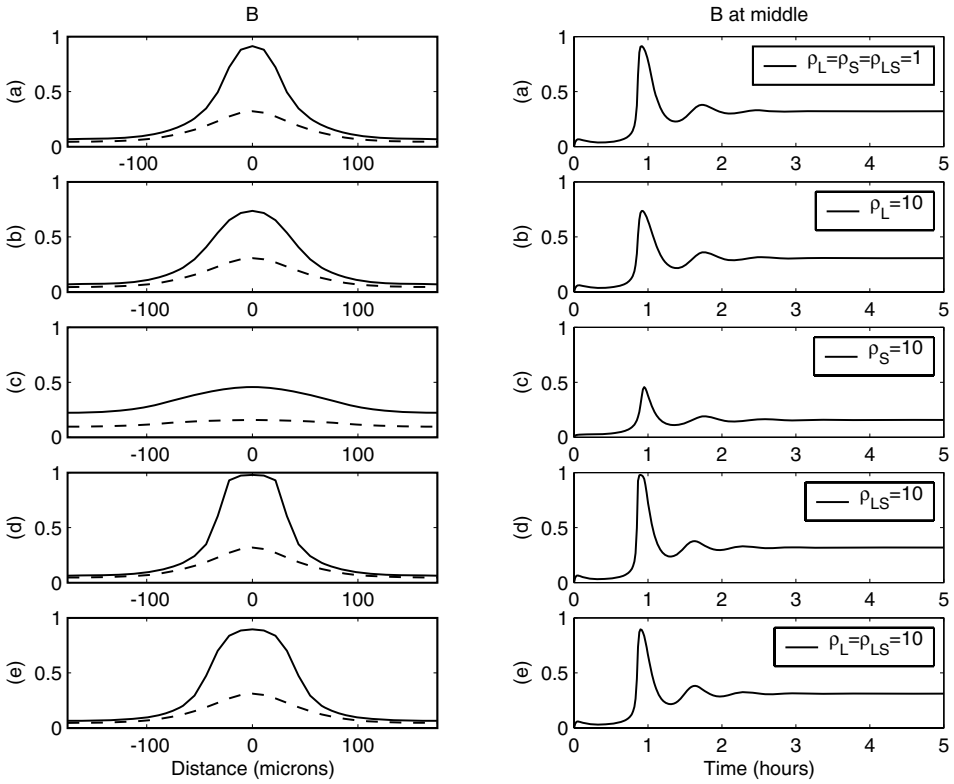


FIG. 8. *Effect of larger diffusion constants on the transient peak and steady state. For the left-hand panels, solid line: transient peak; dotted line: steady state. Parameters are as in Figure* 3 *except for diffusion constants.*

the principal results of our investigation is the establishment of a simple and biologically meaningful necessary and sufficient condition for the existence of a steady state gradient in the system. This condition requires a balance of the production rates of ligands, degradation rate of ligand-receptor complex, and rate of cleavage of ligand-inhibitor complex. For high Sog production rates (relative to the Dpp production rate), a perturbation solution has been obtained in terms of elementary functions. This solution exhibits an intense Dpp-receptor concentration near the dorsal midline. Numerical simulations of the *evolution* of the system confirmed these features of the steady state behavior. In addition, a transient peak of Dpp-receptor concentration at the dorsal midline was found to be even more intense prior to steady state, reaching more than twice the level of the steady state at its peak amount. This transient peak is more sensitively dependent on variation of the production of Sog than the steady state peak. The high Dpp-receptor concentration around the dorsal midline and other features of the system are consistent with experimental observations.

Appendix A.

Proof of Lemma 3.4. We show that there exists $M_1 > 0$ such that $\|\tilde{C}\|_{L^\infty} \leq M_1$. As in the proof of Lemma 3.3, $\|\tilde{C}\|_{L^1} \leq M_2$ for some constant $M_2 > 0$. Integrating

(27) in $(-1,1)$, we get

$$\int_{-1}^{1} h_{LS}\tilde{A}\tilde{D} = \int_{-1}^{1} (f_{LS}+g_{LS})\tilde{C}, \tag{56}$$

which implies that $\int_{-1}^{1}\tilde{A}\tilde{D} \le M_3\int_{-1}^{1}\tilde{C} \le M_2M_3$ for some $M_3>0$. Integrating (27) from -1 to x, we get

$$\rho_{LS}\tilde{C}_{,x} + \lambda\int_{-1}^{x}[h_{LS}\tilde{A}\tilde{D} - (f_{LS}+g_{LS})\tilde{C}] = 0, \quad -1<x<1. \tag{57}$$

Hence

$$\rho_{LS}\|\tilde{C}_{,x}\|_{L^\infty} \le \|h_{LS}\|_{L^\infty}\int_{-1}^{1}\tilde{A}\tilde{D} + \|f_{LS}+g_{LS}\|_{L^\infty}\int_{-1}^{1}\tilde{C} \le M_4 \tag{58}$$

for some constant $M_4>0$. This along with $\int_{-1}^{1}\tilde{C}\le M_2$ implies the L^∞ bound of $\tilde{C}$, which is independent of λ.

Next we show that there exists some constant $M_5>0$ such that $\|\tilde{A}\|_{L^\infty}\le M_5$. To this end, adding up (26) and (27) and integrating from -1 to x, we get

$$\rho_L\tilde{A}_{,x} + \rho_{LS}\tilde{C}_{,x} = \lambda\int_{-1}^{x}\left[\frac{g_Lh_L\tilde{A}}{f_L+g_L+h_L\tilde{A}} + (1-\tau)g_{LS}\tilde{C} - V_L\right], \tag{59}$$

which implies that

$$\|\tilde{A}_{,x}\|_{L^\infty} \le M_6\left(\|\tilde{C}_{,x}\|_{L^\infty} + \|g_L\|_{L^\infty} + \|\rho_{LS}\|_{L^\infty}\int_{-1}^{1}\tilde{C} + \|V_L\|_{L^\infty}\right) := M_7. \tag{60}$$

We claim that there exists some constant $M_8>0$ such that $\int_{-1}^{1}\tilde{A}\le M_8$. To establish this assertion, we argue by contradiction: if not, passing to a subsequence if necessary, we may assume that $\int_{-1}^{1}\tilde{A}\to+\infty$. This together with (60) implies that

$$\left|\frac{\tilde{A}(x)}{\int_{-1}^{1}\tilde{A}} - 1\right| \le \frac{\|\tilde{A}_{,x}\|_{L^\infty}}{\int_{-1}^{1}\tilde{A}} \le \frac{M_7}{\int_{-1}^{1}\tilde{A}} \to 0 \qquad \forall -1\le x\le 1. \tag{61}$$

Hence $\tilde{A}\to+\infty$ uniformly. Similar to (24) we have

$$\int_{-1}^{1} V_L = \int_{-1}^{1}\frac{g_Lh_L\tilde{A}}{f_L+g_L+h_L\tilde{A}} + (1-\tau)\int_{-1}^{1} V_S. \tag{62}$$

By (61) we have $\int_{-1}^{1}\frac{g_Lh_L\tilde{A}}{f_L+g_L+h_L\tilde{A}}\to\int_{-1}^{1}g_L$, which together with (62) implies that $\int_{-1}^{1}V_L = \int_{-1}^{1}g_L + (1-\tau)\int_{-1}^{1}V_S$. However, this contradicts (21). Therefore $\int_{-1}^{1}\tilde{A}$ is uniformly bounded for $\lambda\in(0,1]$. This together with (60) yields $\|\tilde{A}\|_{L^\infty}\le M_5$ for some $M_5>0$.

Finally we show that there exists some constant $M_9>0$ such that $\|\tilde{D}\|_{L^\infty}\le M_9$. We argue by contradiction: suppose not; passing to a subsequence if necessary, we may assume that $\|\tilde{D}\|_{L^\infty}\to\infty$ and $\lambda\to\hat{\lambda}\in[0,1]$. Set $\hat{D}(x) = \frac{\tilde{D}(x)}{\|\tilde{D}\|_{L^\infty}}$.Then $\hat{D}$

satisfies $\hat{D}_{,x}(-1) = \hat{D}_{,x}(1) = 0$, $\|\hat{D}\|_{L^\infty} = 1$, and

$$\rho_{LS}\hat{D}_{,xx} + \lambda\left[-h_{LS}\tilde{A}\hat{D} + \frac{f_{LS}\tilde{C} + V_S}{\|\hat{D}\|_{L^\infty}}\right] = 0, \quad -1 < x < 1. \tag{63}$$

Since $\|\tilde{A}\|_{L^\infty}$, $\|\tilde{A}_{,x}\|_{L^\infty}$ are uniformly bounded, we may assume that $\tilde{A}(x) \to A^*(x)$ uniformly in $[-1,1]$. From (62) and (20) we see $\int_{-1}^{1}\tilde{A} \geq M_{10} > 0$ for some constant M_{10}. Hence $A^* \not\equiv 0$ since $\int_{-1}^{1} A^* \geq M_{10} > 0$. By standard regularity theory we may assume that $\hat{D}(x) \to D^*(x)$ in $C^1[-1,1]$, and D^* is a weak solution of

$$\rho_{LS}D^*_{,xx} - \hat{\lambda}h_{LS}A^*D^* = 0, \quad -1 < x < 1, \quad D^*_{,x}(-1) = D^*_{,x}(1) = 0. \tag{64}$$

Moreover, $D^* \geq 0$ in $[-1,1]$ and $\|D^*\|_{L^\infty} = 1$. If $\hat{\lambda} > 0$, since $A^* \not\equiv 0$, $A^* \geq 0$, by the maximum principle we see that $D^* \equiv 0$, which contradicts $\|D^*\|_{L^\infty} = 1$; if $\hat{\lambda} = 0$, then it follows from (64) that $D^* \equiv 1$, i.e., $\hat{D}(x) \to 1$ uniformly. Dividing (56) by $\|\tilde{D}\|_{L^\infty}$, we have $\int_{-1}^{1} h_{LS}\tilde{A}\hat{D} = \int_{-1}^{1}(f_{LS} + g_{LS})\tilde{C}/\|\tilde{D}\|_{L^\infty}$. Then we obtain $\int_{-1}^{1} h_{LS}A^* = 0$, which implies that $A^* \equiv 0$. Contradiction! This completes the proof of (34). ∎

When $\lambda = 0$, $(\tilde{A}, \tilde{C}, \tilde{D})$ is a solution of (26)–(29) if and only if $\tilde{A}, \tilde{C}$, and $\tilde{D}$ are all constants. It turns out that a particular triple, denoted by $(\hat{A}, \hat{C}, \hat{D})$, is special, where $\hat{A}, \hat{C}, \hat{D}$ are defined as follows: by (20)–(21) it is easy to see that there is a unique positive constant, denoted by $\hat{A}$, such that

$$\int_{-1}^{1} \frac{g_L h_L \hat{A}}{f_L + g_L + h_L\hat{A}} = \int_{-1}^{1} V_L - (1-\tau)\int_{-1}^{1} V_S. \tag{65}$$

Set

$$\hat{D} = \frac{\int_{-1}^{1}(f_{LS} + g_{LS})\int_{-1}^{1} V_S}{\hat{A}\int_{-1}^{1} h_{LS}\int_{-1}^{1} g_{LS}}, \qquad \hat{C} = \frac{\int_{-1}^{1} V_S}{\int_{-1}^{1} g_{LS}}. \tag{66}$$

LEMMA A.1. *Suppose that* (20)–(21) *holds. Let* $(A_\lambda, C_\lambda, D_\lambda)$ *denote positive solutions of* (26)–(29). *Then as* $\lambda \to 0+$, $(A_\lambda, C_\lambda, D_\lambda) \to (\hat{A}, \hat{C}, \hat{D})$ *uniformly.*

Proof. By Lemma 3.4, $(A_\lambda, C_\lambda, D_\lambda)$ are uniformly bounded. By standard regularity theory and the embedding theorem, passing to a subsequence if necessary, we may assume that $(A_\lambda, C_\lambda, D_\lambda) \to (\overline{A}, \overline{C}, \overline{D})$ uniformly, where $\overline{A}$, $\overline{C}$, and $\overline{D}$ satisfy $\overline{A}_{xx} = \overline{C}_{xx} = \overline{D}_{xx} = 0$, and $\overline{A}_x = \overline{C}_x = \overline{D}_x = 0$ at $x = -1, 1$. Therefore $\overline{A}$, $\overline{C}$, $\overline{D}$ are all nonnegative constants. Passing to the limit in (62) (with $\tilde{A}$ being replaced by A_λ), we have $\overline{A} = \hat{A}$. Similarly we can show that $\hat{C} = \overline{C}$ and $\hat{D} = \overline{D}$. Since the limit $(\hat{A}, \hat{C}, \hat{D})$ is unique, the convergence $(A_\lambda, C_\lambda, D_\lambda) \to (\hat{A}, \hat{C}, \hat{D})$ is true for the whole sequence, and the limit is uniform in x. ∎

LEMMA A.2. *There exists some constant* $\delta_1 > 0$ *such that if* $0 < \lambda \leq \delta_1$, (26)–(29) *has a unique positive solution.*

Proof. Set $X = \{u \in C[-1,1] : \int_{-1}^{1} u(x)\,dx = 0\}$, $Z = \{u \in X : u_{,x}(-1) = u_{,x}(1) = 0\}$, and define the projection operator $P : C[-1,1] \to X$ by $Pu = u - \int_{-1}^{1} u(x)\,dx$. For $(\lambda, A_0, a_0, C_0, c_0, D_0, d_0) \in R^1 \times (Z \times R^1)^3$, define $F : R^1 \times (Z \times$

$R^1)^3 \to (X \times R^1)^3$ by

(67)
$$F(\lambda, A_0, a_0, C_0, c_0, D_0, d_0) = \begin{pmatrix} \rho_L A_{0,xx} + \lambda P F_1^+(x, A_0 + a_0, C_0 + c_0, D_0 + d_0) \\ \int_{-1}^1 F_1^+(x, A_0 + a_0, C_0 + c_0, D_0 + d_0)\, dx \\ \rho_{LS} C_{0,xx} + \lambda P F_2(x, A_0 + a_0, C_0 + c_0, D_0 + d_0) \\ \int_{-1}^1 F_2(x, A_0 + a_0, C_0 + c_0, D_0 + d_0)\, dx \\ \rho_S D_{0,xx} + \lambda P F_3(x, A_0 + a_0, C_0 + c_0, D_0 + d_0) \\ \int_{-1}^1 F_3(x, A_0 + a_0, C_0 + c_0, D_0 + d_0)\, dx \end{pmatrix}.$$

By the definition of $\hat{A}$, $\hat{C}$, $\hat{D}$, $F(0, \hat{A}, \hat{C}, \hat{D}, 0, 0, 0) = (0, 0, 0, 0, 0, 0)$. The Fréchet derivative of F with respect to $(A_0, a_0, C_0, c_0, D_0, d_0)$ at $(\lambda, A_0, a_0, C_0, c_0, D_0, d_0) = (0, \hat{A}, 0, \hat{C}, 0, \hat{D}, 0)$ is given by

(68)
$$D_{(A_0,a_0,C_0,c_0,D_0,d_0)} F\big|_{(0,\hat{A},0,\hat{C},0,\hat{D},0)} = \begin{pmatrix} \rho_L \frac{d^2}{dx^2} & 0 & 0 & 0 & 0 & 0 \\ 0 & \rho_{LS} \frac{d^2}{dx^2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \rho_S \frac{d^2}{dx^2} & 0 & 0 & 0 \\ * & * & * & & & \\ * & * & * & M_{3\times 3} & & \\ * & * & * & & & \end{pmatrix},$$

where $M_{3\times 3}$ is the 3×3 matrix

(69)
$$\begin{pmatrix} \int_{-1}^1 \left[\frac{-g_L h_L (f_L + g_L)}{(f_L + g_L + h_L \hat{A})^2} - h_{LS} \hat{D} \right] & \int_{-1}^1 (f_{LS} + \tau g_{LS}) & -(\int_{-1}^1 h_{LS}) \hat{A} \\ (\int_{-1}^1 h_{LS}) \hat{D} & -\int_{-1}^1 (f_{LS} + g_{LS}) & (\int_{-1}^1 h_{LS}) \hat{A} \\ -(\int_{-1}^1 h_{LS}) \hat{D} & \int_{-1}^1 f_{LS} & -(\int_{-1}^1 h_{LS}) \hat{A} \end{pmatrix}.$$

Since the operator $\frac{d^2}{dx^2}$, subject to the no flux boundary condition, is an isomorphism from Z to X, we see that the operator $D_{(A_0,a_0,C_0,c_0,D_0,d_0)} F\big|_{(0,\hat{A},0,\hat{C},0,\hat{D},0)}$ is invertible from $(Z \times R^1)^3$ to $(X \times R^1)^3$ if and only if the matrix $M_{3\times 3}$ is invertible. It is straightforward to check that the determinant of $M_{3\times 3}$ is equal to

(70)
$$\left(\int_{-1}^1 h_{LS} \right) \hat{D} \cdot \int_{-1}^1 (f_{LS} + g_{LS}) \cdot \left(\int_{-1}^1 h_{LS} \right) \hat{A}\, (1 - \gamma_2)(-\gamma_1),$$

where γ_1, γ_2 are defined as

(71)
$$\gamma_1 = \frac{\int_{-1}^1 \frac{g_L h_L (f_L + g_L)}{(f_L + g_L + h_L \hat{A})^2}}{\left(\int_{-1}^1 h_{LS} \right) \hat{D}}, \qquad \gamma_2 = \frac{\int_{-1}^1 f_{LS}}{\int_{-1}^1 (f_{LS} + g_{LS})}.$$

Since $\gamma_1 > 0$ and $0 < \gamma_2 < 1$, $M_{3\times 3}$ is nondegenerate.

By the implicit function theorem, there exists $\delta_2 > 0$ such that if $0 < \lambda \le \delta_2$, there is a unique solution to $F = 0$, denoted by $(A_\lambda(x), a_\lambda(x), C_\lambda(x), c_\lambda(x), D_\lambda(x), d_\lambda(x))$, in some neighborhood of $(\hat{A}, 0, \hat{C}, 0, \hat{D}, 0)$. As $\lambda \to 0+$, $(A_\lambda, a_\lambda, C_\lambda, c_\lambda, D_\lambda, d_\lambda) \to (\hat{A}, 0, \hat{C}, 0, \hat{D}, 0)$ uniformly. In particular, for $0 < \lambda \le \delta_2$, $(A_\lambda + a_\lambda, C_\lambda + c_\lambda, D_\lambda + d_\lambda)$ is the unique positive solution of (26)–(29) in some neighborhood of $(\hat{A}, \hat{C}, \hat{D})$. This and Lemma A.1 imply that, for $0 < \lambda \ll 1$, (26)–(29) has a unique positive solution. ∎

LEMMA A.3. *Let (A^*, C^*, D^*) denote the unique positive solution of* (26)–(29) *for $0 < \lambda \ll 1$. Then for $0 < \lambda \ll 1$, the Fréchet derivative of $T(\lambda)$ with respect to $(\tilde{A}, \tilde{C}, \tilde{D})$ at (A^*, C^*, D^*), denoted by $D_{(\tilde{A},\tilde{C},\tilde{D})}T(\lambda)\big|_{(A^*,C^*,D^*)}$, has no eigenvalue greater than or equal to* 1.

Proof. By (35), $D_{(\tilde{A},\tilde{C},\tilde{D})}T(\lambda)\big|_{(A^*,C^*,D^*)}(\varphi_1, \varphi_2, \varphi_3)$ is given by

$$\begin{pmatrix} L_{\rho_L}^{-1}\Big\{\Big[1 + \lambda\frac{\partial F_1}{\partial \tilde{A}}(x, A^*, C^*, D^*)\Big]\varphi_1 + \lambda\frac{\partial F_1}{\partial \tilde{C}}\varphi_2 + \lambda\frac{\partial F_1}{\partial \tilde{D}}\varphi_3\Big\} \\ L_{\rho_{LS}}^{-1}\Big\{\lambda\frac{\partial F_2}{\partial \tilde{A}}\varphi_1 + \Big[1 + \lambda\frac{\partial F_2}{\partial \tilde{C}}\Big]\varphi_2 + \lambda\frac{\partial F_2}{\partial \tilde{D}}\varphi_3\Big\} \\ L_{\rho_S}^{-1}\Big\{\lambda\frac{\partial F_3}{\partial \tilde{A}}\varphi_1 + \lambda\frac{\partial F_3}{\partial \tilde{C}}\varphi_2 + \Big[\lambda\frac{\partial F_3}{\partial \tilde{D}} + 1\Big]\varphi_3\Big\} \end{pmatrix},$$

where $\frac{\partial F_i}{\partial \tilde{A}}$, $\frac{\partial F_i}{\partial \tilde{C}}$, $\frac{\partial F_i}{\partial \tilde{D}}$ $(i = 1, 2, 3)$ are evaluated at (x, A^*, C^*, D^*).

We argue by contradiction: suppose that Lemma A.3 fails. Passing to a subsequence if necessary, we may assume that for $0 < \lambda \ll 1$ the operator $D_{(\tilde{A},\tilde{C},\tilde{D})}T(\lambda)|_{(A^*,C^*,D^*)}$ has eigenvalue $\mu = \mu(\lambda) \ge 1$, with the corresponding eigenfunction $(\varphi_1, \varphi_2, \varphi_3)$ normalized by $\|\varphi_1\|_{L^\infty} + \|\varphi_2\|_{L^\infty} + \|\varphi_3\|_{L^\infty} = 1$. Then $(\varphi_1, \varphi_2, \varphi_3)$ satisfies

$$-\mu\rho_L\frac{d^2\varphi_1}{dx^2} + (\mu - 1)\varphi_1 = \lambda\left[\frac{\partial F_1}{\partial \tilde{A}}\varphi_1 + \frac{\partial F_1}{\partial \tilde{C}}\varphi_2 + \frac{\partial F_1}{\partial \tilde{D}}\varphi_3\right], \tag{72}$$

$$-\mu\rho_{LS}\frac{d^2\varphi_2}{dx^2} + (\mu - 1)\varphi_2 = \lambda\left[\frac{\partial F_2}{\partial \tilde{A}}\varphi_1 + \frac{\partial F_2}{\partial \tilde{C}}\varphi_2 + \frac{\partial F_2}{\partial \tilde{D}}\varphi_3\right], \tag{73}$$

$$-\mu\rho_S\frac{d^2\varphi_3}{dx^2} + (\mu - 1)\varphi_3 = \lambda\left[\frac{\partial F_3}{\partial \tilde{A}}\varphi_1 + \frac{\partial F_3}{\partial \tilde{C}}\varphi_2 + \frac{\partial F_3}{\partial \tilde{D}}\varphi_3\right], \tag{74}$$

$$(\varphi_1)_{,x} = (\varphi_2)_{,x} = (\varphi_3)_{,x} = 0 \quad \text{at } x = -1, 1, \tag{75}$$

where $\frac{\partial F_i}{\partial \tilde{A}}$, $\frac{\partial F_i}{\partial \tilde{C}}$, $\frac{\partial F_i}{\partial \tilde{D}}$ $(i = 1, 2, 3)$ in (72)–(74) are evaluated at $(\tilde{A}, \tilde{C}, \tilde{D}) = (A^*, C^*, D^*)$.

It is easy to see that $\mu(\lambda) \to 1$ as $\lambda \to 0+$, and the corresponding eigenfunctions $(\varphi_1, \varphi_2, \varphi_3) \to (\overline{\varphi}_1, \overline{\varphi}_2, \overline{\varphi}_3)$ uniformly, where $(\overline{\varphi}_1, \overline{\varphi}_2, \overline{\varphi}_3)$ are constants satisfying $|\overline{\varphi}_1| + |\overline{\varphi}_2| + |\overline{\varphi}_3| = 1$. Set $\mu(\lambda) = 1 + \lambda\mu_1(\lambda)$. Since $\mu(\lambda) \ge 1$, we have $\mu_1(\lambda) \ge 0$. Integrating (72)–(74), we get

$$\int_{-1}^{1}\left[\frac{\partial F_1}{\partial \tilde{A}} - \mu_1\right]\varphi_1 + \int_{-1}^{1}\frac{\partial F_1}{\partial \tilde{C}}\varphi_2 + \int_{-1}^{1}\frac{\partial F_1}{\partial \tilde{D}}\varphi_3 = 0, \tag{76}$$

$$\int_{-1}^{1} \frac{\partial F_2}{\partial \tilde{A}}\varphi_1 + \int_{-1}^{1} \left[\frac{\partial F_2}{\partial \tilde{C}} - \mu_1\right]\varphi_2 + \int_{-1}^{1} \frac{\partial F_2}{\partial \tilde{D}}\varphi_3 = 0, \tag{77}$$

$$\int_{-1}^{1} \frac{\partial F_3}{\partial \tilde{A}}\varphi_1 + \int_{-1}^{1} \frac{\partial F_3}{\partial \tilde{C}}\varphi_2 + \int_{-1}^{1} \left[\frac{\partial F_3}{\partial \tilde{D}} - \mu_1\right]\varphi_3 = 0. \tag{78}$$

We first prove that $\mu_1(\lambda)$ is uniformly bounded for all $0 < \lambda \ll 1$. If not, passing to a subsequence if necessary, we may assume that as $\lambda \to 0+$, $\mu_1(\lambda) \to +\infty$. Divide (76) by μ_1; passing to the limit, we find that $\overline{\varphi}_1 = 0$. Similarly, $\overline{\varphi}_2 = \overline{\varphi}_3 = 0$. However, this contradicts $|\overline{\varphi}_1| + |\varphi_2| + |\overline{\varphi}_3| = 1$. Therefore $\mu_1(\lambda)$ is nonnegative and uniformly bounded. Passing to a subsequence if necessary, we may assume that $\mu_1(\lambda) \to \overline{\mu}_1 \geq 0$ as $\lambda \to 0+$.

Passing to the limit in (76)–(78), by Lemma A.1, $(M_{3\times 3} - \overline{\mu}_1 I_{3\times 3})(\overline{\varphi}_1, \overline{\varphi}_2, \overline{\varphi}_3) = (0,0,0)$. Since $(\overline{\varphi}_1, \overline{\varphi}_2, \overline{\varphi}_3) \neq (0,0,0)$, $|M_{3\times 3} - \overline{\mu}_1 I_{3\times 3}| = 0$. However, direct calculation yields that $|M_{3\times 3} - \overline{\mu}_1 I_{3\times 3}|$ is equal to

$$\begin{aligned}
&- \left(\int_{-1}^{1} h_{LS}\right)\hat{D} \cdot \int_{0}^{1} (f_{LS} + g_{LS}) \cdot \left(\int_{-1}^{1} h_{LS}\right)\hat{A} \\
&\cdot \left\{ \left(\gamma_1 + \frac{\overline{\mu}_1}{(\int_{-1}^{1} h_{LS})\hat{D}}\right) \cdot \left(1 + \frac{\overline{\mu}_1}{\int_{-1}^{1}(f_{LS} + g_{LS})}\right) \cdot \frac{\overline{\mu}_1}{(\int_{-1}^{1} h_{LS})\hat{A}} \right. \\
&\quad + \left[(1-\tau)(1-\gamma_2) + \frac{\overline{\mu}_1}{\int_{-1}^{1}(f_{LS} + g_{LS})}\right] \cdot \frac{\overline{\mu}_1}{\int_{-1}^{1} h_{LS}\hat{A}} \\
&\quad \left. + \left(\gamma_1 + \frac{\overline{\mu}_1}{(\int_{-1}^{1} h_{LS})\hat{D}}\right) \cdot \left(1 - \gamma_2 + \frac{\overline{\mu}_1}{\int_{-1}^{1}(f_{LS} + g_{LS})}\right) \right\},
\end{aligned}$$

which is negative since $\overline{\mu}_1 \geq 0$, $\gamma_1 > 0$, $0 \leq \tau \leq 1$, and $\gamma_2 < 1$. Contradiction! This completes the proof of Lemma A.3. ∎

Proof of Proposition 3.5. By Lemma A.2, for $0 < \lambda \ll 1$, $T(\lambda)$ has a unique fixed point. By Lemma A.3, 1 is not an eigenvalue of $D_{(\tilde{A},\tilde{C},\tilde{D})}T(\lambda)|_{(A^*,C^*,D^*)}$. Hence $\deg\big(I - T(\lambda),\ \Omega,\ (0,0,0)\big) = (-1)^{\beta}$, where β is the number of eigenvalues (counting algebraic multiplicity) of $D_{(\tilde{A},\tilde{C},\tilde{D})}T(\lambda)|_{(A^*,C^*,D^*)}$, which is greater than 1. By Lemma A.3 we see that $\beta = 0$. Hence $\deg(I - T(\lambda), \Omega, (0,0,0)) = 1$ for $0 < \lambda \ll 1$. ∎

Acknowledgments. The authors acknowledge the very helpful discussions with A. Lander and L. Marsh. Part of this work was done when Y. L. was visiting the Department of Mathematics of UCI, and he would like to express appreciation for the hospitality he received.

REFERENCES

[1] L. Wolpert, R. Beddington, J. Brockes, T. Jessel, P. Lawrence, and E. Meyerowitz, *Principles of Development*, 2nd ed., Oxford University Press, Oxford, UK, 2002.

[2] A. A. Teleman, M. Strigini, and S. M. Cohen, *Shaping morphogen gradients*, Cell, 105 (2001), pp. 559–562.

[3] J. B. Gurdon and P. Y. Bourillot, *Morphogen gradient interpretation*, Nature, 413 (2001), pp. 797–803.

[4] A. Lander, Q. Nie, and F. Y. M. Wan, *Do morphogen gradients arise by diffusion?*, Dev. Cell, 2 (2002), pp. 785–796.

[5] E. Bier, *A unity of opposites*, Nature, 398 (1999), pp. 375–376.

[6] H. L. Ashe and M. Levine, *Local inhibition and long-range enhancement of Dpp signal transduction by Sog*, Nature, 398 (1999), pp. 427–431.
[7] M. Oelgeschlager, J. Larrain, D. Geissert, and E. M. Robertis, *The evolutionarily conserved BMP-binding protein twisted gastrulation promotes BMP signalling*, Nature, 405 (2000), pp. 757–762.
[8] J. Ross, O. Shimmi, P. Vilmos, A. Petryk, H. Kim, K. Gaudenez, S. Hermanson, A. Ekker, M. O'Connor, and J. L. Marsh, *Twisted gastrulation is a conserved extracellular BMP antagonist*, Nature, 410 (2001), pp. 479–483.
[9] Y. Lou, Q. Nie, and F.Y.M. Wan, *Nonlinear eigenvalue problems in the stability analysis of morphogen gradients*, Stud. Appl. Math., 113 (2004), pp. 183–215.
[10] A. Lander, Q. Nie, B. Vargas, and F. Y. M. Wan, *Aggregation of a distributed source in morphogen gradient formation*, Stud. Appl. Math., (2005), to appear.
[11] A. Lander, Q. Nie, and F. Y. M. Wan, *Spatially distributed morphogen production and morphogen gradient formation*, Math. Biosci. Eng., (2005), to appear.
[12] A. Lander, Q. Nie, and F. Y. M. Wan, *Internalization and end flux in morphogen gradient formation*, J. Comput. Appl. Math., (2005), to appear.
[13] A. Eldar, R. Dorfman, D. Weiss, H. Ashe, B. Shilo, and N. Barkai, *Robustness of the BMP morphogen gradient in* Drosophila *embryonic patterning*, Nature, 419 (2002), pp. 304–308.
[14] A. Eldar, R. Dorfman, D. Weiss, H. Ashe, B. Shilo, and N. Barkai, *Supplement—Robustness of the BMP morphogen gradient in* Drosophila *embryonic patterning*, Nature, 419 (2002), pp. 304–308.
[15] A. Eldar, D. Rosin, B. Z. Shilo, and N. Barkai, *Self-enhanced ligand degradation underlies robustness of morphogen gradients*, Dev. Cell, 5 (2003), pp. 635–646.
[16] Z. Wang, O. Marcu, M. W. Berns, and J. L. Marsh, *In vivo FCS measurements of ligand diffusion in intact tissues*, Proc. SPIE, 5323 (2004), pp. 177–183.
[17] J. Kao, Q. Nie, A. Teng, F. Y. M. Wan, A. Lander, and I. L. Marsh, *Can morphogen activity be enhanced by its inhibitors?*, in Proceeding of the 2nd MIT Conference on Computational Mechanics, 2003, Elsevier Press, New York, Vol. 2, pp. 1729–1734.
[18] C. Mizutani, Q. Nie, F. Y. M. Wan, Y. Zhang, P. Vilmos, E. Bier, L. Marsh, and A. Lander, *Origin of the BMP activity gradient in the* Drosophila *embryo*, Devel. Cell (2005), to appear.
[19] F. C. Crick, *Diffusion in embryogenesis*, Nature, 225 (1970), pp. 40–42.
[20] M. Kerszberg and L. Wolpert, *Mechanisms for positional signalling by morphogen transport: A theoretical study*, J. Theoret. Biol., 191 (1998), pp. 103–114.
[21] S. Piccolo, E. Agiusa, B. Lu, S. Goodman, L. Dale, and E. De Robertis, *Cleavage of chordin by Xolloid metalloprotease suggests a role for proteolytic processing in the regulation of Spemann organizer activity*, Cell, 91 (1997), pp. 407–416.
[22] G. Marques, M. Musacchio, M. J. Shimell, K. W. Stapleton, K. Cho, and M. O'Connor, *Production of a Dpp activity gradient in the early Drosophila embryo through the opposing actions of the Sog and TLD proteins*, Cell, 91 (1997), pp. 417–426.
[23] M. H. Protter and H. F. Weinberger, *Maximum Principle in Differential Equations*, 2nd ed., Springer-Verlag, Berlin, 1984.
[24] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, 2nd ed., Springer-Verlag, New York, 1994.
[25] J. C. Strikwerda, *Finite Difference Schemes and Partial Differential Equations*, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1989.

SIAM J. APPL. MATH.
Vol. 65, No. 5, pp. 1772–1793

ASYMPTOTIC BEHAVIOR OF INTERNAL ROSSBY WAVES WITH A SHARP DENSITY INTERFACE*

A. OUAHSINE† AND P. A. BOIS‡

Abstract. An asymptotic analysis of a quasi-geostrophic model is presented to investigate the vertically propagating internal Rossby wave structure in the presence of a sharp density interface. For the case of a nonconstant density gradient, the pattern of the vertical structure is nontrivial and depends on the parameter of varying stratification, denoted by δ. As this parameter approaches a critical value from above (i.e., $\delta \approx \delta_{cri}$), the internal wave amplitudes increase continually until wave breaking occurs. When $\delta < \delta_{cri}$, two cases are considered. For the case without a turning-point, the requirement of appropriate interfacial conditions provides a general matching solution, available only for a certain range of δ. In the presence of a turning-point, which describes the solution transition from monotonic to oscillatory behavior or vice versa, three asymptotic forms of the solutions are derived, even for small values of δ.

Key words. asymptotic matching, internal waves, ocean waves, turning-points

AMS subject classifications. 15A15, 15A09, 15A23

DOI. 10.1137/S0036139902414732

1. Introduction.

Internal waves are long waves, with large amplitudes within several kilometers, which are generally produced in the upper layer of the sea surface by tidal fluctuations, differences of density, and atmospheric conditions. Internal waves increase the sea surface current and create a stormy current, which affects fishing grounds, en route vessels, rigs, etc. They may generate slowly varying local currents in open ocean or in fjords. Such induced currents are important factors to be considered when determining the dimensions of offshore platforms operating in deep water areas where internal waves might occur.

When there is a vertical density gradient with a thin density interface, the internal waves can be limited in their spectral density by sporadic local instabilities, and when the waves break, they produce turbulence (Fernando and Hunt [1] and Hannan, Fernando, and List [4]). This process produces a turbulent mixing between fluids of different densities, leading to the overturning of waves by static instability. In environmental fluid mechanics, an understanding of these mixing processes is essential in determining water quality and also the physical properties and concentrations of chemical and biological constituents in the well-mixed layers.

Thus, there is a general need for mathematical models of large-amplitude internal waves propagating along an arbitrary pycnocline or thermocline. In this paper we present an analytical study, to deal with this kind of wave when the density is a continuous function of depth in the presence of a very sharp gradient, taking viscous dissipation into account.

Several attempts have been made to investigate internal waves with respect to the interfacial thickness. Smith and Vallis [2] have presented work on using freely

*Received by the editors September 17, 2002; accepted for publication (in revised form) December 8, 2004; published electronically July 26, 2005.
http://www.siam.org/journals/siap/65-5/41473.html
†Université de Technologie de Compiègne, Laboratoire Roberval, UMR-CNRS 6066, BP 20529, F-60205 Compiègne Cedex, France (ouahsine@utc.fr).
‡UFR Math Pures et Appliquées, LML-UMR CNRS-8107, B.D. Paul Langevin, F-59655 Villeneuve d'Ascq, France (bois@univ-lille1.fr).

decaying geostrophic turbulence to understand and explain the vertical and horizontal flow of energy through a stratified horizontally homogeneous geostrophic fluid. They found that the stratification profile, in particular the presence of a pycnocline, has significant qualitative effects on the efficiency and spectral pathways of energy flow. With uniform stratification, energy in high baroclinic modes transfers directly and almost completely to barotropic mode. In contrast, in the presence of ocean-like stratification, kinetic energy in high baroclinic modes transfers intermediately to the first baroclinic mode, whence it transfers inefficiently to the barotropic mode. The efficiency of transfer to the barotropic mode is reduced as the pycnocline is made increasingly thin.

In his theoretical study, Phillips [5] investigated the degradation of the first internal wave mode and considered the possible occurrence of a dynamic instability in the thermocline where the rate of shear U induced by internal waves reaches its maximum. He postulated the possibility of the occurrence of similar interfacial instabilities when the local gradient Richardson number $R_i = \Delta b L/U^2 < \frac{1}{4}$ for the case of the first internal wave mode at a sharp density interface, where Δb is the characteristic buoyancy jump and L is the characteristic length scale. He added that if the energy supply to the internal wave mode continues after the above limiting condition is reached, a local instability may develop. The possibility of such an instability occurring will be investigated in our study.

Thus, in this paper, in place of R_i we reorganize a critical value of the parameter of varying stratification, namely δ. With further decreases in δ, we deal with a sharp density interface, and hence with the generation of waves that grow until they break and overturn.

The possibility of exponential growth of internal wave amplitudes in horizontally inhomogeneous layers has been suggested by Navrotsky and Simonenko [16] and Navrotsky [17]. They have analyzed the propagation of weakly nonlinear internal waves with slowly varying amplitudes (compared to the wave period) in a layer with a horizontal gradient of buoyancy frequency $dN(z,x)/dx$. They showed that such waves could produce effective mixing within the thermocline. The horizontal variation of the vertical density gradient favorable for the amplification of high-frequency propagating waves can be developed through the deformation of the density field by long topographic or tidal internal waves.

In this analysis, we discuss a linear analysis for deriving the growth rate of the wave amplitude. The instability described here can be interpreted as a criterion in which the wave amplitude grows, according to the parameter δ (see Figure 3.2), to a level such that δ drops below a critical value; then the disturbances can grow exponentially, whence the primary wave amplitude exceeds a certain limiting value to such an extent that an effective discontinuity or front may develop (Basovich and Tsimring [3] and Badulin, Shrira, and Tsimring [6]).

The structure of this paper is as follows. In section 2 equations of motion with continuous density with variable gradient stratification are introduced. The mathematical problem with the asymptotic analysis is formulated, and the method of solving it is discussed. Section 3 considers the development of the solution as the characteristic scale of varying stratification decreases. Section 4 is devoted to the study of the three-layer discontinuous gradient model and to giving stable solutions in the case of the existence of turning-points. Section 5 extends this theory to seek stable solutions in the case of no turning-points. Observational evidence of large-amplitude internal waves and discussions are presented in section 6. Section 7 concludes the paper.

2. Equations of motion and formulation of the problem. The set of dimensionless equations of incompressible stratified fluid dynamics can be written in the Boussinesq approximation in a rotating flow (Ouahsine [8] and Bois [10]) as

$$\begin{aligned} \varrho\left(Ro\left(St\frac{\partial \mathbf{v}}{\partial t}+\mathbf{v}.\mathbf{D}\mathbf{v}+w\frac{\partial \mathbf{v}}{\partial z}\right)+(1+Ro\beta y)k\wedge\mathbf{v}\right)+\frac{Ro}{Fr^2}\mathbf{D}p &= E_v\frac{\partial^2\mathbf{v}}{\partial z^2}+E_h\mathbf{D}^2\mathbf{v},\\ \frac{\partial p}{\partial z}+Bo\varrho &= 0,\\ \mathbf{D}.\mathbf{v}+\frac{\partial w}{\partial z} &= 0,\\ St\frac{\partial \varrho}{\partial t}+\mathbf{v}.\mathbf{D}\varrho+\mathbf{w}\frac{\partial \rho}{\partial z} &= 0. \end{aligned} \tag{2.1}$$

These equations are valid for long wave motions, say $\gamma = H/L \ll 1$, where L and H are the horizontal and vertical characteristic scales, respectively, under the Boussinesq approximation (Pedlosky [12]) and the hydrostatic balance assumptions. Vectors $(\mathbf{i},\mathbf{j},\mathbf{k})$ are unit vectors eastwards, northwards, and locally upwards; $\mathbf{D} = (\partial/\partial x\,,\ \partial/\partial y)$ is the horizontal gradient; t is the time; p is the pressure; and $\beta = (L^2/U)\,(df_0/dy)_{y=0}$ is the beta-plane parameter, where $f_0 = 2\Omega\sin\lambda_0$ is the local Coriolis parameter, λ_0 is the latitude, $\Omega = 7.3\ 10^{-5}$ rad/s is the Earth's rotation, and a_0 is the Earth's radius. Additionally, $\mathbf{v}$ is the horizontal velocity vector (with eastward and northward components u and v), and w is the vertical velocity.

Let $T_0, U_0, \varrho_{00}, \Pi$ be the characteristic scales for time t, for the horizontal velocity (u,v), for the density ϱ, and for the pressure p, respectively. The nondimensional parameters appearing in these equations are the Rossby number $Ro = U_0/f_0L$, the Froude number $Fr = U_0/\sqrt{gH}$, the Strouhal number $St = L/U_0T_0$, the number $Bo = \varrho_{00}gH/\Pi$, and the Ekman numbers for vertical and horizontal kinematics viscosities, respectively $E_v = \nu/f_0H^2$ and $E_h = \nu/f_0L^2$.

The boundary conditions at the bottom (on $z=-1$) are $\mathbf{v}=\mathbf{0}$ and $w=0$, and the initial conditions (at $t=0$) are $\mathbf{v}=\mathbf{v}^0$ and $p=p^0$, where $\mathbf{v}^0$ and p^0 are known.

We assume that, at the sea surface (on $z=0$), the steady state motion is driven by the surface wind stress $[\tau]_w$. It follows that the stress continuity at the ocean-atmosphere interface may be given by (Gill [11])

$$[\tau]_i = [-p\mathbf{n}I_d + 2\mu\mathbf{n}D]_i \quad \text{and} \quad p = p_a, \tag{2.2}$$

where $\mathbf{n}$ is the unit normal vector to the interface, D is the symmetric part of the deformation rate tensor, μ is the eddy viscosity, I_d is the unit vector tensor, and p_a is the atmospheric pressure. The bracket $[u]_i$ denotes the jump of the function u across the interface i. This last condition leads to the following surface boundary condition:

$$\left[-p+2\mu\frac{U}{L}\frac{\partial w}{\partial z}\right]_i = 0, \qquad \left[\mu\frac{U}{L}\frac{\partial \mathbf{v}}{\partial z}\right]_i = 0. \tag{2.3}$$

As the parameters Ro, E_v, E_h, and Fr become too small, the initial condition strategy fails, giving rise to a singular perturbation problem. Eliminating this singularity requires the use of asymptotic analysis. To proceed further with this problem we introduce the following assumptions:

$$Ro = O(Fr), \quad \omega = \frac{Ro^2}{Fr^2}, \quad E_h = \hat{E}_hRo^2, \quad E_v = \hat{E}_vRo^2, \tag{2.4}$$

and we next assume that the parameters $(St,\ \omega, \hat{E}_h\,, \hat{E}_v)$ are fixed and are of $O(1)$. Thus, the asymptotic expansion of $\mathbf{v}, w, p, \rho$ with respect to Ro provides, at the zeroth order, expansions:

$$
\begin{aligned}
&p_0 = p_0(z), \quad \varrho_0 = \varrho_0(z), \quad \mathbf{D}p_0 = 0,\\
&\mathbf{v}_0 = \frac{\omega}{\varrho_0}\mathbf{k}\Lambda\mathbf{D}p_1, \quad \frac{dp_0}{dz} = -Bo\varrho_0.
\end{aligned}
\tag{2.5}
$$

The last two equations are the geostrophic and the hydrostatic assumptions, respectively.

Since the pressure fluctuation is the basic dynamical characteristic of the current, we assume that the wave motion may be specified in terms of the perturbation p_1 deduced, after some calculations from (2.1). Thus at the order R_0 we obtain the so-called quasi-geostrophic potential equation (Pedlosky [12]):

$$
\frac{d_h}{dt}\left[\mathbf{D}^2 p_1 + \frac{\omega}{B_0}\left\{\frac{\partial}{\partial z}\left\{-\frac{\varrho_0(z)}{\varrho_0'(z)}\frac{\partial p_1}{\partial z}\right\} + \frac{\partial p_1}{\partial z}\right\} + \varrho_0(z)\beta\omega y\right] = 0,
\tag{2.6}
$$

where $\frac{d_h}{dt} = St\frac{\partial}{\partial t} + \mathbf{v_0}.\mathbf{D}$ and $\mathbf{D} = \frac{\partial}{\partial x}\,\mathbf{i} + \frac{\partial}{\partial y}\,\mathbf{j}$.

The problem is now reduced to the study of the pressure perturbation p_1 governed by (2.6). The associate boundary conditions are determined supposing that there exist two Ekman boundaries at the ocean-atmosphere interface and at the bottom. To remain within the Ekman-layer thickness as E_v is small, the matching procedure between layers leads, after some calculation, to the following boundary conditions:

$$
\frac{d_h}{dt}\left(\frac{\partial p_1}{\partial z}\right) = \begin{cases} \Gamma(0)\varkappa(0)\,\varpi\,\mathbf{D}^2(\varrho_0^{-1}p_1)_{z=0},\\ \varkappa(-1)\,\varpi\,\mathbf{D}^2(\varrho_0^{-1}p_1)_{z=-1}, \end{cases}
\tag{2.7}
$$

where

$$
\begin{aligned}
&\varkappa(z) = \left(\frac{\hat{E}_v}{2}\right)^{\frac{1}{2}}\varrho_0'(z)\varrho_0(z)^{-\frac{1}{2}},\\
&\varpi = \frac{Bo}{\omega} = O(1),\\
&\Gamma(z) = -1 + \frac{\mu\,\varrho_0(z)^{\frac{1}{2}}}{\mu\,\varrho_0(z)^{\frac{1}{2}} + \mu^a\,\varrho_0^a(z)^{\frac{1}{2}}} < 0,
\end{aligned}
\tag{2.8}
$$

with μ the coefficient of the dynamic viscosity. The superscript "a" refers to the atmospheric terms and indicates the signature of the atmospheric effects in the matching procedure.

In order to analyze how the solution of (2.6) behaves in the presence of the density gradient, we consider only the vertical structure of the pressure perturbation. The study of the vertical structure of the upper ocean can be useful because pressure and bulk temperatures or salinities tend to vary more along a vertical distance of a hundred meters than along a horizontal distance of a thousand kilometers. This holds true over many parts of the world's oceans where it is the case that vertical exchange processes within the water column are likely to affect local conditions much more rapidly than horizontal advection and horizontal mixing. It follows that for our purposes we treat the upper ocean layers as being homogeneous along the horizontal; thus we consider the propagation of monochromatic waves in the form

$$
p_1(x, y, z, t) = P(z)\exp\left\{i(kx + ly - \sigma t)\right\},
\tag{2.9}
$$

where k and l are the characteristic wave numbers and σ is the frequency.

Substituting (2.9) into (2.6) and (2.7), we obtain

$$\frac{d}{dz}\left(\frac{1}{\rho_0'(z)}\frac{dP(z)}{dz}\right)+\frac{\lambda}{\rho_0(z)}P(z)=0, \tag{2.10}$$

where

$$\lambda = K_h^2+\frac{k\beta}{\sigma} \tag{2.11}$$

is a separation parameter to be determined in the analysis below. $K_h^2=k^2+l^2$ is the horizontal wave number, and k and l are the wave numbers in the x- and y-directions; σ is the frequency of the free harmonic waves.

This is an eigenvalue Sturm–Liouville problem for the vertical structure of the pressure perturbation $P(z)$, where λ, assumed to be function of the wave frequency σ from (2.11), is the problem eigenvalue. The associated boundary conditions are

$$\left(-\sigma St+\mathbf{v_0}.\nabla\right)\frac{dP(0)}{dz}-\alpha_2 P(0)=0,$$

$$\left(-\sigma St+\mathbf{v_0}.\nabla\right)\frac{dP(-1)}{dz}-\alpha_1 P(-1)=0, \tag{2.12}$$

where $\mathbf{v_0}$ is given by (2.5), $\nabla = k\mathbf{i}+l\mathbf{j}$, and α_1 and α_2 are given by

$$\begin{aligned}\alpha_2 &= iK_h^2\left(\frac{\hat{E}_v}{2}\right)^{1/2}\Gamma(0)\varrho_0'(0)\varrho_0^{-1/2}(0),\\ \alpha_1 &= iK_h^2\left(\frac{\hat{E}_v}{2}\right)^{1/2}\varrho_0'(-1)\varrho_0^{-1/2}(-1).\end{aligned} \tag{2.13}$$

We note that the horizontal velocity $\mathbf{v}_0$ depends on the pressure perturbation p_1 (2.5). Thus, its presence at the boundary conditions (2.12) is a nonlinear contribution to the eigenvalue problem.

3. Vertical structure of monochromatic internal Rossby waves with the stratification profile. In this section we deal with the set of equations (2.10) and (2.12). The first fact which we must notice is that the solution $P(z)$ depends on the boundary conditions. In other words, we deal with the Sturm–Liouville problem, which is essentially the problem of determining the dependence of the general behavior of the solution $P(z)$ on the parameter λ (or σ) and the dependence of the eigenvalues of λ (or σ) on the complexity of the boundary conditions imposed on $P(z)$. This complexity is due, first, to the presence of the frequency σ in the boundary conditions that render the set of eigenfunctions nonorthogonal. In order to overcome this difficulty, the assumption $St=\varepsilon \ll 1$ is made, where St is the Strouhal parameter. This assumption means that we consider a quasi-stationary flow.

3.1. Case of a variable gradient stratification profile. In the case of a nonuniform density distribution along the vertical, we consider a density distribution in tanh form, which may approximate a typical oceanic profile (see Figure 3.1):

$$\rho_0(z)=\rho_{00}\left[1-N_m^2\tanh\left(\frac{z+h}{\delta}\right)\right], \tag{3.1}$$

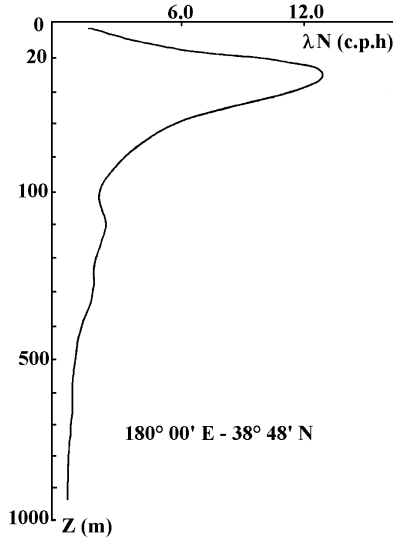


FIG. 3.1. *An example of a typical Brunt–Väisälä frequency profile (taken from* [6]*).*

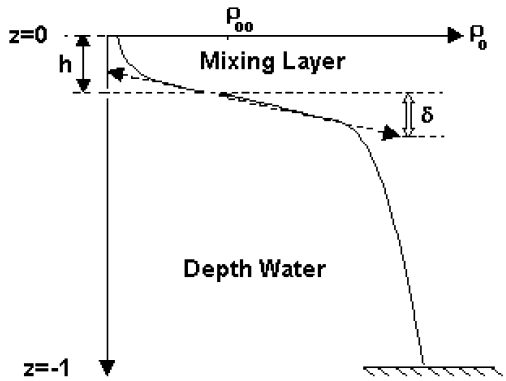


FIG. 3.2. *Schematic presentation of the density distribution in* tanh *form.*

where $N_m^2 \ll 1$ and ρ_{00} are constants. The parameter δ is the varying stratification parameter, which varies in the range $0 < \delta \leq 1$. It may roughly define the thickness of the layer along which $\rho_0(z)$ steeply varies (see Figure 3.2). An experimental study using a density distribution in a tanh profile was considered in Pawlak and Armi [9].

Following a classical procedure (see, for example, [13]), the wave equations (2.10) and (2.12) may be transformed into a suitable form by introducing in place of z a new variable ξ, defined as follows:

$$\xi = \frac{1}{J} \int_{-1}^{z} \left(-\frac{\rho_0(t)}{\rho_0'(t)} \right)^{1/2} dt, \tag{3.2}$$

with

$$J = \int_{-1}^{0} \left(-\frac{\rho_0(t)}{\rho_0'(t)} \right)^{1/2} dt. \tag{3.3}$$

With this form of J, the new variable ξ evolves in the range $[0,1]$ as z evolves in the range $[-1,0]$. Next, in place of the function $P(z)$ we introduce a new function:

$$X(\xi) = P(z) \sqrt[4]{\frac{-1}{\rho_0(z)\rho_0'(z)}}. \tag{3.4}$$

This yields the following modified Sturm–Liouville problem for $X(\xi)$:

$$\frac{d^2X(\xi)}{d\xi^2} - \left(\eta_1^2 + V(\xi,\delta)\right) X(\xi) = 0, \tag{3.5}$$

with

$$\frac{dX(1)}{d\xi} - aX(1) = 0 \quad \text{and} \quad \frac{dX(0)}{d\xi} - bX(0) = 0,$$

where $\eta_1^2 = \lambda.J^2$ (from (2.11) and (3.3)) and

$$V(\xi,\delta) = \left(-\rho_0(z)\rho_0'(z)\right)^{1/4} \frac{d^2}{d\xi^2}\left[\sqrt[4]{\frac{-1}{\rho_0(z)\rho_0'(z)}}\right],$$

$$\begin{aligned} a &= \frac{-\alpha_2 J}{\mathbf{v_0}.\nabla}\frac{\rho_0(0)}{\rho_0'(0)} - \frac{1}{4}(\rho_0\rho_0')\left[\frac{d}{d\xi}\left(\frac{-1}{\rho_0\rho_0'}\right)\right]_{\xi=1}, \\ b &= \frac{-\alpha_1 J}{\mathbf{v_0}.\nabla}\frac{\rho_0(-1)}{\rho_0'(-1)} - \frac{1}{4}(\rho_0\rho_0')\left[\frac{d}{d\xi}\left(\frac{-1}{\rho_0\rho_0'}\right)\right]_{\xi=0}, \end{aligned} \tag{3.6}$$

where $\mathbf{v_0}$ is given by (2.5) and $\nabla = k\mathbf{i} + l\mathbf{j}$. At this stage of analysis, we note that (2.10) and (2.12) correspond to the Sturm–Liouville eigenproblem, which is a problem of determining the dependence of the behavior of the solution $P(z)$ on the parameter λ (or σ). Thus, the presence of the frequency σ in the boundary conditions renders the set of eigenfunctions nonorthogonal. In order to overcome this difficulty, the assumption $St = \epsilon \ll 1$ is made, where St is the Strouhal parameter. This assumption means that we consider a quasi-stationary flow.

To proceed further with this problem, we have to find nontrivial solutions of the homogeneous set of equations (3.5), and we have to determine the values of the separation parameter λ (eigenvalues) for these nontrivial solutions (eigenfunctions). Thus, in the following analysis we consider only the case where η_1^2 is negative: $\eta_1^2 = -\mu_1^2$. Since η_1^2 is negative, we have

$$\lambda J^2 = k^2 + l^2 + \frac{k\beta}{\sigma_1} < 0. \tag{3.7}$$

Using a classical mathematical argument (see Ince [14]), the eigenfunctions $X(\xi)$ of the problem (3.5) can be given by

$$X(\xi) = c_1\cos(\mu_1\xi) + c_2\sin(\mu_1\xi) + \frac{1}{\mu_1}\int_{\varepsilon}^{\xi}\left[V(t,\delta)\sin\left(\mu_1(\xi-t)\right)X(t)\right]dt, \tag{3.8}$$

where $c_1$and c_2 are arbitrary and $0 \le \varepsilon \le 1$. Thus, from the first boundary condition at $\xi = 1$, we deduce that

$$c_2 = \frac{c_1 b}{\mu_1}, \tag{3.9}$$

and from the second condition at $\xi = 0$, we deduce the following eigenvalue equation:

$$\tan g(\mu_1) = \frac{c_1(b-a) + \int_0^1 |V(t,\delta)| \left[\cos(\mu_1 t) + \frac{a}{\mu_1}\sin(\mu_1 t)\right] X(t)dt}{c_1\left(\mu_1 + \frac{b}{\mu_1}\right) - \int_0^1 |V(t,\delta)| \left[\sin(\mu_1 t) - \frac{a}{\mu_1}\sin(\mu_1 t)\right] X(t)dt}. \tag{3.10}$$

The roots of this equation form a discrete spectrum ($\mu_{1k},\ k = 1, 2, \dots$). However, $X(\xi)$ is unknown; thus if μ_{1k} are sufficiently large and on using (3.9), a simple estimate of the solution (3.8) can satisfy an asymptotic form

$$X(\xi) = c_1 \cos(\mu_1 \xi) + \frac{g(\mu_1 \xi)}{\mu_1}, \tag{3.11}$$

where $g(\mu_1\xi)$ is a bounded function as $\mu_1 \longrightarrow \infty$. If we set this asymptotic form (3.11) in (3.10), we obtain

$$\tan(\mu_{1k}) \cong \frac{(b-a) + \frac{1}{2}\int_0^1 |V(t,\delta)|\, dt + O(\frac{1}{k})}{\mu_1 + O(\frac{1}{k})} \qquad (k = 1, 2, \dots). \tag{3.12}$$

By using the fixed point method, we can estimate the solution of the above equation. Thus, we obtain

$$\mu_{1k} \cong k\pi + \frac{(b-a) + \frac{1}{2}\int_0^1 |V(t,\delta)|\, dt}{k\pi} + O\left(\frac{1}{k^2}\right) \qquad (k = 1, 2, \dots). \tag{3.13}$$

The solution of (3.8) can be obtained by successive approximations in the form

$$X(\xi, \mu_1) = \sum_{k=0}^{\infty} x_k(\xi, \mu_1), \tag{3.14}$$

where

$$x_0(\xi, \mu_1) = c_1\left(\cos\mu_1\xi + \frac{b}{\mu_1}\sin\mu_1\xi\right), \tag{3.15}$$

$$x_{k+1}(\xi, \mu_1) = \frac{1}{\mu_1}\int_{\varepsilon}^{\xi} \sin\mu_1(\xi - t)V(t)x_k(t, \mu_1)dt. \tag{3.16}$$

If $|V(\xi)| \le A$, where A is an upper bound, we can prove by induction that

$$|x_{k+1}(\xi, t)| \le \frac{|c_1| + |c_1 b/\mu_1|}{k!}\frac{A^k\,|\xi - \varepsilon|^k}{\mu_1^k}, \qquad k = 1, 2, \dots . \tag{3.17}$$

The solution form, where $k = 0$, is given in (3.15).

We thus see from (3.17) that $x_{k+1}(\xi, t)$ is uniformly convergent in the interval $\xi \in [0, 1]$ as A is a bounded number, and that $x_{k+1}(\xi, t)$ is also an asymptotic expansion of $X(t, \xi)$ as $\mu_1 \to \infty$.

However, since the upper bound A becomes sufficiently large as δ diminishes, this leads to the increasing of $|x_{k+1}(\xi, t)|$ to a large wave amplitude. The same increasing feature may be derived from the Schwartz inequality. This establishes the existence

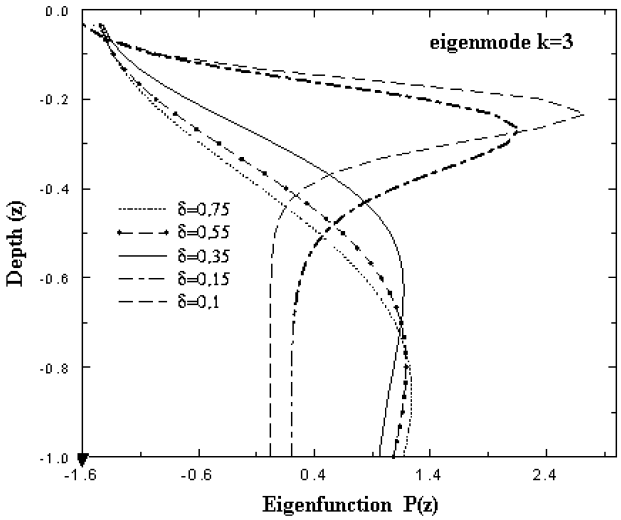


FIG. 3.3. *Variability in the structure eigenfunctions with the parameter δ for the eigenmode $k = 3$. The plots correspond to the numerical solution.*

of a fixed number such that $|x_{k+1}(\xi, t)| < M$. It remains to show from (3.16) that the solution amplitude verifies

$$|x_{k+1}(\xi,t)| \leq |c_1| \left(1 + \frac{b^2}{\mu_1^2}\right)^{1/2} + \frac{M}{\mu_1} \int_0^1 |V(t,\delta)|\, dt \tag{3.18}$$

and hence

$$M \leq |c_1| \sqrt{1 + \frac{b^2}{\mu_1^2}} \frac{1}{1 - \frac{1}{\mu_1} \int_0^1 |V(t,\delta)|\, dt}. \tag{3.19}$$

The validity of the last inequality depends on the behavior of the denominator. In this situation two problems arise. The first of these is that the upper bound must be positive; it follows that the expression in the denominator must be positive in turn. The second problem is the finding of a critical value δ_{cri} to ensure a bounded upper bound M. Then it will be necessary to impose restrictions on δ to prevent the denominator of (3.19) from vanishing. This consists of discarding the condition $\mu_1 = \int_0^1 |V(t,\delta)|\, dt$. This requirement involves the fact that the solution may hold in the interval $\delta_{cri} < \delta < 1$, while in the interval $0 < \delta \leq \delta_{cri}$ the upper bound M becomes sufficiently large so that the wave amplitude dominates the flow dynamics. Thus a significant increase of the energy is acquired as the wave approaches the layer of the maximum density gradient (Figure 3.3). The wave motion focuses at a certain depth $h\,(0 < h < 0.2)$. This depth is determined by the minimum of the characteristic scale of the varying stratification parameter δ (Figure 3.4). The growth of the solution amplitude is accompanied by the transformation of the vertical structure of internal wave modes and renders the vertical wave propagation unstable. The same features were also found in previous laboratory experiments (Stamp and Jacka [7] and Simpson and Linden [15]).

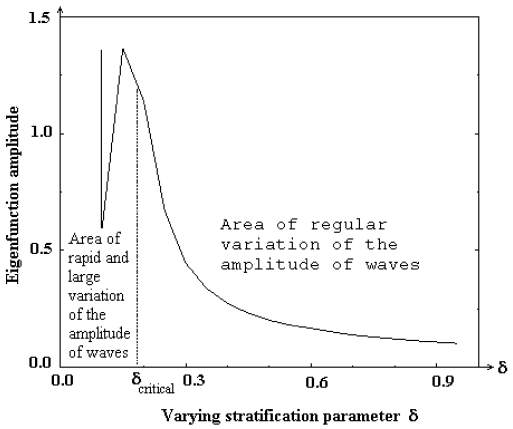


FIG. 3.4. *Growth rate of the eigenmodes as a function of the varying stratification parameter* δ.

The instability described here can be interpreted as a criterion in which the wave amplitudes grow, according to the parameter δ, to a level such that δ drops below a critical value δ_{cri}. Then the disturbances can grow exponentially, whence the primary wave amplitude exceeds a certain limiting value to such an extent that an effective discontinuity or front may develop (see Figures 3.3 and 3.4).

The exponential growth of internal wave amplitudes in horizontally inhomogeneous layers has been studied by Navrotsky and Simonenko [16] and Navrotsky [17]. They showed that such waves could produce effective mixing within the thermocline, and higher modes of internal waves can also play an important role in turbulence generation, especially in a multilayered stratification.

4. Finding the solutions for thin interface layers in the presence of turning-points. In this section we seek to obtain eigenvalues, as the stratification rate parameter δ is very small. For simplicity we proceed under the assumption that $\hat{E}_v \ll 1$ (Kraus [18]). Yet from the preceding section, the analysis of the vertical structure of the pressure perturbation explicitly exploits the thinness of these layers and does not reveal any particular problem due to the presence of the parameter $\hat{E}_v$ at the boundary conditions.

In the following we deal with the set of equations (3.5), which can be written as

$$\frac{d^2X(\xi)}{d\xi^2} + \lambda^2 q(\xi,\delta)X(\xi) = 0, \tag{4.1}$$

with the associated boundary conditions

$$\frac{dX(1)}{d\xi} - aX(1) = 0 \quad \text{and} \quad \frac{dX(0)}{d\xi} - bX(0) = 0, \tag{4.2}$$

where $\lambda^2 = 1/4\delta^4$ with $\delta \ll 1$, and the function $q(\xi,\delta)$ reads

$$q(\xi,\delta) = (\xi-\xi_1)(\xi-\xi_2)F(\xi_1,\xi_2), \tag{4.3}$$

with

$$F(\xi_1,\xi_2) = \frac{1}{\cosh^2\left(\frac{\xi_1-1+h}{\delta}\right)\cosh^2\left(\frac{\xi_2-1+h}{\delta}\right)},$$

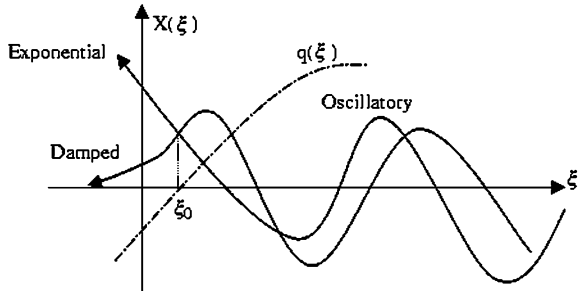


FIG. 4.1. *Schematic presentation of the solution behavior in presence of a turning-point* ξ_0.

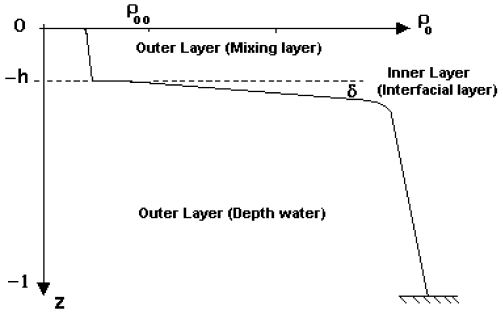


FIG. 4.2. *Three-layer schematic representation of the density, in the case when* $\delta \ll 1$.

$$\xi_1 = 1 - h - \delta \text{ argth } \sqrt{2 - 4\mu_1^2\delta^2}, \tag{4.4}$$

$$\xi_2 = 1 - h + \delta \text{ argth } \sqrt{2 - 4\mu_1^2\delta^2}. \tag{4.5}$$

The function $q(\xi, \delta)$ may vanish if the variable ξ is equal to ξ_1 or ξ_2, called turning-points. The core of the problem is to seek a stable solution, for a given fixed value of μ_1, when the stratification parameter δ is very small and when the function $q(\xi, \delta)$ changes the sign as ξ evolves from 0 to 1 on passing over the turning-points (see Figure 4.1). Elsewhere, as $\delta \to 0$, we deal with three cases (Figure 4.2):

(i) as $\xi > 1 - h$, $\frac{\xi - 1 + h}{\delta} \to +\infty$;
(ii) as $\xi > 1 - h$, $\frac{\xi - 1 + h}{\delta} \to +\infty$;
(iii) as $\xi = 1 - h$, $\frac{\xi - 1 + h}{\delta} = 0$.

Note that from (4.4) and (4.5), $q(\xi, \delta)$ may change the sign in the range

$$(\mu_1\delta) \in \left]\frac{-\sqrt{2}}{2}, \frac{-1}{2}\right[\cup \left]\frac{1}{2}, \frac{\sqrt{2}}{2}\right[. \tag{4.6}$$

To further proceed with this problem we use the WKB approximation. We seek solutions in the form

$$X(\xi) = \exp^{\pm i\phi(\xi)}. \tag{4.7}$$

Substituting the last equation into the wave equation (4.1), this yields the following

dispersion equation:

$$\frac{i\phi''(\xi)}{\phi'^2(\xi)} + \frac{q(\xi)}{\phi'^2(\xi)} - 1 = 0. \tag{4.8}$$

If we choose ϕ such that

$$\frac{\phi''(\xi)}{\phi'^2(\xi)} \ll 1 \quad \text{and} \quad \frac{q(\xi)}{\phi'^2(\xi)} - 1 = 0, \tag{4.9}$$

we deduce from the first condition the inequality

$$\frac{q'(\xi)}{2\sqrt{q(\xi)}} \ll q(\xi), \tag{4.10}$$

which remains verified as $\delta \to 0$. From (4.8) and the second condition in (4.9), ϕ may be approximated as

$$\phi(\xi) \cong \pm \int \sqrt{q(\xi)} + \frac{i}{4} \ln q(\xi); \tag{4.11}$$

then from (4.7), the approximate solution takes the form

$$X(\xi,\delta) \cong \frac{1}{\sqrt[4]{q(\xi)}} \left\{ \alpha \exp\left[i\lambda \int \sqrt{q(\xi)} d\xi \right] + \beta \exp\left[-i\lambda \int \sqrt{q(\xi)} d\xi \right] \right\}, \tag{4.12}$$

where α and β are arbitrary and $\lambda = 1/2\delta^2$.

The validity of this asymptotic form given in (4.12) depends on the sign of the function $q(\xi)$. It has an oscillatory character if $q(\xi)$ is positive and has a monotonic character if $q(\xi)$ is negative. Hence, there exists a point ξ_0 such that, at the vicinity of this point, the transition takes place from one type of behavior to the other (see Figure 4.1).

Thus, if $q(\xi) > 0$, the solution is oscillatory, and may read

$$X_{1(out)}(\xi,\delta) \cong \frac{1}{\sqrt[4]{q(\xi)}} a_1 \sin\left[\lambda \int \sqrt{q(\xi)} d\xi \right] + b_1 \cos\left[-\lambda \int \sqrt{q(\xi)} d\xi \right]. \tag{4.13}$$

If $q(\xi) < 0$, the solution is monotonic, and may read

$$X_{2(out)}(\xi,\delta) \cong \frac{1}{\sqrt[4]{-q(\xi)}} a_2 \exp\left[\lambda \int \sqrt{-q(\xi)} d\xi \right] + b_2 \exp\left[-\lambda \int \sqrt{-q(\xi)} d\xi \right]. \tag{4.14}$$

The solutions $X_{1(out)}(\xi,\delta)$ and $X_{2(out)}(\xi,\delta)$ are valid above and below the turning-points and are called *outer solutions*. In the vicinity of the turning-points ξ_1 and ξ_2, we use Langer's transformation (see Nayfeh [19]):

$$x = \varphi(\xi), \qquad \Phi = \psi(x) X. \tag{4.15}$$

The transformation (4.15) carries (4.1) into

$$\frac{d^2\Phi}{dx^2} + \frac{1}{\varphi'^2}\left(\varphi'' - 2\frac{\varphi'\psi'}{\psi} \right) \frac{d\Phi}{dx} + \frac{1}{\varphi'^2}\left(\lambda^2 q + \psi \left(\frac{\psi'}{\psi^2} \right)' \right) \Phi = 0. \tag{4.16}$$

We determine ψ and φ so that

$$\varphi'' - 2\frac{\varphi'\psi'}{\psi} = 0 \quad \Rightarrow \quad \psi = \sqrt{\varphi'} \tag{4.17}$$

and

$$\frac{-q}{\varphi'^2} = \varphi, \tag{4.18}$$

so that (4.16) becomes

$$\frac{d^2\Phi}{dx^2} - \lambda^2 x\Phi = v(x)\Phi, \tag{4.19}$$

where

$$v(x) = \frac{1}{2}\frac{\varphi'''}{\varphi'^3} - \frac{3}{4}\frac{\varphi''^2}{\varphi'^4}. \tag{4.20}$$

At the vicinity of the turning-point ξ_1 the solutions of (4.18) are

$$\frac{2}{3}(\varphi_1)^{3/2} = \int_{\xi_1}^{\xi} \sqrt{(t-\xi_1)(\xi_2-t)F(\xi_1,\xi_2)}dt \qquad \text{if } \xi > \xi_1, \tag{4.21}$$

$$\frac{2}{3}(-\varphi_1)^{3/2} = \int_{\xi}^{\xi_1} \sqrt{(\xi_1-t)(\xi_2-t)F(\xi_1,\xi_2)}dt \qquad \text{if } \xi < \xi_1. \tag{4.22}$$

In the same manner, at the vicinity of the turning-point ξ_2 the solutions of (4.18) read

$$\frac{2}{3}(-\varphi_2)^{3/2} = \int_{\xi}^{\xi_2} \sqrt{(t-\xi_1)(t-\xi_2)F(\xi_1,\xi_2)}dt \qquad \text{if } \xi > \xi_2, \tag{4.23}$$

$$\frac{2}{3}(\varphi_2)^{3/2} = \int_{\xi_2}^{\xi} \sqrt{(t-\xi_1)(\xi_2-t)F(\xi_1,\xi_2)}dt \qquad \text{if } \xi < \xi_2. \tag{4.24}$$

From (4.21) as $\xi \to \xi_1$, $\varphi_1 \to \sqrt[3]{(\xi_2-\xi_1)F(\xi_1,\xi_2)}.(\xi-\xi_1)$; hence from (4.20) the function $v = O(1)$. Thus as λ is large, this leads to the approximated equation

$$\frac{d^2\Phi}{dx^2} - \lambda^2 x\Phi = 0, \tag{4.25}$$

whose solution is

$$\Phi = c_1 Ai(\lambda^{2/3}x) + c_2 Bi(-\lambda^{2/3}x), \tag{4.26}$$

where c_1 and c_2 are constants of integration. Ai and Bi are the Airy functions of the first and second kind. Applying the previous results and (4.15), the asymptotic solution for $\xi > \xi_1$ reads

$$X_{1(in)}(\xi,\delta) \cong \frac{1}{\sqrt{\varphi_1'}}\left\{c_1^{(1)} Ai\left[\lambda^{2/3}\varphi_1(\xi)\right] + c_2^{(1)} Bi\left[-\lambda^{2/3}\varphi_1(\xi)\right]\right\}. \tag{4.27}$$

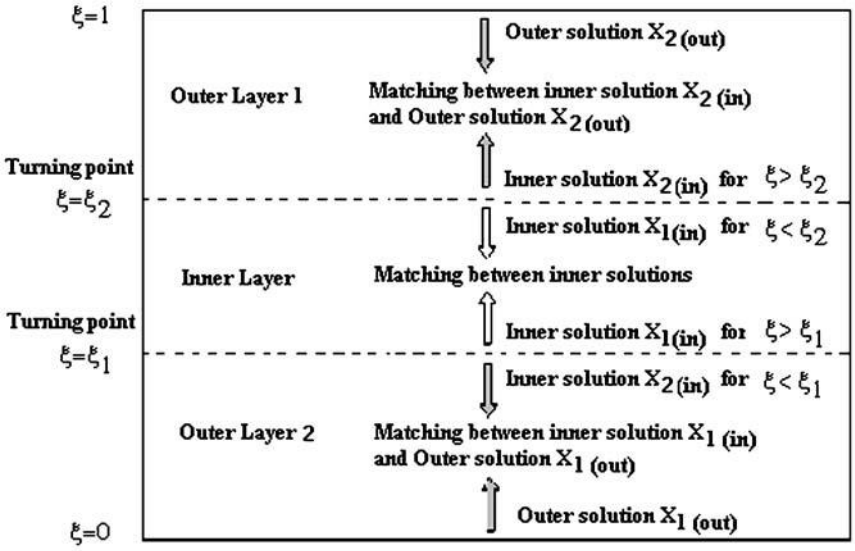


FIG. 4.3. *Sketch of the matching procedure in the presence of turning-points.*

Similarly, the asymptotic solution for $\xi < \xi_2$ reads

$$X_{2(in)}(\xi,\delta) \cong \frac{1}{\sqrt{\varphi_2'}}\left\{c_1^{(2)} Ai\left[\lambda^{2/3}\varphi_2(\xi)\right] + c_2^{(2)} Bi\left[-\lambda^{2/3}\varphi_2(\xi)\right]\right\}. \tag{4.28}$$

These solutions $X_{1(in)}$ and $X_{2(in)}$ (see Figure 4.3) are available for ξ in the range $\xi_1 + \varepsilon < \xi < \xi_2 - \varepsilon$ ($\varepsilon \ll 1$) and are called inner solutions. φ_1 and φ_2 are given by (4.21) and (4.24); $c_1^{(1)}, c_2^{(1)}, c_1^{(2)}, c_2^{(2)}$ are constants of integration.

The matching procedure (Figure 4.3) between the outer solutions given in (4.13) and (4.14) and the inner solutions given in (4.27) and (4.28) may lead to the relationship that exists between the constants $c_1^{(1)}, c_2^{(1)}, c_1^{(2)}, c_2^{(2)}$ and a_1, a_2, b_1, b_2.

4.1. Matching procedure between inner solutions. To match the solutions given in (4.27) and (4.28) we first use the asymptotic forms of large positive arguments of the Airy functions (Nayfeh [19]) given by

$$\begin{aligned} Ai(z) &= \frac{1}{2\sqrt{\pi}} z^{-1/4} \exp\left(-\frac{2}{3}z^{3/2}\right)\left[1 + O\left(\frac{2}{3}z^{3/2}\right)^{-1}\right], \\ Bi(z) &= \frac{1}{2\sqrt{\pi}} z^{-1/4} \exp\left(\frac{2}{3}z^{3/2}\right)\left[1 + O\left(\frac{2}{3}z^{3/2}\right)^{-1}\right]. \end{aligned} \tag{4.29}$$

These last assumptions transform (4.27) and (4.28) into

$$X_{1(in)}(\xi,\delta) \cong \frac{\lambda^{-1/6}}{\sqrt[4]{-q(\xi)}\sqrt{\pi}}\left\{\frac{c_1^{(1)}}{2}\exp\left(-\frac{2}{3}\lambda\varphi_1^{3/2}\right) + c_2^{(1)}\exp\left(\frac{2}{3}\lambda\varphi_1^{3/2}\right)\right\}, \tag{4.30}$$

$$X_{2(in)}(\xi,\delta) \cong \frac{\lambda^{-1/6}}{\sqrt[4]{-q(\xi)}\sqrt{\pi}}\left\{\frac{c_1^{(2)}}{2}\exp\left(-\frac{2}{3}\lambda\varphi_2^{3/2}\right) + c_2^{(2)}\exp\left(\frac{2}{3}\lambda\varphi_2^{3/2}\right)\right\}. \tag{4.31}$$

Equating (4.30) and (4.31) gives

(4.32)
$$\frac{c_1^{(1)}}{2}\exp\left(-\frac{2}{3}\lambda\varphi_1^{3/2}\right)+c_2^{(1)}\exp\left(\frac{2}{3}\lambda\varphi_1^{3/2}\right)=\frac{c_1^{(2)}}{2}\exp\left(-\frac{2}{3}\lambda\varphi_2^{3/2}\right)+c_2^{(2)}\exp\left(\frac{2}{3}\lambda\varphi_2^{3/2}\right).$$

If we introduce

$$\Delta=\frac{2}{3}\lambda(\varphi_1^{3/2}+\varphi_2^{3/2})=\int_{\xi_1}^{\xi_2}\sqrt{(t-\xi_1)(\xi_2-t)F(\xi_1,\xi_2)}dt, \tag{4.33}$$

we obtain from (4.32)

$$\begin{aligned}\frac{c_1^{(1)}}{2}+c_2^{(1)}&=\frac{c_1^{(2)}}{2}\exp(-\Delta)+c_2^{(2)}\exp(\Delta),\\ c_2^{(1)}-\frac{c_1^{(1)}}{2}&=\frac{c_1^{(2)}}{2}\exp(-\Delta)-c_2^{(2)}\exp(\Delta),\end{aligned} \tag{4.34}$$

so that

$$\begin{aligned}c_2^{(1)}&=c_1^{(2)}\exp(-\Delta),\\ c_2^{(2)}&=c_1^{(1)}\exp(-\Delta).\end{aligned} \tag{4.35}$$

Thus for (4.30) and (4.31) to have a bounded solution as λ is large, it is necessary that $c_2^{(1)}$ and $c_2^{(2)}$ vanish. It follows that Δ must be strictly positive.

Thus, on using the ξ_1 and ξ_2 expressions given by (4.4) and (4.5), the integration of (4.33) gives

$$\Delta=\frac{2}{3}\lambda(\varphi_1^{3/2}+\varphi_2^{3/2})=\frac{\pi}{2}\frac{\delta^2\mathrm{argth}^2\sqrt{2-4\mu_1^2\delta^2}}{\cosh(\frac{\xi_1-1+h}{\delta})\cosh(\frac{\xi_2-1+h}{\delta})} \tag{4.36}$$

and shows that Δ is positive for all values of $\delta>0$.

4.2. Matching procedure between inner and outer solutions. To match the outer solutions (4.13) and (4.14) with the inner solutions (4.27) and (4.28) (see Figure 4.3), we use a method similar to that employed in subsection 4.1, the asymptotic forms of large but negative arguments of the Airy functions given by

$$\begin{aligned}Ai(-z)&=\frac{1}{\sqrt{\pi}}z^{-1/4}\sin\left(\frac{2}{3}z^{3/2}+\frac{\pi}{4}\right),\\ Bi(-z)&=\frac{1}{\sqrt{\pi}}z^{-1/4}\cos\left(\frac{2}{3}z^{3/2}+\frac{\pi}{4}\right).\end{aligned} \tag{4.37}$$

These last assumptions transform (4.27) and (4.28) into

$$\begin{aligned}X_{1(in)}(\xi,\delta)\cong{}&\frac{\lambda^{-1/6}}{\sqrt[4]{-q(\xi)}\sqrt{2\pi}}(c_1^{(1)}-c_2^{(1)})\sin\left[\frac{2}{3}\lambda(-\varphi_1)^{2/3}\right]\\ &+(c_1^{(1)}+c_2^{(1)})\cos\left[\frac{2}{3}\lambda(-\varphi_1)^{2/3}\right],\end{aligned} \tag{4.38}$$

$$X_{2(in)}(\xi,\delta) \cong \frac{\lambda^{-1/6}}{\sqrt[4]{-q(\xi)}\sqrt{2\pi}} \left\{ (c_3^{(1)} - c_4^{(1)}) \sin\left[\frac{2}{3}\lambda(-\varphi_2)^{2/3}\right] + (c_3^{(1)} + c_4^{(1)}) \cos\left[\frac{2}{3}\lambda(-\varphi_2)^{2/3}\right] \right\}. \tag{4.39}$$

We shall now connect these last asymptotic representations with those obtained in the outer regions given by (4.13) and (4.14). We obtain the following relationships:

$$(a_1, b_1) = \frac{\lambda^{-1/6}}{\sqrt{2\pi}} \left\{ (c_1^{(1)} + c_2^{(1)}), (c_1^{(1)} - c_2^{(1)}) \right\}, \tag{4.40}$$

$$(a_2, b_2) = \frac{\lambda^{-1/6}}{\sqrt{2\pi}} \left\{ (c_1^{(2)} + c_2^{(2)}), (c_1^{(2)} - c_2^{(2)}) \right\}, \tag{4.41}$$

where $c_2^{(1)}$ and $c_2^{(2)}$ are given in (4.35)

5. Solving the solutions for thin interface without turning-points. From (3.5) and from the density profile given by (3.1), the wave equation reads

$$\frac{d^2X(\eta)}{d\eta^2} - \frac{1}{4}\left(\tanh^2\eta - [2 - 4\delta^2\mu_1^2]\right) X(\eta) = 0, \tag{5.1}$$

where $\eta = \frac{\xi - 1 + h}{\delta}$ and $\delta \ll 1$. The associated boundary condition with assumption $\hat{E}_v \ll 1$ reads

$$X(\eta) = 0 \quad \text{as} \quad \eta \to \pm\infty.$$

Note that, for a given fixed value of μ_1, the expression $\tanh^2\eta - (2 - 4\delta^2\mu_1^2)$ cannot change sign in the range

$$(\mu_1\delta) \in \left[\frac{-1}{2}, \frac{1}{2}\right]. \tag{5.2}$$

Using the transformation

$$X(\eta) = \frac{1}{\cosh^a\eta}\Phi(u), \tag{5.3}$$

where a is a factor to be determined in what follows and u is given by

$$u = \frac{1}{2}(1 + \tanh\eta), \tag{5.4}$$

the substitution of these transformations into (5.1) leads to the following hypergeometric differential equation for Φ:

$$u(u-1)\Phi''(u) + (1+a)(1-2u)\Phi'(u) - \left[\frac{1}{4} + a(a+1)\right]\Phi(u) = 0, \tag{5.5}$$

where $u = 0$ and $u = 1$ are singular points as $\eta \to \pm\infty$ and a verifies

$$a = \pm\frac{1}{4}\sqrt{1 - 4\mu_1^2\delta^2}. \tag{5.6}$$

In the following we will keep only the negative sign of the parameter a to ensure bounded solutions as u is equal to 0 or 1. Thus the analytical solution (Abramovitz and Stegun [20] and Bender and Orzag [21]), which is finite at $u = 0$ ($\eta \to -\infty$) satisfying the condition $\Phi(u = 0) = 0$, is given by

$$\Phi_0(u) = Au^{-a} F\left(\frac{1}{2}, \frac{1}{2}; 1 - a; u\right), \tag{5.7}$$

where F is the hypergeometric function given by Gauss series

$$F(\alpha, \beta; \gamma, z) = \sum_{n=0}^{N} \frac{(\alpha)_n (\beta)_n}{(\gamma)_n} \frac{z^n}{n!}. \tag{5.8}$$

The analytical solution, which is finite at $u = 1$ ($\eta \to +\infty$) satisfying the condition $\Phi(u = 1) = 0$, is given by

$$\Phi_1(u) = B(1 - u)^{-a} u^{-a} F\left(\frac{1}{2} + a, \frac{1}{2} + a; 1 - a; 1 - u\right), \tag{5.9}$$

where A and B are constants of integration. They may be determined using the normalization formula $\int_0^1 \Phi^2 du = 1$.

We note that the factor $(-a)$ present in (5.7) and (5.9) must be positive, in order that the solutions $\Phi_0(u)$ and $\Phi_1(u)$ be finite as u is equal to 0 or 1.

For the solution to be continuous at $u = 1$, we shall transform (5.7) in the neighborhood of $u = 1$ into two independent solutions, by using the linear transformation formulas given in [20, p. 559]. This step is necessary to transform (5.7) in power of $(1 - u)$ into a convenient form for matching (5.7) with (5.9). We obtain

$$\begin{aligned} \Phi_{01}(u) = \alpha_0^{(1)} &\left[Au^{-a} F\left(\frac{1}{2}, \frac{1}{2}; 1 + a; 1 - u\right)\right] \\ &+ \alpha_0^{(2)} \left[A(1 - u)^{-a} u^{-a} F\left(\frac{1}{2} - a, \frac{1}{2} - a; 1 - a; 1 - u\right)\right], \end{aligned} \tag{5.10}$$

with

$$\alpha_0^{(1)} = \frac{\Gamma(1 - a)\,\Gamma(-a)}{\Gamma(\frac{1}{2} - a)\Gamma(\frac{1}{2} - a)}, \tag{5.11}$$

$$\alpha_0^{(2)} = \frac{\Gamma(1 - a)\,\Gamma(a)}{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2})}, \tag{5.12}$$

where Γ is the gamma function.

In order to determine a uniformly valid solution as u varies from 0 to 1, we must connect the solutions given in (5.9) and (5.10).

However, the matching will be valid only if the first term of (5.10) vanishes, i.e., $\alpha_0^{(1)} = 0$. It follows that, since the gamma function satisfies the property

$$\frac{1}{\Gamma(-n)} = 0 \qquad \text{for } n \geq 0, \tag{5.13}$$

then for n any positive integer ($n \geq 0$), the coefficient $\alpha_0^{(1)}$ (see 5.11) may vanish by setting

$$\frac{1}{2} - a = -n \qquad (n = 0, 1, \dots). \tag{5.14}$$

On using (5.6), this leads to the following relationship between the parameters δ and μ_1:

$$(\mu_1 \delta)_n = \pm\sqrt{n^2 + n + \frac{1}{4}} \qquad (n = 0, 1, 2, \dots). \tag{5.15}$$

As $n \geqslant 1$ the values of $(\mu_1\delta)_n$ become much larger than the lowest and highest limits of the interval given by (5.2). If this is the case, we have to deal with a turning-point problem. It turns out that the last analysis falls down.

Thus, the criterion of our analysis validity given in (5.2) will be satisfied only for the case $n = 0$. If this limiting requirement is met, then, from (5.15), the relationship between the eigenvalue μ_1 and the interface thickness δ takes the form $(\mu_1\delta)_0 = \pm\frac{1}{2}$. The associated solution $\Phi_{01}(u)$ reads

$$\Phi_{01}(u) = \alpha_0^{(2)} \left[A(1-u)^{-a} u^{-a} F\left(-n, -n; 1-a; 1-u\right)\right]. \tag{5.16}$$

6. Observational evidence and discussion. Several of the characteristics of large-amplitude internal waves in the presence of a strong stratification or of a thin interface, as discussed in the previous sections, can be observed in the real ocean.

In the Strait of Gibraltar, experiments were conducted in 1985–1986 to examine the structure of the interface layer between the inflowing Atlantic waters and the outflowing Mediterranean waters in this strait [26]. It was found that the interface is 60–100 m thick, with a strong vertical salinity gradient identified by fitting individual salinity profiles to a piecewise-linear, three-layer model. The interface is deeper, thicker, fresher, and colder on the west end of the strait than in the Narrows, where there is a minimum in thickness and a maximum in salinity gradient. Property variations in all three layers are also cast in terms of the three principal water types involved in the exchange. The complexity of interaction between the interface and the upper and lower layers argues against the use of two-layer models to characterize the exchange through the Strait of Gibraltar (see Bray, Ochoa, and Kinder [26] for more details).

In the Mozambique Channel, internal waves were observed in the narrowest passage between Mozambique and Madagascar [27]. By using the ADCP-observations and CTD-profiles,[1] a strong pycnocline has been found. Below the pycnocline, this is the cross-channel averaged value of $N(z)$. For the upper layer the cross-channel differences in pycnocline depth lead to smearing of the pycnocline when $N(z)$ is averaged. The pattern becomes particularly complicated due to beam scattering at the

[1] ADCP = acoustic Doppler current profiler; CTD = conductivity, temperature, and depth recorder.

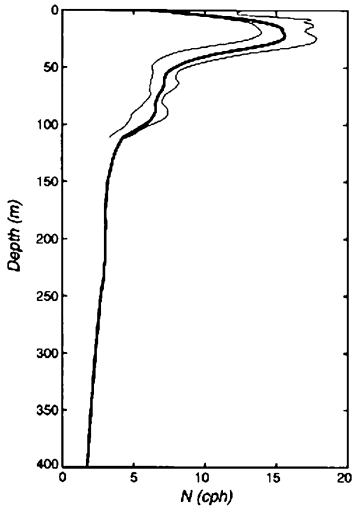


FIG. 6.1. *Measured mean buoyancy frequency profile N (cph) over the upper slope and shelf during summer PRIMER (taken from* [23]*; reproduced/modified by permission of American Geophysical Union). Also shown are the maximum and minimum N values at each depth from the horizontally averaged SeaSoar sections.*

pycnocline and the multiple reflections back into the basin. The authors argue that multiple reflections in a basin of sloping walls can in principle lead to the formation of internal-wave attractors, as was shown in Maas et al. [28].

In several places, for example, in the Andaman Sea (Osborne and Burch [29]), near a shelfbreak of the Sea of Japan (Navrotsky et al. [22]), in the Sulu Sea (Apel et al. [24]), and at the Mascarene Ridge (Konyaev, Sabinin, and Serebryany [30]), large-amplitude internal solitary waves are observed; that is, their amplitudes and the typical length scale of the vertical stratification are of the same order.

Experimental work has been presented by Colosi et al. [23] to describe the internal tide and high-frequency internal waves observed in the moored array data during the summer Shelfbreak Primer study. The summer Shelfbreak Primer study was conducted between July 26 and August 5, 1996. They found that in situ measurements and synthetic aperture radar (SAR) imagery show that packets of high-frequency nonlinear internal waves are generated near the shelfbreak during stratified conditions (late spring to early fall) and tend to propagate onshelf, while the SeaSoar data showed the presence of a shallow thermocline (pycnocline) over the shelf (see Figure 6.1), which allowed large-amplitude high-frequency internal waves of depression to form and propagate onshelf. The internal waves within the packets tend to become soliton-like, with large amplitudes, short wavelengths, and high frequencies near the local buoyancy frequency N.

Large-amplitude internal waves were also observed by Yessy and Masaki [31]. They used the monitoring result of internal wave detection in ERS-1/2 SAR and Topex/Poseidon (T/P) images over the southwest coast of Japan (see Figure 6.2). The left panel of Figure 6.2 shows two packets of internal waves at the south coast of Tsushima Island. In the right panel, the vertical transect of the internal wave clearly identifies eight crests of waves. The internal waves were propagated southeastward with lengths of wave crest around 17 km and length between crests from 375 m to 750 m.

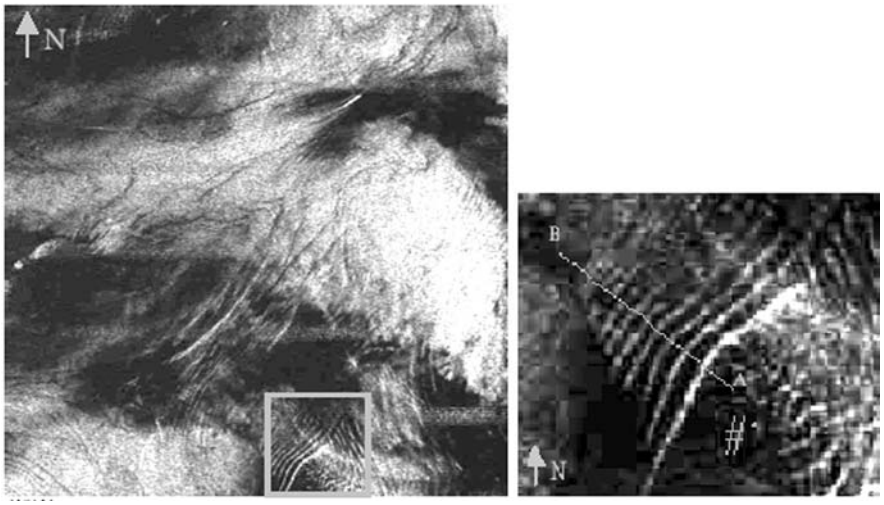


FIG. 6.2. *Internal waves around the southwest coast of Japan observed using ERS-*1/2 *SAR and T/P images (date: August,* 1993*). The right panel corresponds to the enlargement of rectangle area in the left panel (image from* [31] *by permission, and courtesy of the European Remote Sensing satellite).*

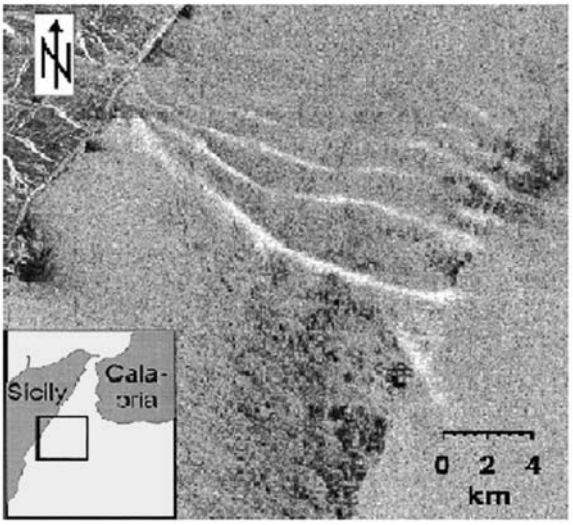


FIG. 6.3. *Radar image of a region south of the Strait of Messina, acquired by the SAR of the ERS-*2 *on August* 22, 1997*. The image shows sea surface manifestations of a train of southward propagating internal solitary waves (image taken from* [32] *and used by permission of the American Meteorological Society).*

Figure 6.3 presents a radar image showing sea surface manifestations of a train of southward propagating internal solitary waves south of the Strait of Messina in the Mediterranean Sea (see Vlasenko, Brandt, and Rubino [32] for details).

7. Conclusion. The primary purpose of this paper is to study the motion of long internal waves in a density-stratified fluid even for a very thin density interface. A theoretical model is then presented to establish stable solutions of these waves

in the presence or absence of turning-points at the interfacial layer. A prominent feature of our investigation is finding asymptotically stable and bounded solutions as the density gradient is very large. The results show that, in the case of variable stratification, a transformation of the vertical structure of internal wave occurs. This structure transformation is accompanied by the growth of the wave amplitude and steepness at the pycnocline. Here, there is a close similarity between this feature and the phenomenon of trapping internal waves by inhomogeneity of stratification discussed in [6]. Waves with the initial frequency σ_0 are trapped by a layer with the maximum effective Brunt-Väisälä frequency ($N_{ef} = N(z)/(1 - kU/\sigma_0)$) equal to σ_0. This trapping may concentrate the waves in a bandwidth in the vicinity of the pycnocline and give rise to a triple-layered area.

The presented model takes into account this layering feature, and permits us to obtain satisfactory results in the more complex situation even where turning-points exist. The model takes advantage of the analysis of internal waves filtered in quasi-geostrophic equations and based on Navier–Stokes equations under the Boussinesq and hydrostatic assumptions, with vertically varying stratification. Successive estimation of the velocity field and pressure leads to a single equation of the vertical structure of the pressure perturbation. An eigenvalue problem of Sturm–Liouville type is presented for the modal structures and frequencies of the eigenmode oscillations. The specification of these eigenmodes requires determination from boundary conditions that are complicated by the presence of the eigenvalues. This complication gives rise to the nonorthogonality of the boundary value problem verified by the pressure perturbation. Thus, to deal with this problem, an asymptotic analysis is used. The results show that a significant growth of the solution amplitude appears as the characteristic scale of a varying stratification parameter named δ approaches a critical value δ_{cri}, giving rise to the formation of an interfacial wave that dominates the flow. To deal with this interface, two cases are discussed of describing bounded internal solutions even for a very thin interface. In the first case it is assumed that the turning-points exist; thus we derive inner solutions at the vicinity of the turning-points and outer solutions above and below these points. A requirement of appropriate interfacial conditions provides a general stable matching solution, even for very small values of the parameter δ. In the second case no turning-points are allowed, and a two-layer discontinuous gradient system is examined. When matching solutions above and below the interface, a critical value of the parameter δ is required to obtain a uniform and valid solution in the entire domain.

REFERENCES

[1] H. J. S. Fernando and J. C. R. Hunt, *Turbulence and mixing at shear-free density interfaces. Part* I. *A theoretical model*, J. Fluid Mech., 347 (1997), pp. 197–234.

[2] K. Shafer Smith and G. Vallis, *The scales and equilibration of midocean eddies: Freely evolving flow*, J. Phys. Oceanogr., 131 (2000), pp. 554–571.

[3] A. Ya. Basovich and L. Sh. Tsimring, *Internal waves in a horizontally inhomogeneous flow*, J. Fluid Mech., 142 (1983), pp. 233–249.

[4] I. A. Hannoun, H. J. S. Fernando, and E. J. List, *Turbulence structure near a sharp density interface*, J. Fluid Mech., 189 (1988), pp. 189–209.

[5] O. M. Phillips, *Dynamics of the Upper Ocean*, Cambridge University Press, Cambridge, UK, 1977.

[6] S. I. Badulin, V. I. Shrira, and L. S. H. Tsimring, *The trapping and vertical focusing of internal waves in a pycnocline due to the horizontal inhomogeneities of density and currents*, J. Fluid. Mech., 158 (1985), pp. 199–218.

[7] A. P. Stamp and M. Jacka, *Deep water internal solitary waves*, J. Fluid. Mech., 305 (1995), pp. 347–371.

[8] A. Ouahsine, *Vertical structure and stability in a mathematical model of the ocean internal waves*, Internat. J. Engrg. Sci., 34 (1996), pp. 1311–1326.
[9] G. Pawlak and L. Armi, *Vortex dynamics in spatially accelerating shear layer*, J. Fluid. Mech., 376 (1998), pp. 347–371.
[10] P. A. Bois, *Asymptotic aspects for the Boussinesq approximation for gases and liquids*, J. Geophys. Astrophys. Fluid Dyn., 58 (1991), pp. 45–55.
[11] A. E. Gill, *Atmosphere-Ocean Dynamics*, International Geophysics Series 30, Academic Press, New York, 1982.
[12] J. Pedlosky, *Geophysical Fluid Dynamic*, 2nd ed., Springer-Verlag, Heidelberg, 1987.
[13] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, McGraw-Hill, New York, 1954.
[14] E. L. Ince, *Ordinary Differential Equations*, Dover Publications, New York, 1956.
[15] J. E. Simpson and P. F. Linden, *Frontogenesis in a fluid with horizontal density gradients*, J. Fluid. Mech., 202 (1989), pp. 1–16.
[16] V. V. Navrotsky and S. V. Simonenko, *Generation of internal waves near the shelf boundary*, in Proceedings of the Conference on Pacific Ocean Environment and Exploration, Okinawa, Japan, 1992, Vol. 2, pp. 1269–1274.
[17] V. V. Navrotsky, *Mixing caused by internal waves and turbulence: A comparative analysis*, J. Marine Systems, 21 (1999), pp. 131–145.
[18] E. B. Kraus, *Modelling and Prediction of the Upper Layers of the Ocean*, Pergamon Press, Oxford, UK, 1977.
[19] A. H. Nayfeh, *Perturbation Methods*, John Wiley & Sons, New York, 1973.
[20] M. Abramovitz and I. Stegun, *Handbook of Mathematical Functions*, Dover Publications, New York, 1965.
[21] C. M. Bender and S. A.Orzag, *Advanced Mathematical Methods for Scientists and Engineers*, Internat. Ser. Pure Appl. Math., McGraw-Hill, New York, 1978.
[22] V. V. Navrotsky, I. D. Lozovatsky, E. P. Pavlova, and H. J. S. Fernando, *Observations of internal waves and thermocline splitting near a shelf break of the Sea of Japan (East Sea)*, Continental Shelf Res., 24 (2004), pp. 1375–1395.
[23] J. A. Colosi, R. C. Beardsley, J. F. Lynch, G. Gawarkiewicz, C.-S. Chiu, and A. Scotti, *Observations of nonlinear internal waves on the outer New England continental shelf during the summer Shelfbreak Primer study*, J. Geophys. Res., 106 (2001), pp. 9587–9601.
[24] J. R. Apel, J. R. Holbrook, A. K. Liu, and J. J. Tsai, *The Sulu Sea internal soliton experiment*, J. Phys. Oceanogr., 15 (1985), pp. 1625–1651.
[25] P. Brandt, A. Rubino, W. Alpers, and J. O. Backhaus, *Internal waves in the Strait of Messina studied by a numerical model and synthetic aperture radar images from the ERS* 1/2 *satellites*, J. Phys. Oceanogr., 27 (1997), pp. 648–663.
[26] N. A. Bray, J. Ochoa, and T. H. Kinder, *The role of the interface in exchange through the Strait of Gibraltar*, J. Geophys. Res., 100 (1995), pp. 10755–10776.
[27] A. M. M. Manders, L. R. M. Maas, and T. Gerkema, *Observations of internal tides in the Mozambique Channel*, J. Geophys. Res., 109 (2004), paper C12034.
[28] L. R. M. Maas, D. Benielli, J. Sommeria, and F.-P. A. Lam, *Observation of an internal wave attractor in a confined, stably stratified fluid*, Nature, 388 (1997), pp. 557–561.
[29] A. R. Osborne and T. I. Burch, *Internal solitons in the Andaman Sea*, Science, 208 (1980), pp. 451–269.
[30] K. V. Konyaev, K. D. Sabinin, and A. N. Serebryany, *Large amplitude internal waves at the Mascarene Ridge in the Indian Ocean*, Deep-Sea Res., 42 (1995), pp. 2075–2091.
[31] A. Yessy and O. Masaki, *Internal wave observation in southwest coast of Japan using ERS-* 1/2 *and Topex/Poseidon images*, in Proceedings of the 23rd Asian Conference on Remote Sensing (ACRS 2002), Kathmandu, Nepal, 2002.
[32] V. Vlasenko, P. Brandt, and A. Rubino, *Structure of large-amplitude internal solitary waves*, J. Phys. Oceanogr., 30 (2000), pp. 2172–2185.

SIAM J. APPL. MATH.
Vol. 65, No. 5, pp. 1794–1816

BIFURCATION ANALYSIS OF AN SIRS EPIDEMIC MODEL WITH GENERALIZED INCIDENCE*

M. E. ALEXANDER† AND S. M. MOGHADAS†

Abstract. An SIRS epidemic model, with a generalized nonlinear incidence as a function of the number of infected individuals, is developed and analyzed. Extending previous work, it is assumed that the natural immunity acquired by infection is not permanent but wanes with time. The nonlinearity of the functional form of the incidence of infection, which is subject only to a few general conditions, is biologically justified. The stability analysis of the associated equilibria is carried out, and the threshold quantity ($\mathscr{R}_0$) that governs the disease dynamics is derived. It is shown that $\mathscr{R}_0$, called the basic reproductive number, is independent of the functional form of the incidence. Local bifurcation theory is applied to explore the rich variety of dynamical behavior of the model. Normal forms are derived for the different types of bifurcation that the model undergoes, including Hopf, saddle-node, and Bogdanov–Takens. The first Lyapunov coefficient is computed to determine various types of Hopf bifurcation, such as forward or backward, subcritical or supercritical. The existence of a saddle-node bifurcation, at the turning point of backward bifurcation, is established by applying Sotomayor's theorem. The Bogdanov–Takens normal form is used to formulate the local bifurcation curve for a family of homoclinic orbits arising when a Hopf and a saddle-node bifurcation merge. These theoretical results are detailed and numerically illustrated for two different kinds of incidence, corresponding to unbounded and saturated contact rates. The coexistence of two limit cycles, due to the occurrence of a backward subcritical Hopf bifurcation, is also demonstrated. These results lead to the determination of ranges for the periodicity behavior of the model based on two critical parameters: the basic reproductive number and the rate of loss of natural immunity.

Key words. epidemic models, nonlinear incidence, Hopf bifurcation, saddle-node bifurcation, Bogdanov–Takens bifurcation

AMS subject classifications. Primary, 34C23, 92D25; Secondary, 34D23

DOI. 10.1137/040604947

1. Introduction.

In modeling of communicable diseases, there are several factors that substantially affect the dynamical behavior of the models. Recent studies have shown that the incidence rate is a major factor in producing the rich dynamics of epidemic models [1, 8, 13, 16, 23, 24, 25, 30, 31]. These studies have described interesting mathematical phenomena, such as bistability and periodicity, observed in data for some infectious diseases. In the bistability phenomenon (backward bifurcation), the model exhibits multiple endemic equilibria even when the classical requirement of the basic reproductive number being less than unity is satisfied. Disease eradication may then depend on other agents such as the initial conditions of the subpopulations. In the other scenario (periodicity), the model exhibits oscillatory behavior due to the existence of periodic solutions. This is an important dynamical feature of the model, as it shows periodically high level of incidence with a large number of infected individuals, which can be severely damaging to the population. Models with bistability or periodicity are numerous in the literature, and the reader may consult [1, 14, 15, 19, 21, 26, 30, 31].

In most classical models of epidemics, the incidence rate is taken to be mass

*Received by the editors March 9, 2004; accepted for publication (in revised form) January 5, 2005; published electronically July 26, 2005. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).
http://www.siam.org/journals/siap/65-5/60494.html
†Institute for Biodiagnostics, National Research Council Canada, Winnipeg, MB, Canada R3B 1Y6 (murray.alexander@nrc-cnrc.gc.ca, seyed.moghadas@nrc-cnrc.gc.ca).

action incidence with bilinear interactions i.e., βIS, where β is the probability of transmission per contact and S and I represent the numbers of susceptible and infected individuals, respectively. These models typically do not admit bistability or periodicity, as they have at most one endemic equilibrium; the disease will be eradicated if the basic reproductive number is less than one, and will persist otherwise [3, 23]. Although these simple models lead to general conclusions for long-term disease dynamics, they may not provide sufficient details of complexity in the population behavior to contribute to a better understanding of epidemiological patterns and disease control. For instance, the nonlinearity in the incidence could lead to complex dynamics of epidemic models. In a previous study [1], we have shown that simple models with nonlinear incidence rates can exhibit periodic oscillations or backward bifurcations without time-dependent coefficients or being cyclic.

There are reasons for using nonlinear incidence rates in the process of disease modeling. Yorke and London [33] showed that the incidence rate $\beta(1 - cI)IS$ with positive c and time-dependent β is consistent with the results of the simulations for measles outbreaks. Capasso and Serio [6] used a saturated incidence rate $\beta IS/(1 + \beta\delta I)$, $\delta > 0$, to prevent the unboundedness of the contact rate. The effect of behavioral changes has been incorporated by Liu and colleagues [24, 25] through the use of a nonlinear incidence rate $\kappa I^l S/(1 + \alpha I^h)$ with $\kappa,\ l,\ \alpha,\ h > 0$. Recent studies have also pointed out several reasons for the nonlinearity of the incidence rates, including the change in the contact rate with increasing likelihood of infection from multiple exposures, severity and stage of the infection, and recruitment of infected individuals [13, 16, 30, 31].

This paper extends some previous studies [1, 30, 31] on the dynamics of simple SIR (Susceptible-Infected-Recovered) epidemic models by incorporating a generalized incidence rate in an SIRS (Susceptible-Infected-Recovered-Susceptible) model which satisfies some realistic assumptions. The transitions between subpopulations are mathematically expressed by the following differential equations:

$$\frac{dS}{dt} = \Pi - \beta[1 + f(I;\nu)]\frac{IS^p}{N} - \mu S + \delta R, \tag{1.1}$$

$$\frac{dI}{dt} = \beta[1 + f(I;\nu)]\frac{IS^p}{N} - (\mu + \alpha)I, \tag{1.2}$$

$$\frac{dR}{dt} = \alpha I - (\mu + \delta)R, \tag{1.3}$$

where $p > 0$; $N \equiv S + I + R$ is the total population size; Π is the rate of recruitment of individuals (including newborns and immigrants) into the susceptible class; μ is the natural death rate; β is the probability of infection per contact per unit time; α is the recovery rate; δ is the rate at which recovered individuals lose their immunity (acquired by infection); and $f(I;\nu) \in C^3(\mathbb{R})$ for $I,\ \nu \geq 0$, is a nonlinear function which satisfies the following assumptions:

(A1) $f(0;\nu) = f(I;0) = 0$,
(A2) $\partial f/(\partial\nu) > 0$ for $\nu > 0$,
(A3) $\partial f/(\partial I) > 0$ for $I > 0$,
(A4) $\partial^2 f/(\partial I^2) \leq 0$ for $I > 0$.

The term S^p in the infection rate $\beta[1 + f(I;\nu)]IS^p/N$ is a specific form of the general case $\psi(S)$ satisfying $\psi(0) = 0$ and $\psi'(S) \geq 0$ for $S \geq 0$ (see [27]). We note from (A1) that, for ν and I small, the term $\beta I/N$ (proportional mixing) in the infection rate dominates, while for large enough ν and I, the nonlinear term $f(I;\nu)I/N$ dominates. When $\nu = 0$, the incidence rate reduces to the case of proportional mixing, in which

case the model exhibits at most one endemic equilibrium; the disease dies out if the basic reproductive number is less than unity and invades otherwise. Therefore, ν may be considered as a parameter measuring departure from a proportional mixing rate of incidence (see [1]). Under the assumptions (A1)–(A4), the nonlinear function $f(I;\nu)$ covers both unbounded and bounded (saturated) cases that are frequently used. The nonlinearity may be due to several factors, such as crowding of infected individuals, multiple pathways to infection, stage of infection and its severity, or protective measures taken by susceptible individuals [1, 30, 31]. Some of the specific forms of $f(I;\nu)$ appearing in the literature and satisfying (A1)–(A4) are $f(I;\nu) = \nu I^q$; $f(I;\nu) = \nu I^q/(1+\nu I^q)$ with $0 < q \leq 1$, $\nu > 0$ [1, 31]; $f(I,\nu) = \nu I^q(1+\kappa I)^p/(1+\nu I^q)$ with $p+q=1,\ p,\kappa \geq 0$ [1, 31]; and $f(I;\nu) = 1 - e^{-\nu I}$ [17].

The model (1.1)–(1.3) is studied in [1] for the special case $p = 1$ and $\delta = 0$. It has been shown that this model not only exhibits backward bifurcation but also may undergo a Hopf bifurcation leading to the appearance of two concentric limit cycles. Similar models with nonlinear incidence rates have also been studied by a number of authors. Ruan and Wang [30] analyzed an SIRS model with the incidence rate $\kappa I^l S/(1+\alpha I^h)$ (introduced by Liu and colleagues [24, 25]), where $l = h = 2$, and showed that it undergoes a Bogdanov–Takens bifurcation. The case where $\alpha = 0$ was discussed by Hethcote and van den Driessche [16], Liu, Hethcote, and Levin [24], and Liu, Levin, and Iwasa [25]. Derrick and van den Driessche [8] discussed bifurcation behavior of an SIRS model with the incidence rate $\beta I^2 S/N^2$ [28]. Their study uses a modification of the Bogdanov–Takens–Carr procedure to blow up a Bogdanov point, and Melnikov's method to determine a locus of approximate values along which a homoclinic orbit can be perturbed. The uniqueness of periodic solutions for this model has been shown by Alwash [2].

This paper focuses on the detailed dynamics analysis of the model (1.1)–(1.3). We show that it may exhibit two endemic equilibria, giving rise to the phenomenon of bistability, for some ranges of parameter values. The local stability of these equilibria is investigated, which enables us to classify the types of model equilibria (e.g., attractor, saddle, or repeller). The different kinds of bifurcation the model undergoes for a general incidence rate f satisfying (A1)–(A4) are discussed. The normal form of the model is derived and used to determine the conditions for the existence of various types of Hopf bifurcation (sub- or supercritical, forward or backward). The existence of multiple limit cycles is also discussed. A saddle-node bifurcation is analyzed using Sotomayor's theorem at the turning point of backward bifurcation, in which two endemic equilibria merge. Finally, the existence of a homoclinic orbit is proven by using the Bogdanov–Takens normal form of the model which gives, among others, the local representation of a homoclinic bifurcation curve. The results of our analysis are applied in detail to two different examples of the function f: unbounded and bounded. Numerical simulations for these examples are presented to illustrate the theoretical results.

The paper is organized as follows. The existence of the model equilibria is discussed in section 2. The stability analysis of the equilibria is carried out in section 3 by reducing the model to a two dimensional system. Bifurcation behavior of the reduced model is analyzed in section 4. The results are detailed and numerically illustrated for two examples in section 5. The paper ends with a brief discussion of the results in section 6.

2. Model equilibria. In the absence of the disease ($I = 0$), the model has a unique disease-free equilibrium (DFE), given by $\mathscr{E}_0 = (\Pi/\mu, 0, 0)$. In order to find the

endemic equilibria in the presence of the disease ($I \neq 0$), we note that the equation for the total population is given by $dN/dt = \Pi - \mu N$. Since $N \to \Pi/\mu$ as $t \to \infty$, it follows that at any equilibrium $\mathscr{E}^* = (S^*, I^*, R^*)$, $N^* = S^* + I^* + R^* = \Pi/\mu$, and

$$\Omega = \left\{ (S, I, R) : \ S, I, R \geq 0, \ S + I + R = \frac{\Pi}{\mu} \right\}$$

is a positively invariant region for the model. Henceforth, we restrict our attention to the dynamics of the model in Ω.

Solving equations (1.2) and (1.3) for S and R in terms of I gives, at equilibrium,

$$S^p = \frac{\mu + \alpha}{\beta[1 + f(I;\nu)]} \frac{\Pi}{\mu}, \tag{2.1}$$

$$R = \frac{\alpha I}{\mu + \delta}. \tag{2.2}$$

Substituting (2.1) and (2.2) into (1.1) leads to the following equation for I at equilibrium:

$$\Pi - \frac{\mu(\mu + \alpha + \delta)}{\mu + \delta} I - \mu \left(\frac{\mu + \alpha}{\beta[1 + f(I;\nu)]} \frac{\Pi}{\mu} \right)^{\frac{1}{p}} = 0. \tag{2.3}$$

Defining

$$\mathscr{R}_0 = \frac{\beta}{\mu + \alpha} \left(\frac{\Pi}{\mu} \right)^{p-1}, \tag{2.4}$$

one can see that the roots of (2.3) are the fixed points of the equation:

$$\phi(I;\nu) \equiv \kappa \left(1 - \frac{1}{\mathscr{R}_0^{\frac{1}{p}} [1 + f(I;\nu)]^{\frac{1}{p}}} \right) = I, \tag{2.5}$$

where $\kappa = (\mu + \delta)\Pi/[\mu(\mu + \alpha + \delta)]$. In order to determine the number of endemic equilibria, we consider some properties of the function ϕ listed below:

(i) $\phi_0 \equiv \phi(0;\nu) = \kappa \left(1 - \dfrac{1}{\mathscr{R}_0^{\frac{1}{p}}} \right)$.

(ii) $\dfrac{\partial \phi}{\partial I} = \dfrac{\kappa f_I(I;\nu)}{p \mathscr{R}_0^{\frac{1}{p}} [1 + f(I;\nu)]^{1+\frac{1}{p}}} > 0.$

(iii) $\dfrac{\partial^2 \phi}{\partial I^2} = \kappa \dfrac{f_{II}(I;\nu)[1 + f(I;\nu)] - (1 + 1/p) f_I^2(I;\nu)}{p \mathscr{R}_0^{\frac{1}{p}} [1 + f(I;\nu)]^{2+\frac{1}{p}}} < 0.$

If $\mathscr{R}_0 > 1$, then $\phi_0 > 0$. It follows from (ii) that ϕ is an increasing function and hence $\lim_{I \to \infty} \phi(I;\nu) \equiv \phi_\infty$ is positive and finite. Since ϕ is concave down (by (iii)), there exists a unique $I^* > 0$ such that $\phi(I^*;\nu) = I^*$. If $\mathscr{R}_0 = 1$, then $\phi_0 = 0$ and $\partial\phi/(\partial I)(0) > 1 (\leq 1)$ if $f_I(0;\nu) > p/\kappa (\leq p/\kappa)$. Thus, $\phi(I;\nu) = I$ has a unique positive root if $f_I(0;\nu) > p/\kappa$, and no positive root otherwise. Finally, suppose $\mathscr{R}_0 < 1$, so that $\phi_0 < 0$. In this case, if $\lim_{I \to \infty} f(I;\nu) \equiv f_\infty(\nu) \leq 1/\mathscr{R}_0 - 1$, then $\phi_\infty < 0$, and

thus no positive roots of $\phi(I;\nu) = I$ exist. If $f_\infty(\nu) > 1/\mathscr{R}_0 - 1$, then depending on the functional form of ϕ, the equation $\phi(I;\nu) = I$ may have no, one, or two positive roots (see [1, Figure 1]). Therefore, we have the following theorem.

THEOREM 2.1. (a) *If $\mathscr{R}_0 > 1$, then the model* (1.1)–(1.3) *has a unique endemic equilibrium.*

(b) *If $\mathscr{R}_0 = 1$, then the model* (1.1)–(1.3) *has a unique endemic equilibrium if $f_I(0;\nu) > p/\kappa$ and no endemic equilibrium if $f_I(0;\nu) \le p/\kappa$.*

(c) *If $\mathscr{R}_0 < 1$, then*

(i) *the model* (1.1)–(1.3) *has no endemic equilibria whenever $f_\infty(\nu) \le 1/\mathscr{R}_0 - 1$;*

(ii) *the model* (1.1)–(1.3) *may have no, one, or two endemic equilibria if $f_\infty(\nu) > 1/\mathscr{R}_0 - 1$.*

We now determine the conditions under which the model exhibits two endemic equilibria when $\mathscr{R}_0 < 1$. Let $G(I;\nu) = \phi(I;\nu) - I$, and suppose that $f_I(0;\nu) > p/\kappa$, which implies $(\partial G/\partial I)|_{I=0} > 0$. Then $\lim_{I\to\infty} \partial G/\partial I = -1$ and $\partial^2 G/\partial I^2 = \partial^2\phi/\partial I^2 < 0$ for $I > 0$. Hence $\partial G/\partial I$ is a monotone decreasing function of I, and thus there exists a unique $I_0^* > 0$ such that $(\partial G/\partial I)|_{I=I_0^*} = 0$. This implies that G is an increasing function on $(0, I_0^*]$ and decreasing on (I_0^*, ∞) with maximum value at I_0^*. Since $G(0) < 0$ when $\mathscr{R}_0 < 1$, and $G(0) = 0$ when $\mathscr{R}_0 = 1$, by continuity, it follows that there exists $\mathscr{R}^*$ with $\mathscr{R}^* < 1$ for which $\partial G/\partial I$ has a unique root I_0^* with $(\partial G/\partial I)|_{I=I_0^*} = G(I_0^*) = 0$. Then for some values of $\mathscr{R}_0 \in (\mathscr{R}^*, 1)$, G has two roots I_1^* and I_2^* with $I_1^* < I_2^*$ (corresponding to two endemic equilibria of the model). Assuming $(\partial I/\partial \mathscr{R}_0)|_{I=I_1^*} > 0$ and taking into account that $(\partial\phi/\partial I)|_{I=I_1^*} > 1$, we have

$$\frac{\partial I}{\partial \mathscr{R}_0}\Big|_{I=I_1^*} > \frac{\partial I}{\partial \mathscr{R}_0}\Big|_{I=I_1^*} + \frac{\kappa}{p\mathscr{R}_0^{1+\frac{1}{p}}[1+f(I;\nu)]^{\frac{1}{p}}} > \frac{\partial I}{\partial \mathscr{R}_0}\Big|_{I=I_1^*},$$

which is a contradiction. This implies that $(\partial I/\partial \mathscr{R}_0)|_{I=I_1^*} < 0$, and thus the number of infected individuals at the low endemic equilibrium reduces when $\mathscr{R}_0$ increases. Similarly, it can be shown that $(\partial I/\partial \mathscr{R}_0)|_{I=I_2^*} > 0$, and hence the number of infected individuals at the high endemic equilibrium increases when $\mathscr{R}_0$ increases. Therefore, the quantity $\mathscr{R}^*$ for which $(\partial G/\partial I)|_{I=I_0^*} = G(I_0^*) = 0$ is unique and we have the following theorem.

THEOREM 2.2. *If $f_I(0;\nu) > p/\kappa$, then there exists a unique $\mathscr{R}^*$ with $\mathscr{R}^* < 1$ such that the model* (1.1)–(1.3) *has no endemic equilibrium if $\mathscr{R}_0 < \mathscr{R}^*$, a unique endemic equilibrium if $\mathscr{R}_0 = \mathscr{R}^*$, and two endemic equilibria if $\mathscr{R}^* < \mathscr{R}_0 < 1$.*

Remark 2.1. The case $f_I(0;\nu) > p/\kappa$ in Theorem 2.2 corresponds to backward (transcritical) bifurcation at E_0, which leads to the existence of multiple endemic equilibria for $\mathscr{R}_0 \in (\mathscr{R}^*, 1)$ (see Figure 4.1). There is also a forward transcritical bifurcation at E_0 (when $\mathscr{R}_0 = 1$) if $f_I(0;\nu) < p/\kappa$.

3. Reduced model and stability analysis. Since Ω is a positively invariant region for the model (1.1)–(1.3), assuming that the size of the population has reached its limiting value, i.e., $N \equiv \Pi/\mu = S + I + R$, and using $R = \Pi/\mu - S - I$ in (1.1), we can eliminate R from the equations. This gives the following reduced model:

$$\frac{dS}{dt} = \Pi - \left(\frac{\mu}{\Pi}\right)\beta[1+f(I;\nu)]IS^p - \mu S + \delta\left(\frac{\Pi}{\mu} - S - I\right), \tag{3.1}$$

$$\frac{dI}{dt} = \left(\frac{\mu}{\Pi}\right)\beta[1+f(I;\nu)]IS^p - (\mu+\alpha)I. \tag{3.2}$$

The equilibria of the reduced model correspond to those of the model (1.1)–(1.3). The DFE of (3.1)–(3.2), given by $E_0 = (\Pi/\mu, 0)$, has the corresponding Jacobian

$$J_0 = \begin{pmatrix} -(\mu+\delta) & -\beta\left(\frac{\Pi}{\mu}\right)^{p-1} - \delta \\ 0 & (\mu+\alpha)(\mathscr{R}_0 - 1) \end{pmatrix},$$

with the eigenvalues $\lambda_1 = -(\mu+\delta)$ and $\lambda_2 = (\mu+\alpha)(\mathscr{R}_0 - 1)$. Therefore, E_0 is locally asymptotically stable (LAS) if $\mathscr{R}_0 < 1$, and unstable if $\mathscr{R}_0 > 1$. The threshold quantity $\mathscr{R}_0$ is called the *basic reproductive number*. Biologically, this quantity is defined to be the average number of new infectious cases produced by one infected case introduced into a wholly susceptible population [3]. Mathematically, it determines the condition under which the DFE is LAS. In the following, we discuss how the critical threshold $\mathscr{R}_0$ governs disease dynamics.

Suppose $\mathscr{R}_0 < 1$. If the model has no endemic equilibrium, then (from the Poincaré–Bendixson theorem) no periodic orbits exist in Ω. Since Ω is a bounded positively invariant region and E_0 is the only equilibrium in Ω, the local stability of E_0 implies that every solution initiating in Ω approaches E_0.

THEOREM 3.1. *Suppose $\mathscr{R}_0 < 1$. If the model has no endemic equilibrium, then the DFE is globally asymptotically stable (GAS).*

This theorem implies that E_0 is GAS whenever $\mathscr{R}_0 < \mathscr{R}^*$. Note that if $f(I;\nu)$ is bounded and $\mathscr{R}_0 < 1/[1 + f_\infty(\nu)]$, then $\phi_\infty < 0$, and hence the model has no endemic equilibrium. Thus, the DFE is GAS.

Now suppose that $f_I(0;\nu) > p/\kappa$ and $\mathscr{R}^* < \mathscr{R}_0 < 1$ where the model exhibits two endemic equilibria. Then, we have the following result.

THEOREM 3.2. *If $\mathscr{R}^* < \mathscr{R}_0 < 1$, then one of the endemic equilibria is a saddle and the other is either an attractor or a repeller.*

Proof. Let $E_1^* = (S_1^*, I_1^*)$ and $E_2^* = (S_2^*, I_2^*)$ denote the equilibria of the reduced model with, respectively, low and high number of infected individuals when $\mathscr{R}^* < \mathscr{R}_0 < 1$. We shall show that E_1^* is always a saddle point. The corresponding Jacobian of the reduced model at a typical endemic equilibrium $E^* = (S^*, I^*)$ is given by

$$J^*(\nu) = \begin{pmatrix} -a - (\mu+\delta) & -b - \delta \\ a & b - (\mu+\alpha) \end{pmatrix},$$

where

$$a = p\left(\frac{\mu}{\Pi}\right)\beta[1 + f(I^*;\nu)]I^* S^{*p-1}, \tag{3.3}$$

$$b = \left(\frac{\mu}{\Pi}\right)\beta[1 + f(I^*;\nu) + I^* f_I(I^*;\nu)]S^{*p}. \tag{3.4}$$

The eigenvalues of $J^*(\nu)$ are the roots of the characteristic polynomial $P(\lambda) = \lambda^2 + B\lambda + C$, with

$$B = a - b + 2\mu + \alpha + \delta, \tag{3.5}$$

$$C = a(\mu + \alpha + \delta) - b(\mu + \delta) + (\mu + \alpha)(\mu + \delta). \tag{3.6}$$

Using (2.1) for S^*, the expression for C can be reduced to

$$C = a(\mu + \alpha + \delta)\left(1 - \frac{\partial\phi}{\partial I}\Big|_{I=I^*}\right). \tag{3.7}$$

Since $\partial\phi/(\partial I) > 1$ at E_1^*, it follows that $C|_{I=I_1^*} < 0$, and hence $P(\lambda)$ has two real roots with opposite signs. This implies that E_1^* is a saddle point. Noting that $\partial\phi/(\partial I) < 1$ at E_2^* (for any $\mathscr{R}_0 > \mathscr{R}^*$), it follows from (3.7) that $C|_{I=I_2^*} > 0$. This implies that E_2^* cannot be a saddle point. Therefore, the endemic equilibrium with high number of infected individuals is either an attractor or a repeller. ∎

4. Bifurcation analysis. In this section, different kinds of bifurcation will be discussed. We will undertake the stability analysis of the equilibria to obtain normal forms of the model for Hopf, saddle-node, and Bogdanov–Takens bifurcations.

4.1. Hopf bifurcation. Since $C|_{I=I_2^*} > 0$ at $E_2^* = (S_2^*, I_2^*)$, any exchange of stability of E_2^* corresponds to a Hopf bifurcation, which occurs when $B|_{I=I_2^*} = 0$ in (3.5). Suppose there exists ν^c such that $B(\nu^c) \equiv B|_{I=I_2^*(\nu^c)} = 0$ and $dB/(d\nu)|_{\nu^c} \neq 0$. Then, the model undergoes a Hopf bifurcation in which the roots of $P(\lambda)$ cross the imaginary axis as ν passes through ν^c such that $d(Re[\lambda(\nu)])/(d\nu) = -(1/2)(dB/d\nu) \neq 0$ at ν^c. The kind of Hopf bifurcation will be determined by the sign of the first Lyapunov coefficient of the normal form of the model. In the following, we will obtain this normal form for a neighborhood of E_2^*.

Consider the transformations $S = S^* + x$ and $I = I^* + y$ about a typical endemic equilibrium $E^* = (S^*, I^*)$ (where $C|_{I=I^*} > 0$), as the origin of coordinates (x, y). Using the expressions (2.1) and (2.5), the model (3.1)–(3.2) transforms to

$$\begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} = J^*(\nu) \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} -M(x,y,\nu) \\ M(x,y,\nu) \end{pmatrix}, \tag{4.1}$$

where

$$M(x,y;\nu) = \left(\frac{\mu}{\Pi}\right)\beta\Big\{[1+f(I^*;\nu)]I^* + p[1+f(I^*;\nu)+I^* f_I(I^*;\nu)]xyS^{*p-1} + [(I^*+y)f_2(y,I^*;\nu) + y^2 f_I(I^*;\nu)](S^{*p} + pxS^{*p-1})\Big\}\sum_{k=2}^{\infty}\begin{pmatrix} p \\ k \end{pmatrix} x^k S^{*p-k}$$

and $f_2(y, I^*; \nu)$ denotes second and higher order terms in y of the expression

$$f(I^*+y;\nu) = f(I^*;\nu) + yf_I(I^*;\nu) + f_2(y,I^*;\nu).$$

Note that $\text{trace}(J^*(\nu^c)) = -B(\nu^c) = 0$. Let $\mathbf{u}$ be an eigenvector of $J^*(\nu^c)$ corresponding to the eigenvalue $i\omega_c$, i.e., $J^*(\nu^c)\mathbf{u} = i\omega_c\mathbf{u}$, where $\mathbf{u} = (u^1, u^2)^T \in \mathbb{C}^2$ and $\omega_c = \sqrt{C(I^*(\nu^c))}$. A simple calculation gives $\mathbf{u} = (b+\delta, -b+\mu+\alpha-i\omega_c)^T$. Letting $\mathbf{u} = Re(\mathbf{u}) + i\text{Im}(\mathbf{u})$ and defining $Q = [\text{Re}(\mathbf{u}), \text{Im}(\mathbf{u})]$, it follows that

$$Q^{-1}J^*(\nu^c)Q = \begin{pmatrix} 0 & \omega_c \\ -\omega_c & 0 \end{pmatrix}.$$

Defining $(\xi, \eta)^T = Q^{-1}(x, y)^T$, the normal form of the system (4.1) is obtained as

$$\begin{pmatrix} \dot{\xi} \\ \dot{\eta} \end{pmatrix} = \begin{pmatrix} 0 & \omega_c \\ -\omega_c & 0 \end{pmatrix}\begin{pmatrix} \xi \\ \eta \end{pmatrix} - \frac{1}{b+\delta}\begin{pmatrix} 1 \\ \frac{\mu+\alpha+\delta}{\omega_c} \end{pmatrix}\tilde{M}(\xi,\eta;\nu), \tag{4.2}$$

where $\tilde{M}(\xi, \eta; \nu) = M((b+\delta)\xi, (-b+\mu+\alpha)\xi - \omega_c\eta; \nu)$. Since

$$\frac{d}{d\nu}\text{trace}(J^*(\nu))\Big|_{\nu=\nu^c} = -\frac{dB}{d\nu}\Big|_{\nu=\nu^c} \neq 0, \tag{4.3}$$

it follows that E^* is LAS for $\nu < \nu^c$ (respectively, $\nu > \nu^c$) and unstable for $\nu > \nu^c$ (respectively, $\nu < \nu^c$) if $dB/(d\nu)\big|_{\nu=\nu^c} < 0$ (respectively, $dB/(d\nu)\big|_{\nu=\nu^c} > 0$), and the model undergoes a Hopf bifurcation at $\nu = \nu_c$ [11, Theorem 8.6]. Evaluating the first Lyapunov coefficient (σ) [11, 22] of system (4.2) at $(0, 0, \nu^c)$ gives

$$16(b+\delta)^2\sigma = -(b+\delta)\left[\tilde{M}_{\xi\xi\xi} + \tilde{M}_{\xi\eta\eta} + \frac{\mu+\alpha+\delta}{\omega_c}(\tilde{M}_{\xi\xi\eta} + \tilde{M}_{\eta\eta\eta})\right]$$

$$- \frac{1}{\omega_c}\left\{\left[1 - \left(\frac{\mu+\alpha+\delta}{\omega_c}\right)^2\right]\tilde{M}_{\xi\eta}(\tilde{M}_{\xi\xi} + \tilde{M}_{\eta\eta}) - \frac{\mu+\alpha+\delta}{\omega_c}(\tilde{M}_{\xi\xi}^2 - \tilde{M}_{\eta\eta}^2)\right\}, \tag{4.4}$$

where

$$\tilde{M}_{\xi\xi\xi}(0,0,\nu^c) = \{Q_{11}^3 M_{xxx} + 3Q_{11}^2 Q_{21} M_{xxy} + 3Q_{11}Q_{21}^2 M_{xyy} + Q_{21}^3 M_{yyy}\}\big|_{(0,0,\nu^c)},$$

$$\tilde{M}_{\eta\eta\eta}(0,0;\nu^c) = Q_{22}^3 M_{yyy}\big|_{(0,0,\nu^c)}, \quad \tilde{M}_{\xi\eta}(0,0;\nu^c) = Q_{22}\{Q_{11}M_{xy} + Q_{21}M_{yy}\}\big|_{(0,0,\nu^c)},$$

$$\tilde{M}_{\xi\eta\eta}(0,0;\nu^c) = Q_{22}^2\{Q_{11}M_{xyy} + Q_{21}M_{yyy}\}\big|_{(0,0,\nu^c)}, \quad \tilde{M}_{\eta\eta}(0,0;\nu^c) = Q_{22}^2 M_{yy}\big|_{(0,0,\nu^c)},$$

$$\tilde{M}_{\xi\xi\eta}(0,0;\nu^c) = Q_{22}\{Q_{11}^2 M_{xxy} + 2Q_{11}Q_{21}M_{xyy} + Q_{21}^2 M_{yyy}\}\big|_{(0,0,\nu^c)},$$

$$\tilde{M}_{\xi\xi}(0,0;\nu^c) = \{Q_{11}^2 M_{xx} + 2Q_{11}Q_{21}M_{xy} + Q_{21}^2 M_{yy}\}\big|_{(0,0,\nu^c)},$$

with $Q_{11} = b + \delta$, $Q_{21} = -b + \mu + \alpha$, and $Q_{22} = -\omega_c$. Using these expressions in (4.4), we have the following theorem.

THEOREM 4.1. *If $\sigma \neq 0$, then a curve of periodic solutions bifurcates from the endemic equilibrium E^* such that*

(a) *for $\sigma < 0$, the model undergoes a supercritical Hopf bifurcation if $dB/(d\nu)|_{\nu=\nu^c} < 0$ and a backward supercritical Hopf bifurcation if $dB/(d\nu)|_{\nu=\nu^c} > 0$;*

(b) *for $\sigma > 0$, the model undergoes a subcritical Hopf bifurcation if $dB/(d\nu)|_{\nu=\nu^c} < 0$ and a backward subcritical Hopf bifurcation if $dB/(d\nu)|_{\nu=\nu^c} > 0$.*

A supercritical (backward supercritical) Hopf bifurcation in case (a) of Theorem 4.1 leads to the appearance (disappearance) of a stable limit cycle when ν passes through ν^c. In case (b), the model exhibits an unstable limit cycle around E^* when ν passes through ν^c. Thus, every solution starting inside the limit cycle approaches E^* (see [1, Figure 2]).

Remark 4.1. Note that E_0 is unstable whenever $\mathscr{R}_0 > 1$, and the unique endemic equilibrium is either stable or unstable. Since Ω is a bounded positively invariant set, if the model undergoes a subcritical Hopf bifurcation, then (from the Poincaré–Bendixson theorem) every solution initiating outside the limit cycle must approach a stable limit cycle. This shows that when $\mathscr{R}_0 > 1$, two concentric limit cycles can coexist: the inner one unstable and the outer one stable (see [1]). We shall discuss the existence of multiple limit cycles in section 5.2.

4.2. Saddle-node bifurcation. In order to show that the model may undergo a saddle-node bifurcation, we will take advantage of Sotomayor's theorem [12, Theorem 3.4.1]. Considering $\mathscr{R}_0$ as the bifurcation parameter, it follows from Theorem 2.2 that at $\mathscr{R}_0 = \mathscr{R}^*$, the model has a unique endemic equilibrium $E_0^* = (S_0^*, I_0^*)$ for which the Jacobian $J_{\mathscr{R}^*}^*(E_0^*)$ has a simple eigenvalue 0 and an eigenvalue $\lambda = -B_{\mathscr{R}^*}(E_0^*)$. Note that $\partial\phi/(\partial I)\big|_{I=I_0^*} = 1$, and hence $C_{\mathscr{R}^*}(E_0^*) = 0$ (see (3.7)). Let $\mathbf{V} = (v_1, v_2)^T$ and $\mathbf{W} = (w_1, w_2)$ be right and left eigenvectors of $J_{\mathscr{R}^*}^*(E_0^*)$, respectively, corresponding to the zero eigenvalue. Then, a simple calculation yields

$$\mathbf{V} = \big[b+\delta, -a-(\mu+\delta)\big]^T, \qquad \mathbf{W} = \big[a, a+\mu+\delta\big]. \tag{4.5}$$

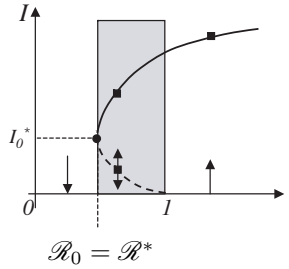


FIG. 4.1. *A backward bifurcation at $\mathscr{R}_0 = 1$ with a dashed curve for the location of the saddle point and solid curve for the location of the other endemic equilibrium. The saddle-node bifurcation occurs at $E_0^* = (S_0^*, I_0^*)$ when $\mathscr{R}_0$ passes through $\mathscr{R}^*$.*

Let $\mathscr{G} = (g_1, g_2)$, where g_1 and g_2 are, respectively, the right-hand sides of equations in (3.1)–(3.2). By considering β as a function of $\mathscr{R}_0$, we have

$$\mathscr{L}_1 \equiv \mathbf{W} \cdot \frac{\partial \mathscr{G}}{\partial \mathscr{R}_0}\Big|_{(E_0^*, \mathscr{R}^*)} = \frac{(\mu+\alpha)(\mu+\delta)I_0^*}{\mathscr{R}^*} > 0. \tag{4.6}$$

Using (2.1), it can be seen that

$$\begin{aligned}\mathscr{L}_2 \equiv \mathbf{W} \cdot [D^2_{(S,I)}\mathscr{G}(\mathbf{V},\mathbf{V})]\Big|_{(E_0^*, \mathscr{R}^*)} &= (w_1 - w_2)\left[\frac{\partial^2 g_1}{\partial S^2}v_1^2 + \frac{\partial^2 g_1}{\partial I^2}v_2^2 + 2\frac{\partial^2 g_1}{\partial S \partial I}v_1 v_2\right] \\ = -(\mu+\delta)\frac{\mu\beta}{\Pi}\Big\{ &- p(p-1)[1+f(I_0^*;\nu)]I_0^* S_0^{*p-2}(b+\delta)^2 \\ &- \big[2f_I(I_0^*;\nu) + I_0^* f_{II}(I_0^*;\nu)\big](a+\mu+\delta)^2 S_0^{*p} \\ &+ 2p\big[1+f(I_0^*;\nu) + I_0^* f_I(I_0^*;\nu)\big](b+\delta)(a+\mu+\delta)S_0^{*p-1}\Big\}.\end{aligned}$$

Then, since $C_{\mathscr{R}^*}(E_0^*) = 0$, after some manipulations, it follows from (A4) that

$$\begin{aligned}S_0^*\mathscr{L}_2 &= -(\mu+\delta)\left\{a(1+p)(b+\delta)^2 - \left(\frac{\mu\beta}{\Pi}\right)I_0^* S_0^{*p+1} f_{II}(I_0^*;\nu)(a+\mu+\delta)^2\right\} \\ &< -a(1+p)(\mu+\delta)(b+\delta)^2 < 0.\end{aligned} \tag{4.7}$$

Thus, from Sotomayor's theorem [12], there is a smooth curve of equilibria in $\mathbb{R}^2$ passing through E_0^*, tangent to the line $\mathscr{R}_0 = \mathscr{R}^*$ (see Figure 4.1). Since $\mathscr{L}_1 > 0$ and $\mathscr{L}_2 < 0$, there are no equilibria near E_0^* when $\mathscr{R}_0 < \mathscr{R}^*$ and two equilibria when $\mathscr{R}_0 > \mathscr{R}^*$. In fact, from (4.6) and (4.7), the local phase portraits of the model (3.1)–(3.2) are topologically equivalent to those of $\dot{v} = (\mathscr{R}_0 - \mathscr{R}^*) - (v - I_0^*)^2$. Therefore, we have the following theorem.

THEOREM 4.2. *If $f_I(0;\nu) > p/\kappa$, then the model undergoes a saddle-node bifurcation at E_0^* when $\mathscr{R}_0$ passes through the critical value $\mathscr{R}^*$.*

4.3. Homoclinic bifurcation. In this section, the Bogdanov–Takens bifurcations of the model (3.1)–(3.2) are discussed, from which the local representation of a homoclinic bifurcation curve is derived. Let $\mathscr{R}_{BT}$ and ν_{BT} denote the bifurcation parameters at which the model simultaneously undergoes a saddle-node bifurcation and a Hopf bifurcation. Here, we first consider the conditions for determining $\mathscr{R}_{BT}$

and ν_{BT}. Using (3.3) and (3.4), it can be seen that

$$\text{trace } J^*(\nu) = p(\mu+\alpha)(\mu+\delta)(u-1)\left(1-\frac{\partial\phi}{\partial I}\right), \tag{4.8}$$

$$\det J^*(\nu) = p(\mu+\alpha)(u-1)\frac{\partial\phi}{\partial I} - (\mu+\delta)\left(1+\frac{p(\mu+\alpha)(u-1)}{\mu+\alpha+\delta}\right), \tag{4.9}$$

where $u = \mathscr{R}_0^{1/p}[1+f(I;\nu)]^{1/p}$. A simple calculation yields that if trace $J^*(\nu) = \det J^*(\nu) = 0$, then

$$u = 1 + \frac{(\mu+\delta)(\mu+\alpha+\delta)}{p\alpha(\mu+\alpha)}. \tag{4.10}$$

Thus, it follows from (2.5) that $I_0^* = \kappa(1-1/u)$. Noting that $\partial\phi/\partial I = 1$ at the Bogdanov point (S_0^*, I_0^*), it can be seen that

$$I_0^* f_I(I_0^*;\nu_{BT}) = \frac{p(u-1)u^p}{\mathscr{R}_{BT}}. \tag{4.11}$$

Therefore, for given functional form of $f(I;\nu)$, equations (4.10) and (4.11) provide conditions for determining $\mathscr{R}_{BT}$ and ν_{BT}. These general equations will be applied to two specific forms of $f(I;\nu)$ in section 5, where closed form expressions for the Bogdanov point(s) are given.

We continue the analysis of the homoclinic bifurcation by deriving the normal form at the Bogdanov point. At this point, $B = C = 0$, and it follows from (3.5) and (3.6) that

$$\alpha a = (\mu+\delta)^2. \tag{4.12}$$

Using the transformations $S = S_0^* + \xi$ and $I = I_0^* + \eta$ (at $\mathscr{R}_0 = \mathscr{R}_{BT}$, $\nu = \nu_{BT}$), the model (3.1)–(3.2) becomes

$$\begin{pmatrix} \dot\xi \\ \dot\eta \end{pmatrix} = J(E_0^*)\begin{pmatrix} \xi \\ \eta \end{pmatrix} + \begin{pmatrix} -M(\xi,\eta;\nu_{BT}) \\ M(\xi,\eta;\nu_{BT}) \end{pmatrix}, \tag{4.13}$$

where, by (4.12),

$$J(E_0^*) = \frac{1}{\alpha}\begin{pmatrix} -(\mu+\delta)(\mu+\alpha+\delta) & -(\mu+\alpha+\delta)^2 \\ (\mu+\delta)^2 & (\mu+\delta)(\mu+\alpha+\delta) \end{pmatrix}.$$

Since $J(E_0^*) \neq 0$, there exist real linearly independent vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ such that $J(E_0^*)\mathbf{x}_1 = 0$ and $J(E_0^*)\mathbf{x}_2 = \mathbf{x}_1$. These vectors are given by

$$\mathbf{x}_1 = [-(\mu+\alpha+\delta), \mu+\delta]^T, \qquad \mathbf{x}_2 = [-1,1]^T.$$

Similarly, there exist vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ such that $[J(E_0^*)]^T\mathbf{y}_1 = 0$ and $[J(E_0^*)]^T\mathbf{y}_2 = \mathbf{y}_1$, where $[J(E_0^*)]^T$ is the transposed matrix. These vectors may be expressed as

$$\mathbf{y}_1 = \left(\frac{1}{\alpha}\right)[\mu+\delta, \mu+\alpha+\delta]^T, \qquad \mathbf{y}_2 = \left(\frac{1}{\alpha}\right)[-1,-1]^T.$$

It is easy to verify that $\mathbf{x}_1\cdot\mathbf{y}_2 = \mathbf{x}_2\cdot\mathbf{y}_1 = 1$ and $\mathbf{x}_2\cdot\mathbf{y}_2 = \mathbf{x}_1\cdot\mathbf{y}_1 = 0$. By defining $\tilde{Q} = [\mathbf{x}_1, \mathbf{x}_2]$, any $(\xi,\eta)^T$ can be uniquely represented by $(\xi,\eta)^T = \tilde{Q}(\theta_1,\theta_2)^T$, for some real $\theta_1, \theta_2 \in \mathbb{R}$, from which the new coordinates (θ_1,θ_2) are obtained as

$$\theta_1 = -\frac{\xi+\eta}{\alpha}, \qquad \theta_2 = \frac{\mu+\delta}{\alpha}\xi + \frac{\mu+\alpha+\delta}{\alpha}\eta.$$

Considering (4.1) for all $\mathscr{R}_0$ and ν with $|\mathscr{R}_0 - \mathscr{R}_{BT}|$ and $|\nu - \nu_{BT}|$ small (recall that $B = C = 0$ at $\mathscr{R}_0 = \mathscr{R}_{BT}$, $\nu = \nu_{BT}$), and expanding its right-hand side as a Taylor series in (θ_1, θ_2) at $(0, 0)$ gives

$$\begin{aligned}
\frac{d\theta_1}{dt} &= \theta_2, \qquad (4.14)\\
\frac{d\theta_2}{dt} &= -C\theta_1 - B\theta_2\\
&\quad + \big[r^2 d_{11}(I^*, S^*) - rs d_{12}(I^*, S^*) + s^2 d_{22}(I^*, S^*)\big]\theta_1^2\\
&\quad + \big[d_{11}(I^*, S^*) - d_{12}(I^*, S^*) + d_{22}(I^*, S^*)\big]\theta_2^2\\
&\quad + \big[2r d_{11}(I^*, S^*) - (r+s) d_{12}(I^*, S^*) + 2s d_{22}(I^*, S^*)\big]\theta_1\theta_2\\
&\quad + O(\|(\theta_1, \theta_2)\|^3), \qquad (4.15)
\end{aligned}$$

where $r = \mu + \alpha + \delta$, $s = \mu + \delta$, and

$$\begin{aligned}
d_{11}(I^*, S^*) &= \frac{1}{2}\frac{\mu\beta}{\Pi} p(p-1)[1 + f(I^*; \nu)] I^* S^{*p-2},\\
d_{12}(I^*, S^*) &= \frac{\mu\beta}{\Pi} p[1 + f(I^*; \nu) + I^* f_I(I^*; \nu)] S^{*p-1},\\
d_{22}(I^*, S^*) &= \frac{1}{2}\frac{\mu\beta}{\Pi}[2 f_I(I^*; \nu) + I^* f_{II}(I^*; \nu)] S^{*p}.
\end{aligned}$$

For the sake of convenience, we define

$$\begin{aligned}
\mathscr{K}_{11} &= r^2 d_{11}(I^*, S^*) - rs d_{12}(I^*, S^*) + s^2 d_{22}(I^*, S^*),\\
\mathscr{K}_{12} &= 2r d_{11}(I^*, S^*) - (r+s) d_{12}(I^*, S^*) + 2s d_{22}(I^*, S^*),\\
\mathscr{K}_{22} &= d_{11}(I^*, S^*) - d_{12}(I^*, S^*) + d_{22}(I^*, S^*).
\end{aligned}$$

Assume that $\mathscr{K}_{12} \neq 0$ at the Bogdanov point. Then, there is a neighborhood of $(S_0^*, I_0^*; \mathscr{R}_{BT}, \nu_{BT})$ in which $\mathscr{K}_{12} \neq 0$. By setting $\Theta_1 = \theta_1 - \chi$, where $\chi = B/\mathscr{K}_{12}$, denoting Θ_1 as θ_1, and using a time reparametrization $dt = (1 - \mathscr{K}_{22}\theta_1)d\tau$, it is easy to check that (4.14) and (4.15) become

$$\begin{aligned}
\frac{d\theta_1}{d\tau} &= (1 - \mathscr{K}_{22}\theta_1)\theta_2, \qquad (4.16)\\
\frac{d\theta_2}{d\tau} &= (1 - \mathscr{K}_{22}\theta_1)\Big\{ -\chi(C - \mathscr{K}_{11}\chi) - (C - 2\mathscr{K}_{11}\chi)\theta_1\\
&\qquad + \mathscr{K}_{11}\theta_1^2 + \mathscr{K}_{22}\theta_2^2 + \mathscr{K}_{12}\theta_1\theta_2 + O(\|(\theta_1, \theta_2)\|^3)\Big\}. \qquad (4.17)
\end{aligned}$$

Introduce new variables $\Theta_1 = \theta_1$ and $\Theta_2 = (1 - \mathscr{K}_{22}\theta_1)\theta_2$ and rename Θ_1, Θ_2 as θ_1, θ_2, respectively. Then, (4.16) and (4.17) become

$$\begin{aligned}
\frac{d\theta_1}{d\tau} &= \theta_2, \qquad (4.18)\\
\frac{d\theta_2}{d\tau} &= -\chi(C - \mathscr{K}_{11}\chi) + \big[2\chi(C - \mathscr{K}_{11}\chi)\mathscr{K}_{22} - (C - 2\mathscr{K}_{11}\chi)\big]\theta_1 + \mathscr{K}_{12}\theta_1\theta_2\\
&\quad - \Big[\chi(C - \mathscr{K}_{11}\chi)\mathscr{K}_{22}^2 - 2(C - 2\mathscr{K}_{11}\chi)\mathscr{K}_{22} - \mathscr{K}_{11}\Big]\theta_1^2 + O(\|(\theta_1, \theta_2)\|^3). \qquad (4.19)
\end{aligned}$$

Let $\mathscr{J} = \chi(C - \mathscr{K}_{11}\chi)\mathscr{K}_{22}^2 - 2(C - 2\mathscr{K}_{11}\chi)\mathscr{K}_{22} - \mathscr{K}_{11}$. Since $\chi \to 0$, $C \to 0$ as $\mathscr{R}_0 \to \mathscr{R}_{BT}$ and $\nu \to \nu_{BT}$, it follows from (A4) that

$$\lim_{\substack{\mathscr{R}_0 \to \mathscr{R}_{BT} \\ \nu \to \nu_{BT}}} \mathscr{J} > \frac{1}{2S_0^*}(\mu + \alpha + \delta)\big[(\mu + \alpha + \delta)a + p(\mu + \delta)(a + \mu + \delta)\big] > 0.$$

Thus, since $\mathscr{K}_{12} \neq 0$, by making the change of variables $\Theta_1 = \mathscr{K}_{12}^2\theta_1/\mathscr{J}$, $\Theta_2 = \mathscr{K}_{12}^3\theta_2/\mathscr{J}^2$, and $t = \mathscr{J}\tau/\mathscr{K}_{12}$ in a small neighborhood of the origin and renaming Θ_1, Θ_2 as θ_1, θ_2, respectively, we have

(4.20)
$$\frac{d\theta_1}{dt} = \theta_2,$$
$$\frac{d\theta_2}{dt} = -\mathscr{K}_{12}^4\chi(C-\mathscr{K}_{11}\chi)/\mathscr{J}^3 + \mathscr{K}_{12}^2\big[2\chi(C-\mathscr{K}_{11}\chi)\mathscr{K}_{22} - (C-2\mathscr{K}_{11}\chi)\big]\theta_1/\mathscr{J}^2$$
(4.21)
$$-\theta_1^2 + \theta_1\theta_2 + O(\|(\theta_1,\theta_2)\|^3).$$

Therefore, from Theorem 8.4 and (8.52)–(8.54) in [22], the following theorem is established.

THEOREM 4.3. *Assume $\mathscr{K}_{12} \neq 0$ at the Bogdanov point. Then, the reduced model* (3.1)–(3.2) *has the following bifurcation behavior in a small neighborhood of E_0^*:*

(a) *there is a saddle-node bifurcation curve*

$$SN = \Big\{(\mathscr{R}_0,\nu):\ 4\chi(C-\mathscr{K}_{11}\chi)\mathscr{J} + \big[2\chi(C-\mathscr{K}_{11}\chi)\mathscr{K}_{22} - (C-2\mathscr{K}_{11}\chi)\big]^2 = 0\Big\};$$

(b) *there is a Hopf bifurcation curve*

$$H = \Big\{(\mathscr{R}_0,\nu):\ (\chi = 0 \Leftrightarrow B = 0),\ C > 0\Big\};$$

(c) *there is a homoclinic bifurcation curve*

$$P = \Big\{(\mathscr{R}_0,\nu):\ 2\chi(C-\mathscr{K}_{11}\chi)\mathscr{K}_{22} < C - 2\mathscr{K}_{11}\chi,$$
$$25\,\chi(C-\mathscr{K}_{11}\chi)\mathscr{J} - 6\big[2\chi(C-\mathscr{K}_{11}\chi)\mathscr{K}_{22} - (C-2\mathscr{K}_{11}\chi)\big]^2$$
$$= o(\|(\mathscr{R}_0 - \mathscr{R}_{BT}, \nu - \nu_{BT})\|^2)\Big\}.$$

Remark 4.2. The Bogdanov–Takens normal form in (4.21)–(4.21) is based on the assumption $\mathscr{K}_{12} \neq 0$ in a neighborhood of the Bogdanov point. It is worth noting that

$$\lim_{\substack{\mathscr{R}_0 \to \mathscr{R}_{BT} \\ \nu \to \nu_{BT}}} \mathscr{K}_{12} = \frac{1}{\alpha S_0^*}\Big\{-(\mu+\alpha+\delta)(\mu+\delta)^2 - p[(\mu+\alpha+\delta)^2(\mu+\delta) - \alpha^2(\mu+\alpha)]\Big\}$$
$$+ \left(\frac{\mu}{\Pi}\right)\beta(\mu+\delta)I_0^*S_0^{*p+1}f_{II}(I_0^*;\nu_{BT}).$$

Thus, for the parameter values for which

$$-(\mu+\alpha+\delta)(\mu+\delta)^2 - p[(\mu+\alpha+\delta)^2(\mu+\delta) - \alpha^2(\mu+\alpha)] < 0,$$

there is a neighborhood of $E_0^* = (S_0^*, I_0^*)$ in which $\mathscr{K}_{12} < 0$.

5. Examples. In this section, we shall detail the results of our analysis for two different forms of incidence rate. Numerical simulations are also presented to illustrate these results.

5.1. Unbounded case: $f(I;\nu) = \nu I^q$ ($\nu > 0$, $0 < q \leq 1$). In this case, if $p = 1$, then it can be seen from (3.5) that at a positive endemic equilibrium,

$$B = \frac{1}{(\mu+\alpha+\delta)(1+f)}\tilde{B}, \tag{5.1}$$

where

$$\tilde{B} = (\mu+\alpha)(\mu+\delta)\mathscr{R}_0 f^2 + \big\{(\mu+\delta)[2(\mu+\alpha)\mathscr{R}_0+\delta] - q(\mu+\alpha)(\mu+\alpha+\delta)\big\}f \\ +(\mu+\delta)[(\mu+\alpha)\mathscr{R}_0+\delta]. \tag{5.2}$$

Thus, $B(f) = 0$ if and only if $\tilde{B}(f) = 0$ and has two real roots for f (including the case of multiplicity 2) if and only if

$$\mathscr{R}_0 \leq \frac{(z_0-1)^2}{4z_1} \equiv \ell_0, \tag{5.3}$$

where

$$z_0 = \frac{\delta(\mu+\delta)}{q(\mu+\alpha)(\mu+\alpha+\delta)}, \qquad z_1 = \frac{\mu+\alpha}{\delta}z_0. \tag{5.4}$$

These roots are positive if and only if $z_0 < 1$, or equivalently,

$$q > \frac{\delta(\mu+\delta)}{(\mu+\alpha)(\mu+\alpha+\delta)}. \tag{5.5}$$

Suppose now that (5.3) and (5.5) hold. Then, there are positive ν_1, ν_2 such that $B(\nu_1) = B(\nu_2) = 0$. Differentiating B with respect to ν at ν_i and using (2.5) gives

$$\frac{dB}{d\nu}\Big|_{\nu=\nu_i} = \frac{1}{(\mu+\alpha+\delta)(1+f)}\frac{d\tilde{B}}{d\nu}\Big|_{\nu=\nu_i}, \qquad i = 1,2,$$

where, after some calculations, it can be seen that

$$\begin{aligned}\frac{d\tilde{B}}{d\nu}\Big|_{\nu=\nu_i} &= \big\{2(\mu+\alpha)(\mu+\delta)\mathscr{R}_0(1+f) - q(\mu+\alpha)(\mu+\alpha+\delta) + \delta(\mu+\delta)\big\}\frac{df}{d\nu}\Big|_{\nu=\nu_i} \\ &= \frac{2\delta(\mu+\delta)z_1^2\mathscr{R}_0}{z_0}\left\{f + 1 - \frac{1-z_0}{2z_1\mathscr{R}_0}\right\}\left(\frac{\mathscr{R}_0(1+f)^2 f I^*}{\kappa\nu\big[\mathscr{R}_0(1+f)^2 - (1+f) - qf\big]}\right)\Bigg|_{\nu=\nu_i}.\end{aligned}$$

Let $h_0 = (1-z_0)/(2z_1\mathscr{R}_0) - 1$ and $g = \mathscr{R}_0(1+f)^2 - (1+f) - qf$. If $\mathscr{R}_0 \geq 1$, then $g \geq f^2 + (1-q)f > 0$, and therefore,

$$\operatorname{sign}\left(\frac{dB}{d\nu}\Big|_{\nu=\nu_i}\right) = \operatorname{sign}(f-h_0)\big|_{\nu=\nu_i}, \qquad i = 1,2. \tag{5.6}$$

Let f_1 and f_2 (with $f_1 < f_2$) be positive roots of (5.1) when (5.3) and (5.5) hold. Then, if $\mathscr{R}_0 < \ell_0$, it follows that

$$\tilde{B}(h_0) = (\mu+\alpha)(\mu+\delta)\left[1 - \frac{(1-z_0)^2}{4z_1\mathscr{R}_0}\right] < (\mu+\alpha)(\mu+\delta)\left[1 - \frac{(1-z_0)^2}{4z_1\ell_0}\right] = 0,$$

and hence $f_1 < h_0 < f_2$. Therefore, from (5.6) it can be seen that $dB/(d\nu) < 0\,(> 0)$ at ν_1 (at ν_2). Thus, Theorem 4.1 implies that at E_2^* (where $C > 0$), the model undergoes a Hopf bifurcation which is forward at ν_1 and backward at ν_2.

THEOREM 5.1. *Suppose $1 \le \mathscr{R}_0 < \ell_0$ and $z_0 < 1$. Then the model undergoes a forward (backward) Hopf bifurcation at ν_1 (at ν_2). This bifurcation is supercritical if $\sigma < 0$ and subcritical if $\sigma > 0$.*

Suppose now $\mathscr{R}_0 < 1$. Thus, g has a unique positive root, namely f_0, such that $g < 0$ for $0 < f < f_0$ and $g > 0$ for $f > f_0$. Therefore, we have the following theorem.

THEOREM 5.2. *Suppose $\mathscr{R}^* < \mathscr{R}_0 < \min(\ell_0, 1)$ and $z_0 < 1$.*

(a) *If f_i lies in the interval between f_0 and h_0, then $dB/(d\nu) < 0$, and the model undergoes a forward supercritical (subcritical) Hopf bifurcation if $\sigma < 0$ (if $\sigma > 0$).*

(b) *If f_i lies outside the interval between f_0 and h_0, then $dB/(d\nu) > 0$, and the model undergoes a backward supercritical (subcritical) Hopf bifurcation if $\sigma < 0$ (if $\sigma > 0$).*

Remark 5.1. It can be seen, after some algebra, that for $p > 0$, $B = 0$ if and only if

$$1 - \frac{1}{w} = \tilde{z}_0 + \tilde{z}_1 u, \tag{5.7}$$

where $w = 1 + f$, $u = [\mathscr{R}_0(1 + f)]^{1/p}$, and

$$\tilde{z}_0 = \frac{(\mu + \delta)[\delta - (p-1)(\mu + \alpha)]}{q(\mu + \alpha)(\mu + \alpha + \delta)}, \qquad \tilde{z}_1 = \frac{p(\mu + \delta)}{q(\mu + \alpha + \delta)}. \tag{5.8}$$

Note that for the case $p = 1$, $\tilde{z}_0$ and $\tilde{z}_1$ reduce to z_0 and z_1, respectively, in (5.4). Equation (5.7) has a positive root of multiplicity 2 if and only if the tangency condition $\tilde{z}_1 u = p/w$ is satisfied. Then, it follows from (5.7) that there is a unique positive root u_* given by

$$u_* = \frac{p(1 - \tilde{z}_0)}{\tilde{z}_1(1 + p)}, \tag{5.9}$$

as long as $\tilde{z}_0 < 1$. This corresponds to a unique critical value of $\mathscr{R}_{crit} = \tilde{z}_1 u_*^{p+1}/p$ such that there is no Hopf bifurcation point if $\mathscr{R}_0 > \mathscr{R}_{crit}$, and two Hopf bifurcation points if $\mathscr{R}_0 < \mathscr{R}_{crit}$ and sufficiently close to $\mathscr{R}_{crit}$. These bifurcation points merge at the critical value $\mathscr{R}_{crit}$. Note that if $p = 1$, the threshold $\mathscr{R}_{crit}$ reduces to ℓ_0 in (5.3).

In the following, we consider the case where the model simultaneously undergoes a saddle-node and a Hopf bifurcation. It is easy to see that $If_I(I; \nu_{BT}) = qf(I; \nu_{BT})$, and hence from the definition of u and (4.11), we have

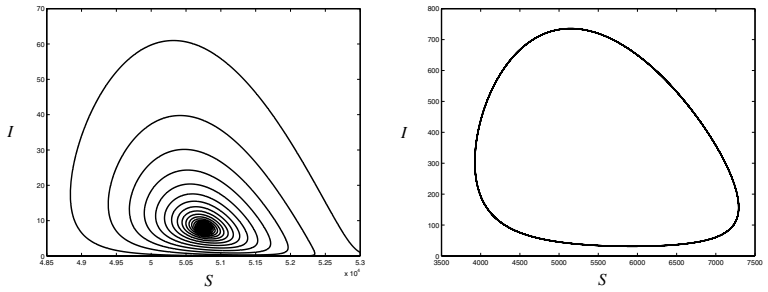
$$q\left(\frac{u^p}{\mathscr{R}_{BT}} - 1\right) = \frac{p}{\mathscr{R}_{BT}}(u - 1)u^p, \tag{5.10}$$

which implies that

$$\mathscr{R}_{BT} = \frac{u^p}{q}(p + q - pu). \tag{5.11}$$

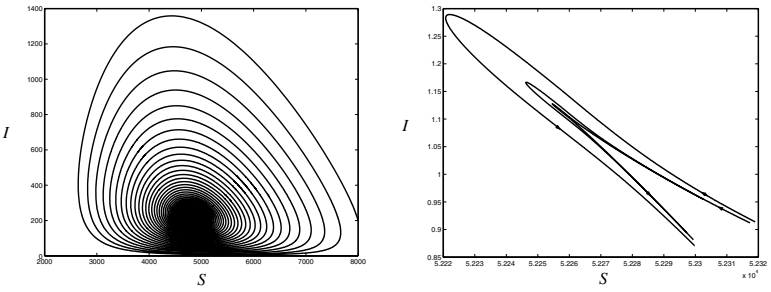
Thus, using the expression $I = \kappa(1 - 1/u)$ (see equation (2.5)), we find

$$\nu_{BT} = \frac{f(I; \nu)}{I^q} = \frac{1}{I^q}\left(\frac{u^p}{\mathscr{R}_{BT}} - 1\right). \tag{5.12}$$



(A) (B)

FIG. 5.1. (A) *Phase portrait for the stable endemic equilibrium in Example* 1 *with the parameter values* $\Pi = 1050$, $\mu = 0.02$, $\alpha = 26$, $q = 0.05$, $\delta = 0.1$, *and* $(\mathscr{R}_0, \nu) = (0.95, 0.08)$. (B) *Phase portrait for a stable periodic orbit in Example* 1 *with the parameter values:* $\Pi = 1050$, $\mu = 0.02$, $\alpha = 26$, $q = 0.05$, $\delta = 0.1$, *and* $(\mathscr{R}_0, \nu) = (0.95, 7)$.



(A) (B)

FIG. 5.2. (A) *Phase portrait for the stable endemic equilibrium in Example* 1 *with the parameter values* $\Pi = 1050$, $\mu = 0.02$, $\alpha = 26$, $q = 0.05$, $\delta = 0.1$, *and* $(\mathscr{R}_0, \nu) = (0.95, 8)$. (B) *Phase portrait at the Bogdanov point in Example* 1 *with the parameter values* $\Pi = 1050$, $\mu = 0.02$, $\alpha = 26$, $q = 0.05$, $\delta = 0.1$, *and* $(\mathscr{R}_{BT}, \nu_{BT}) = (0.9115413570, 0.1015835081)$.

It is worth noting that from (4.10), (5.11), and (5.12) it follows that the Bogdanov point for this example is unique.

To numerically illustrate the results of this bifurcation analysis, the model was simulated with parameter values estimated for a measles infection [1, 5]: $\Pi = 1050$, $\mu = 0.02$, $\alpha = 26$, $\delta = 0.1$, $q = 0.05$, and $p = 1$. With these values and $\mathscr{R}_0 = 0.95$, we have $\nu_1 = 9.49167 \times 10^{-2}$ and $\nu_2 = 7.14126$, corresponding to two Hopf bifurcation points. Numerical calculations show that $dB/d\nu < 0$, $\sigma < 0$ at ν_1, and $dB/d\nu > 0$, $\sigma < 0$ at ν_2, so that the Hopf bifurcation is supercritical at ν_1 and backward supercritical at ν_2. Figures 5.1(A)-(B) and 5.2(A) show phase portraits of the model for $\nu = 0.08 < \nu_1$ (with the stable endemic equilibrium), $\nu_1 < \nu = 7 < \nu_2$ (with the stable limit cycle), and $\nu = 8 > \nu_2$ (with a stable endemic equilibrium), respectively. The limit cycle created by the Hopf bifurcation at ν_1 shrinks and disappears when ν passes through ν_2 (see Figure 5.3(A)). Figure 5.2(B) shows the phase portrait at the Bogdanov point with $(\mathscr{R}_{BT}, \nu_{BT}) = (0.9115413570, 0.1015835081)$, which geometrically has cusp orbits. Homoclinic orbits exist as a codimension-1 family in every neighborhood of the Bogdanov point in parameter space.

In order to provide more intuition into the theoretical analysis of the model, bifurcation curves are displayed in Figure 5.4, with the same parameter values as used

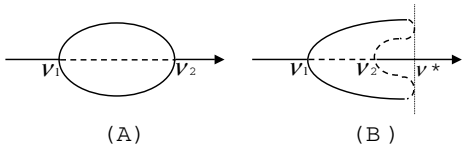


FIG. 5.3. *Bifurcation diagrams for* (A) *a supercritical Hopf bifurcation at* ν_1 *and a backward supercritical Hopf bifurcation at* ν_2*;* (B) *a supercritical Hopf bifurcation at* ν_1 *and a backward subcritical Hopf bifurcation at* ν_2*. In* (A)*, the second Hopf bifurcation at* ν_2 *causes the limit cycle created by the first Hopf bifurcation at* ν_1 *to shrink and disappear. In* (B)*, the second Hopf bifurcation at* ν_1 *leads to the appearance of an unstable limit cycle in the presence of the stable limit cycle created by the first Hopf bifurcation at* ν_1*. These limit cycles merge at a critical value of* $\nu = \nu^*$ *and disappear for* $\nu > \nu^*$*. Solid lines and curves denote stable branches, and dashed lines and curves denote unstable branches.*

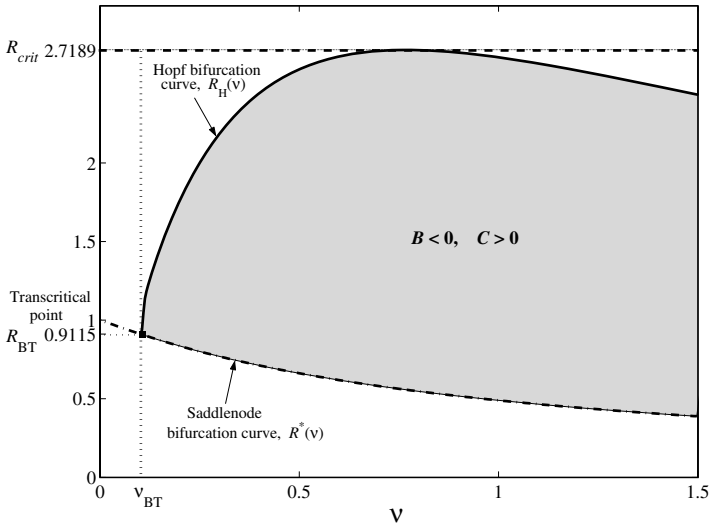


FIG. 5.4. *Bifurcation curves of the model for Example* 1*, with the parameter values* $\Pi = 1050$*,* $\mu = 0.02$*,* $\alpha = 26$*,* $q = 0.05$*,* $\delta = 0.1$*,* $p = 1$*, and* $0 \le \nu \le 1.5$*. Solid curve shows* $\mathscr{R}_H(\nu)$ *corresponding to the Hopf bifurcation along which* $B = 0$*, where* B *is defined in* (3.5)*. Dashed curve shows* $\mathscr{R}^*(\nu)$ *corresponding to the saddle-node bifurcation along which* $C = 0$*, where* C *is defined in* (3.6)*. These two curves meet at the unique Bogdanov point* $(\nu_{BT}, \mathscr{R}_{BT}) = (0.1015835081, 0.9115413570)$*. A saddle-node bifurcation occurs on the* $\mathscr{R}^*(\nu)$ *curve for any* ν*, while a Hopf bifurcation occurs on the* $\mathscr{R}_H(\nu)$ *curve (at the endemic equilibrium with high number of infected individuals) only for* $\nu > \nu_{BT}$*. Above the* $\mathscr{R}_H(\nu)$ *curve, the model exhibits no Hopf bifurcation. In the grey area* ($B < 0,\ C > 0$) *between the two curves, no bifurcation behavior occurs, and below the* $\mathscr{R}^*(\nu)$ *curve,* E_0 *is GAS.*

above. This figure shows the $\mathscr{R}^*(\nu)$ and $\mathscr{R}_H(\nu)$ curves for a range of values of the bifurcation parameter ν, along which saddle-node and Hopf bifurcations, respectively, occur. Figure 5.4 also shows the saddle-node and Hopf bifurcation branches, locally described in parts (a) and (b) of Theorem 4.3, passing through the (unique) Bogdanov point.

5.2. Bounded case: $f(I;\nu) = \nu I^q/(1+\nu I^q)$ ($\nu > 0$, $0 < q \le 1$). With this function, if $p = 1$, then at a positive endemic equilibrium

$$B = \frac{1}{(\mu+\alpha+\delta)(1+h)(1+2h)}\tilde{\tilde{B}}, \tag{5.13}$$

where

$$\tilde{\tilde{B}} = (\mu+\delta)\big[(\mu+\alpha)\mathscr{R}_0+\delta\big] + \big\{(\mu+\delta)[4(\mu+\alpha)\mathscr{R}_0+3\delta] - q(\mu+\alpha)(\mu+\alpha+\delta)\big\}h$$
$$+2(\mu+\delta)\big[2(\mu+\alpha)\mathscr{R}_0+\delta\big]h^2, \tag{5.14}$$

and $h \equiv \nu I^q$. A simple calculation shows that $\tilde{\tilde{B}}(h) = 0$ has two real roots h (including the case of multiplicity 2) if and only if

$$\mathscr{R}_0 \le \frac{\big[(3-2\sqrt{2})z_0 - 1\big]\big[(3+2\sqrt{2})z_0 - 1\big]}{8z_1}, \tag{5.15}$$

where z_0 and z_1 are defined in (5.4). If (5.15) holds, then the roots of $\tilde{\tilde{B}}(h) = 0$ are positive if and only if $z_0 < 3 - 2\sqrt{2}$, or equivalently,

$$q > \frac{(3+2\sqrt{2})\delta(\mu+\delta)}{(\mu+\alpha)(\mu+\alpha+\delta)}. \tag{5.16}$$

Suppose (5.15) and (5.16) hold and ν_1 and ν_2 correspond, respectively, to the two positive roots of $\tilde{\tilde{B}}(h) = 0$, namely h_1 and h_2 (with $h_1 < h_2$). Then,

$$\frac{dB}{d\nu}\Big|_{\nu=\nu_i} = \frac{1}{(\mu+\alpha+\delta)(1+h)(1+2h)}\frac{d\tilde{\tilde{B}}}{d\nu}\Big|_{\nu=\nu_i},$$

where

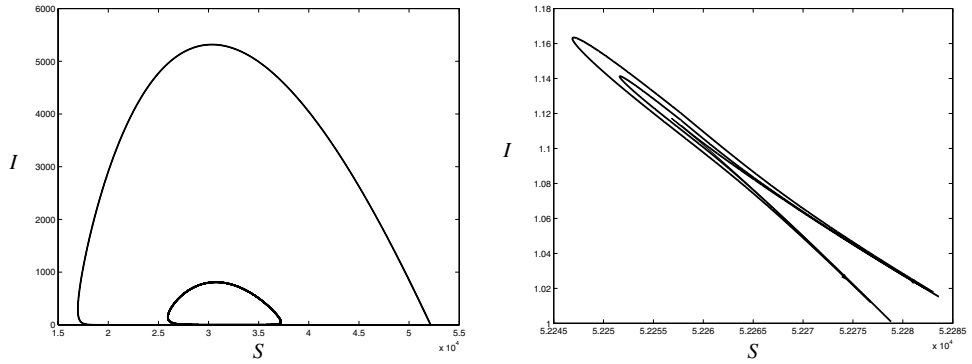
$$\frac{d\tilde{\tilde{B}}}{d\nu}\Big|_{\nu=\nu_i} = \delta(\mu+\delta)\left\{\left(\frac{8z_1\mathscr{R}_0}{z_0}+4\right)h + \frac{4z_1\mathscr{R}_0}{z_0} - \frac{1}{z_0} + 3\right\}$$
$$\times \frac{h(1+2h)}{\nu}\underbrace{\Big(\frac{(2\mathscr{R}_0-1)h + \mathscr{R}_0 - 1}{\mathscr{R}_0(1+2h)^2 - (1+h)(1+2h) - qh}\Big)\Big|_{\nu=\nu_i}}_{\Delta}.$$

Let $\tilde{h}_0 = (1-z_0)/[4(2z_1\mathscr{R}_0 + z_0)] - 1/2$. If $\mathscr{R}_0 \ge 1$, then $\Delta > 0$, and thus, $\text{sign}(dB/d\nu|_{\nu=\nu_i}) = \text{sign}(h-\tilde{h}_0)|_{\nu=\nu_i}, i = 1,2$. A simple calculation shows that $\tilde{\tilde{B}}(\tilde{h}_0) < 0$ and hence $h_1 < \tilde{h}_0 < h_2$. Therefore, $\text{sign}(dB/d\nu) = (-1)^i$ at $\nu_i, i = 1,2$. Consequently, the model undergoes a forward (backward) Hopf bifurcation at ν_1 (at ν_2). Whether these bifurcations are sub- or supercritical is determined by the sign of σ.

Suppose now $\mathscr{R}^* < \mathscr{R}_0 < 1$. From (2.5) and $\partial\phi/\partial I\big|_{I=I_0^*} = 1$, it is easy to see that $\mathscr{R}^* > 1/2$. Note that if $\mathscr{R}_0 \le 1/2$, then the model has no endemic equilibrium and the DFE is GAS. Let $L(h) = \mathscr{R}_0(1+2h)^2 - (1+h)(1+2h) - qh$. Then $L(h)$ has a unique positive root $\tilde{\tilde{h}}_0$ such that $L(h) < 0$ for $0 < h < \tilde{\tilde{h}}_0$ and $L(h) \ge 0$ for $h \ge \tilde{\tilde{h}}_0$. Defining $h_c = (1-\mathscr{R}_0)/(2\mathscr{R}_0 - 1)$, it can be shown that, regardless of whether h_c is less than or greater than $\tilde{\tilde{h}}_0$, $\Delta < 0$ in the interval between h_c and $\tilde{\tilde{h}}_0$, and $\Delta > 0$ outside this interval (see [1, Figure 9]). Therefore, $\text{sign}(dB/d\nu|_{\nu=\nu_i}) = \text{sign}(h-\tilde{h}_0)\ \text{sign}(\Delta)|_{\nu=\nu_i}, i = 1,2$.

Remark 5.2. It can be seen, after some manipulations, that for $p > 0$, $B = 0$ if and only if

$$-w + 3 - \frac{2}{w} = \tilde{z}_0 + \tilde{z}_1 u, \tag{5.17}$$



(A) (B)

FIG. 5.5. (A) *the coexistence of two limit cycles in Example* 2 *with the parameter values* $\Pi = 1050$, $\mu = 0.02$, $\alpha = 26$, $q = 0.05$, $\delta = 0.1$, $\mathscr{R}_0 = 1.05 > 1$, *and* $\nu = 1.2 > \nu_2$. *The limit cycle with large amplitude is stable, and the one with small amplitude is unstable, both surrounding the stable endemic equilibrium.* (B) *Phase portrait at the Bogdanov point in Example* 2 *with the parameter values* $\Pi = 1050$, $\mu = 0.02$, $\alpha = 26$, $q = 0.05$, $\delta = 0.1$, *and* $(\mathscr{R}_{BT}, \nu_{BT}) = (0.5612316061, 3.743081478)$.

where $w = 1 + f$, $u = [\mathscr{R}_0(1 + f)]^{1/p}$, and $\tilde{z}_0$ and $\tilde{z}_1$ are defined in (5.8). Equation (5.17) has a positive root of multiplicity 2 if and only if the tangency condition

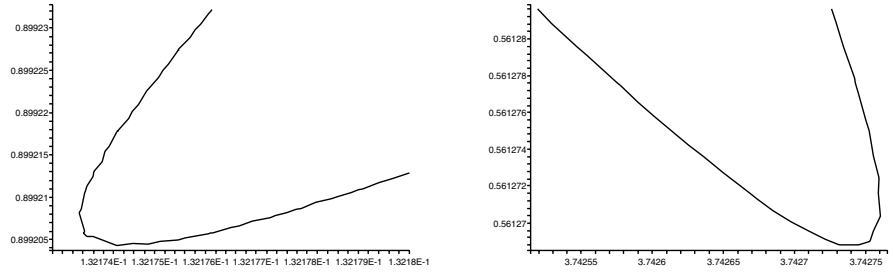
$$(p-1)w^2 + (3 - \tilde{z}_0)w - 2(p+1) = 0 \tag{5.18}$$

is satisfied. If $p < 1$, then (5.18) has two positive roots w_1, w_2 whenever $\tilde{z}_0 < 3 - 2\sqrt{2(1-p^2)}$, with $w_1 + w_2 = (3-\tilde{z}_0)/(1-p)$ and $w_1 w_2 = 2(1+p)/(1-p)$. Substituting these into (5.17) gives $u_1 + u_2 = -2p^2(3 - \tilde{z}_0)/[\tilde{z}_1(1-p^2)]$ and $\tilde{z}_1^2 u_1 u_2 = p^2[8 - (3 - \tilde{z}_0)^2]/(1-p^2)$. Therefore, if $\tilde{z}_0 \in (3 - 2\sqrt{2}, 3 - 2\sqrt{2(1-p^2)})$, then $u_1 u_2 > 0$, so that u_1 and u_2 are both negative even though $w_1, w_2 > 0$. If $0 < \tilde{z}_0 < 3 - 2\sqrt{2}$, then $u_1 u_2 < 0$, and thus there is a unique positive root u satisfying (5.17). In summary, from equations (5.17), (5.18), and the above discussion, we have

(i) for $p < 1$, there is a unique positive root u, provided $0 < \tilde{z}_0 < 3 - 2\sqrt{2}$;
(ii) for $p = 1$, there is a unique positive root u, provided $\tilde{z}_0 < 3$;
(iii) for $p > 1$, there is always a unique positive root u.

The conditions for the uniqueness of u provide a unique critical value of $\tilde{\mathscr{R}}_{crit} = u_*^p / w_*$ (where u_* and w_* are the unique solutions of (5.17) and (5.18)) such that there is no Hopf bifurcation point if $\mathscr{R}_0 > \tilde{\mathscr{R}}_{crit}$ and two Hopf bifurcation points if $\mathscr{R}_0 < \tilde{\mathscr{R}}_{crit}$ and sufficiently close to $\tilde{\mathscr{R}}_{crit}$. These bifurcation points merge at the critical value $\tilde{\mathscr{R}}_{crit}$.

The simulations were also run with the same parameter values as used in Example 1. With these values and $\mathscr{R}_0 = 1.05 > 1$, we have Hopf bifurcations at $\nu_1 = 0.147656$ and $\nu_2 = 1.12033$. Numerical calculations show that $dB/d\nu < 0\,(> 0)$, $\sigma < 0\,(> 0)$ at ν_1 (at ν_2), so that the Hopf bifurcation is supercritical at ν_1 and backward subcritical at ν_2 (Figure 5.3(B)). The bifurcation at ν_2 leads to the appearance of an unstable limit cycle in the presence of the stable limit cycle created by the Hopf bifurcation at ν_1 (see Remark 4.1). Figure 5.5(A) shows the phase portrait of the model for the coexistence of two limit cycles for $\nu = 1.2 > \nu_2$; the limit cycle with large amplitude is stable, and the one with small amplitude is unstable. Similar results were obtained for $\mathscr{R}_0 = 0.95 < 1$ with $\nu_1 = 0.127445$ ($\sigma < 0$) and $\nu_2 = 1.36961$ ($\sigma > 0$).



ν (A) ν (B)

FIG. 5.6. *Homoclinic branches ($\mathscr{R}_0$ vs ν) of the Bogdanov–Takens bifurcation for Example 2, as described in Theorem 4.3(c).*

We now consider the case where the model undergoes a Bogdanov–Takens bifurcation. It is easy to see that $If_I(I;\nu_{BT}) = qf(I;\nu_{BT})[1-f(I;\nu_{BT})]$, and hence from the definition of u and (4.11), we have

$$q\left(\frac{u^p}{\mathscr{R}_{BT}}-1\right)\left(2-\frac{u^p}{\mathscr{R}_{BT}}\right)=\frac{p}{\mathscr{R}_{BT}}(u-1)u^p, \tag{5.19}$$

which leads to the following quadratic equation:

$$q\wp^2+[p(u-1)-3q]\wp+2q=0, \tag{5.20}$$

where $\wp = u^p/\mathscr{R}_{BT}$. Let

$$\gamma \equiv p(u-1)=\frac{(\mu+\delta)(\mu+\alpha+\delta)}{\alpha(\mu+\alpha)}.$$

Then, $\Delta \equiv \gamma^2-6\gamma q+q^2=[\gamma-(3+2\sqrt{2})q][\gamma-(3-2\sqrt{2})q]>0$ if $\gamma>(3+2\sqrt{2})q$ or $\gamma<(3-2\sqrt{2})q$. Therefore, (5.20) has two positive roots if $\Delta>0$ and $\gamma<3q$, which reduce to the following condition:

$$\frac{(\mu+\delta)(\mu+\alpha+\delta)}{\alpha(\mu+\alpha)}<(3-2\sqrt{2})q. \tag{5.21}$$

The roots of (5.20) are given by

$$\wp_\pm=\frac{3q-\gamma\pm\sqrt{[(3+2\sqrt{2})q-\gamma][(3-2\sqrt{2})q-\gamma]}}{2q},$$

and thus $\mathscr{R}_{BT}^{\pm}=u^p/\wp_\pm$, where u is given by (4.10). Finally, using $I=\kappa(1-1/u)$ (see (2.5)), the expression for ν_{BT} is obtained:

$$\nu_{BT}^{\pm}=\frac{1}{I^q}\left(\frac{\wp_\pm-1}{2-\wp_\pm}\right).$$

A simple calculation yields that $\nu_{BT}^{\pm}>0$ for both positive roots of (5.20) given by $\wp_\pm$, provided that (5.21) is satisfied. This analysis shows that, in general, the values of the parameters in the parameter space $(\mathscr{R}_0,\nu)$ for which the model undergoes the Bogdanov–Takens bifurcation may not be unique. Figure 5.5(B) illustrates the phase portrait at the Bogdanov point with $(\mathscr{R}_{BT},\nu_{BT})=(0.5612316061, 3.743081478)$. Figure 5.6(A)–(B) representsbifurcation curves along which a homoclinic orbit exists in

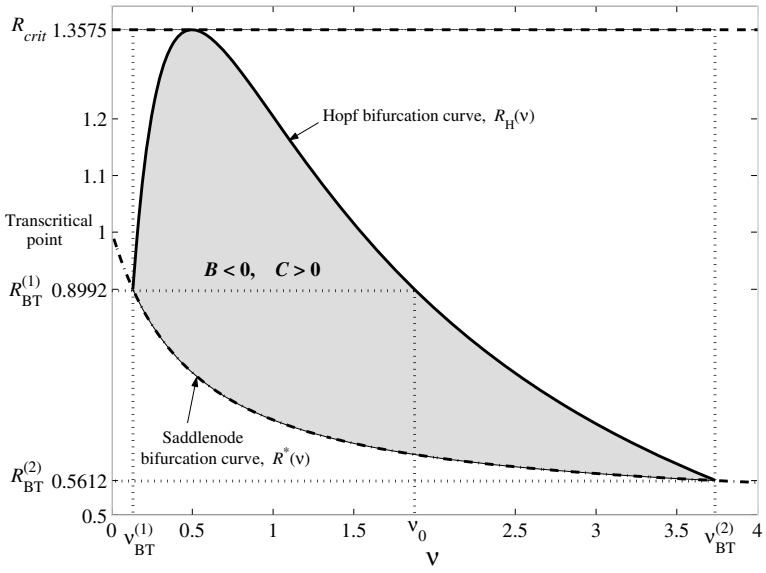


FIG. 5.7. *Bifurcation curves of the model for Example* 2*, with the parameter values* $\Pi = 1050$*,* $\mu = 0.02$*,* $\alpha = 26$*,* $q = 0.05$*,* $\delta = 0.1$*,* $p = 1$*, and* $0 \le \nu \le 4$*. Solid curve shows* $\mathscr{R}_H(\nu)$ *corresponding to the Hopf bifurcation along which* $B = 0$*, where* B *is defined in* (3.5)*. Dashed curve shows* $\mathscr{R}^*(\nu)$ *corresponding to the saddle-node bifurcation along which* $C = 0$*, where* C *is defined in* (3.6)*. These two curves meet at the Bogdanov points* $(\nu_{BT}^{(1)}, \mathscr{R}_{BT}^{(1)}) = (0.1321652741, 0.8991721941)$ *and* $(\nu_{BT}^{(2)}, \mathscr{R}_{BT}^{(2)}) = (3.743081478, 0.5612316061)$*. A saddle-node bifurcation occurs on the* $\mathscr{R}^*(\nu)$ *curve for any* ν*, while a Hopf bifurcation occurs on* $\mathscr{R}_H(\nu)$ *curve (at the endemic equilibrium with high number of infected individuals) only for* $\nu_{BT}^{(1)} < \nu < \nu_{BT}^{(2)}$*. Above the* $\mathscr{R}_H(\nu)$ *curve, the model exhibits no Hopf bifurcation. There are two Hopf bifurcation points for* $\nu \in (\nu_{BT}^{(1)}, \nu_0)$*, a unique point for* $\nu \in (\nu_0, \nu_{BT}^{(2)})$*, and no points outside of these ranges. In the grey area* ($B < 0$*,* $C > 0$) *between the two curves, no bifurcation behavior occurs, and below the* $\mathscr{R}^*(\nu)$ *curve* E_0 *is GAS.*

neighborhoods of the two Bogdanov points $(\mathscr{R}_{BT}, \nu_{BT}) = (0.8991721941, 0.1321652741)$ and $(\mathscr{R}_{BT}, \nu_{BT}) = (0.5612316061, 3.743081478)$. As for Example 1, a diagram showing saddle-node and Hopf bifurcation curves for Example 2 is given in Figure 5.7. In this case, the two bifurcation curves merge at each of the Bogdanov points.

6. Discussion. In this paper, we focused on the bifurcation analysis of an SIRS epidemic model with generalized nonlinear incidence. This study extends our previous work on the Hopf bifurcation analysis of a similar model when $p = 1$ and $\delta = 0$ [1]. Stability analysis of the model equilibria enabled us to completely analyze their local bifurcation behavior, such as Hopf, saddle-node, and Bogdanov–Takens bifurcations. We computed the first Lyapunov coefficient to determine the various types of Hopf bifurcation the model undergoes. A saddle-node bifurcation at the threshold $\mathscr{R}^*$ of the bistability region was established by applying Sotomayor's theorem. Using the Bogdanov–Takens normal form in the parameter space $(\mathscr{R}_0, \nu)$, the local representation of a homoclinic bifurcation curve was also derived. Finally, we detailed and numerically illustrated our results with two examples (bounded and unbounded) of the nonlinear function $f(I; \nu)$. These results provide the conditions for the occurrence of Hopf bifurcations in terms of two major parameters: the basic reproductive number $(\mathscr{R}_0)$ and the rate of loss of immunity acquired by infection (δ).

The parameter δ is biologically important, as its inverse represents the mean period of natural immunity following recovery. Although this parameter may change due to several factors, such as age and the response of the immune systems of recovered individuals, in general it depends on the type of the disease being modeled. For example, natural measles infection induces life-long immunity [4, 32], while natural immunity following recovery from influenza is temporary and extremely variable among different age-categories in a population [7, 9]. Here, we would like to explore the effect of this factor in producing oscillatory behavior of the model. Let us first consider Example 1. From (5.5), it can be seen that there is a critical rate ($\delta_{\max}$) such that if $\delta > \delta_{\max}$, then no Hopf bifurcation occurs, and therefore the model exhibits no periodicity behavior. This critical rate is given by

$$\delta_{\max} = \frac{1}{2}\big(\alpha + \sqrt{\alpha^2 + 4(\mu+\alpha)^2}\big).$$

If $\delta > \delta_{\max}$, then no value of $q \in (0,1]$ satisfies (5.5). By making the assumption $\alpha \gg \mu$ (which is reasonable for most curable diseases), $\delta_{\max} \approx 1.618\alpha$. The corresponding maximum rate for Example 2 can be obtained from (5.16) as

$$\delta_{\max} = \frac{1}{2(3+2\sqrt{2})}\Big(\alpha - (2+2\sqrt{2})\mu + \sqrt{[\alpha - (2+2\sqrt{2})\mu]^2 + 4(3+2\sqrt{2})(\mu+\alpha)^2}\Big),$$

with $\delta_{\max} \approx 0.509\alpha$ for $\alpha \gg \mu$. As an immediate consequence of these approximations, it can be seen that not only does $\delta_{\max}$ depend on the type of the disease, but also it depends greatly on the factors affecting the magnitude of the incidence rate, and hence on the functional form of $f(I;\nu)$.

Our illustrations, based on the parameter values estimated for measles infection (see [1, 5] and the references therein), have been verified by many clinical studies. For instance, a clinical study of measles in Poland reports an epidemic outbreak between November, 1997, and July, 1998 (with 2255 cases), despite high vaccination coverage (95%) since the 1980s [18]. The results of this study confirm that reducing $\mathscr{R}_0$ to values less than unity may fail to control the spread of the disease, which is associated with the mathematical phenomenon of backward bifurcation. Furthermore, our model exhibits oscillatory behavior, which is consistent with what has been observed for some infectious diseases such as measles, whooping cough, and rubella [10, 20, 29]. This is due to the fact that the immunity acquired following recovery from these diseases is long-lasting ($\delta \approx 0$), so that $\delta < \delta_{\max}$. This prediction is confirmed by numerical experiments (using data from England) in [29], demonstrating periods of 2 and 2–3 years for measles and whooping cough dynamics, respectively.

This study has further epidemiological implications by providing a threshold quantity $\mathscr{R}^* < 1$ (where a saddle-node bifurcation occurs at $\mathscr{R}_0 = \mathscr{R}^*$) such that the disease dies out if $\mathscr{R}_0 < \mathscr{R}^*$. Unlike the basic reproductive number, the threshold $\mathscr{R}^*$ depends on the functional form of the incidence rate as well as other parameters. If $\mathscr{R}^* < \mathscr{R}_0 < 1$, then the model exhibits two endemic equilibria which may compete with the stable disease-free equilibrium. Therefore, the long-term disease dynamics may depend on the initial values of the subpopulations. In both Examples 1 and 2, we showed that for $\mathscr{R}_0 > \mathscr{R}^*$ there exists a unique threshold $\mathscr{R}_{crit}$ such that if $\mathscr{R}_0 > \mathscr{R}_{crit}$, no Hopf bifurcation occurs. Combining the above discussion for the rate of loss of immunity, the ranges for the feasibility of periodicity behavior of the model are obtained as $\delta < \delta_{\max}$ and $\mathscr{R}^* < \mathscr{R}_0 < \mathscr{R}_{crit}$.

Acknowledgments. The authors would like to thank the referees for their insightful comments which have improved the paper.

REFERENCES

[1] M. E. Alexander and S. M. Moghadas, *Periodicity in an epidemic model with a generalized nonlinear incidence*, Math. Biosci., 189 (2004), pp. 75–96.
[2] M. A. M. Alwash, *Limit cycles in a disease transmission model*, Appl. Math. Comput., 86 (1997), pp. 85–92.
[3] R. M. Anderson and R. M. May, *Infectious Diseases of Humans*, Oxford University Press, London/New York, 1991.
[4] R. E. Behrman and R. M. Kliegman, *Nelson Essentials of Paediatrics*, Saunders, Philadelphia, 1998.
[5] B. M. Bolker and B. T. Grenfell, *Chaos and biological complexity in measles dynamics*, Philos. Trans. R. Soc. Lond. B, 251 (1993), pp. 75–81.
[6] V. Capasso and G. Serio, *A generalization of the Kermack–McKendrick deterministic epidemic model*, Math. Biosci., 42 (1978), pp. 43–61.
[7] N. J. Cox and K. Subbarao, *Influenza*, Lancet, 354 (1999), pp. 1277–1282.
[8] W. R. Derrick and P. van den Driessche, *Homoclinic orbits in a disease transmission model with nonlinear incidence and nonconstant population*, Discrete Contin. Dyn. Syst. Ser. B, 3 (2003), pp. 299–309.
[9] D. J. D Earn, J. Dushoff, and S. A. Levin, *Ecology and evolution of the flu*, Trends Ecol. Evol., 17 (2002), pp. 334–340.
[10] D. J. D. Earn, P. Rohani, B. M. Bolker, and B. T. Grenfell, *A simple model for complex dynamical transitions in epidemics*, Science, 287 (2000), pp. 667–670.
[11] P. Glendinning, *Stability, Instability, and Chaos*, Cambridge University Press, New York, 1994.
[12] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Appl. Math. Sci. 42, Springer-Verlag, New York, 1983.
[13] A. B. Gumel and S. M. Moghadas, *A qualitative study of a vaccination model with nonlinear incidence*, Appl. Math. Comput., 143 (2003), pp. 409–419.
[14] K. P. Hadeler and C. Castillo-Chavez, *A core group model for disease transmission*, Math. Biosci., 128 (1994), pp. 41–55.
[15] K. P. Hadeler and P. van den Driessche, *Backward bifurcation in epidemic control*, Math. Biosci., 146 (1997), pp. 15–35.
[16] H. W. Hethcote and P. van den Driessche, *Some epidemiological models with nonlinear incidence*, J. Math. Biol., 29 (1991), pp. 271–287.
[17] V. S. Ivlev, *Experimental Ecology of the Feeding of Fishes*, Yale University Press, New Haven, CT, 1961.
[18] W. Janaszek, N. J. Gay, and W. Gut, *Measles vaccine efficacy during an epidemic in* 1998 *in the highly vaccinated population of Poland*, Vaccine, 21 (2003), pp. 473–478.
[19] C. M. Kribs-Zaleta, *Center manifolds and normal forms in epidemic models*, in Mathematical Approaches for Emerging and Reemerging Infectious Diseases, Inst. Math. Appl. 125, Springer-Verlag, New York, 2002, pp. 269–286.
[20] M. J. Keeling, P. Rohani, and B. T. Grenfell, *Seasonally forced disease dynamics explored as switching between attractors*, Phys. D, 148 (2001), pp. 317–335.
[21] C. M. Kribs-Zaleta and J. X. Velasco-Hernández, *A simple vaccination model with multiple endemic states*, Math. Biosci., 164 (2000), pp. 183–201.
[22] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, Appl. Math. Sci. 112, Springer-Verlag, New York, 1995.
[23] S. A. Levin, T. G. Hallam, and L. J. Gross, *Applied Mathematical Ecology*, Springer-Verlag, New York, 1989.
[24] W. M. Liu, H. W. Hethcote, and S. A. Levin, *Dynamical behavior of epidemiological models with nonlinear incidence rates*, J. Math. Biol., 25 (1987), pp. 359–380.
[25] W. M. Liu, S. A. Levin, and Y. Iwasa, *Influence of nonlinear incidence rates upon the behavior of SIRS epidemiological models*, J. Math. Biol., 23 (1986), pp. 187–204.
[26] S. M. Moghadas, *Modelling the effect of imperfect vaccines on disease epidemiology*, Discrete Contin. Dyn. Syst. Ser. B, 4 (2004), pp. 999–1012.
[27] H. N. Moreira and W. Yuquan, *Global stability in an $S \to I \to R \to I$ model*, SIAM Rev., 39 (1997), pp. 496–502.
[28] A. Nold, *Heterogeneity in disease-transmission modeling*, Math. Biosci., 52 (1980), pp. 227–240.

[29] P. Rohani, M. J. Keeling, and B. T. Grenfell, *The interplay between determinism and stochasticity in childhood diseases*, Amer. Naturalist, 159 (2002), pp. 469–481.
[30] S. Ruan and W. Wang, *Dynamical behavior of an epidemic model with a nonlinear incidence rate*, J. Differential Equations, 188 (2003), pp. 135–163.
[31] P. van den Driessche and J. Watmough, *A simple SIS epidemic model with a backward bifurcation*, J. Math. Biol., 40 (2000), pp. 525–540.
[32] H. C. Whittle, P. Aaby, B. Samb, H. Jensen, J. Bennet, and F. Simondon, *Effect of subclinical infection on maintaining immunity against measles in vaccinated children in West Africa*, Lancet, 353 (1999), pp. 98–102.
[33] J. A. Yorke and W. P. London, *Recurrent outbreaks of measles, chickenpox and mumps* II, Amer. J. Epidemiol., 98 (1973), pp. 469–482.

SIAM J. APPL. MATH.
Vol. 65, No. 5, pp. 1817–1837

ASPECTS OF TOTAL VARIATION REGULARIZED L^1 FUNCTION APPROXIMATION*

TONY F. CHAN† AND SELIM ESEDOĞLU†

Abstract. The total variation–based image denoising model of Rudin, Osher, and Fatemi [*Phys. D*, 60, (1992), pp. 259–268] has been generalized and modified in many ways in the literature; one of these modifications is to use the L^1-norm as the fidelity term. We study the interesting consequences of this modification, especially from the point of view of geometric properties of its solutions. It turns out to have interesting new implications for data-driven scale selection and multiscale image decomposition.

Key words. total variation, denoising, scale space

AMS subject classifications. 94A08, 65K10

DOI. 10.1137/040604297

1. Introduction.

Variational models for image reconstruction have had great success. One of the best known and influential examples is the total variation–based model of Rudin, Osher, and Fatemi (ROF) [22]. This model and its variants have been a very active research topic. The idea behind the model is to exhibit the reconstructed image as the minimizer of the following energy:

$$\int_D |\nabla u| + \lambda \int_D (f-u)^2 \, dx. \tag{1.1}$$

The functional is to be minimized over all $u \in L^2(D)$. Here D is a domain in $\mathbf{R}^N$, $N \geq 2$, with Lipschitz boundary; it represents, for example, the computer screen. In this paper, we will work with $D = \mathbf{R}^N$ for convenience. The function $f(x)$ represents the observed and possibly degraded image and is taken to be in $L^2(D)$. The second integral in the functional is the *fidelity* term; it encourages the solution $u(x)$ that is being sought to approximate the *observed image* $f(x)$. The first integral in the functional is the *regularization* term; it is the essential novelty of the ROF model, as it allows for the reconstruction of images with discontinuities across hypersurfaces. Nevertheless, it disfavors oscillations and is responsible for the elimination of noise in applications to noisy images.

The standard ROF model (1.1) is well known to have certain limitations. One important issue is the loss of contrast in solutions even for noise-free observed images. For example, Strong and Chan studied in [25] the case when the observed image $f(x)$ is a disk and showed that the solution to (1.1), for any given λ, is of the form $cf(x)$, where $c \in [0, 1)$ is a constant. We never get $c = 1$, no matter how large the constant λ is chosen. More generally, given any observed image $f(x)$ and $\lambda > (2\|f\|_*)^{-1}$, it can be shown [15] for the corresponding solution $u(x)$ that $\|f - u\|_* = \frac{1}{2\lambda}$. Here, $\|\cdot\|_*$ denotes the dual norm of total variation. (See [15] for definition of the dual norm and

*Received by the editors February 19, 2004; accepted for publication (in revised form) November 19, 2004; published electronically July 26, 2005. This work was supported in part by NSF contract DMS-9973341, NSF contract ACI-0072112, NSF contract DMS-0410085, ONR contract N00014-03-1-0888, and NIH contract P20 MH65166.

http://www.siam.org/journals/siap/65-5/60429.html

†Mathematics Department, UCLA, Box 951555, Los Angeles, CA 90095 (TonyC@college.ucla.edu, esedoglu@math.ucla.edu).

proofs of the statements just mentioned.) It is in general desirable for image denoising algorithms to have a large class of "noise-free" images that they leave invariant. For the standard ROF model, as these results show, that class consists of only the trivial image $f(x) := 0$.

Recently, work of Meyer inspired research into understanding the role of the fidelity term better. It highlighted the fact that the choice of a suitable fidelity term can have far reaching consequences. For example, following up on Meyer's ideas, Vese and Osher [27] and then Osher, Sole, and Vese [21] came up with variants of the original model that replace the fidelity term with weaker norms. It is shown in these works that this modification allows for much better separation of the high frequency component of images, such as noise and texture, from the piecewise smooth or "cartoon" part.

In this paper, we ask related but rather different questions. We study a version of the ROF model that uses the L^1-norm as a measure of fidelity between the observed and denoised images. Given an observed image $f(x) \in L^1(\mathbf{R}^N)$, this model is based on the following variational problem:

$$\inf_{u(x)\in BV(\mathbf{R}^N)} \int_{\mathbf{R}^N} |\nabla u| + \lambda \int_{\mathbf{R}^N} |u(x) - f(x)|\, dx. \tag{1.2}$$

Our goal in this paper is to explore the consequences of this modest modification on the standard ROF model. In particular, we shall obtain some results that allow us to contrast the modified model (1.2) with the standard one (1.1). Also, the new understanding we develop about the nature of the scale space, lack of uniqueness of solutions, and lack of continuous dependence on data will suggest applications beyond mere removal of noise for the modified model: We will argue that some of these ordinarily undesirable characteristics can be real assets. Indeed, it turns out that the L^1 fidelity-based model has many desirable, and some unexpected, consequences in applications such as multiscale image decomposition and data-driven parameter selection.

Some distinctions between the modified model (1.2) and the standard ROF model (1.1) are immediate:

- The way the fidelity and regularization terms scale with respect to each other in the modified and standard models is different. In particular, unlike the standard model, the modified model is contrast invariant in the following sense: If $u(x)$ is a solution of the modified model for the observed image $f(x)$, then $cu(x)$ is a solution of the modified model for the observed image $cf(x)$.
- The original model is strictly convex, and therefore its solution (the minimizer of the functional) is unique. The modified model is not strictly convex, leading to nonuniqueness of minimizers. This makes the scale space generated by the modified model qualitatively very different—and, as explained in sections 6 and 7, for certain purposes more suitable—than that of the standard ROF model.

We concentrate especially on the scale space and geometric features of the decomposition technique derived from this model. The analytical and numerical results presented in this paper suggest the following major advantages of the L^1 fidelity-based model over the standard one:

- The regularization imposed on solutions by the L^1 model is more geometric. By "more geometric" we mean that the regularization process has less dependence on the contrast of image features than on their shapes. Indeed,

as some of our analytical results show, the L^1 model almost decouples the level sets of the given image from each other and treats them independently of their associated level (grayscale value).

- As distinct from the standard model, small features in the image maintain their contrast even as the fidelity parameter λ is lowered, maintaining good contrast until they suddenly disappear.
- An unexpected consequence of the modification is that it suggests a data-driven scale selection technique: It seems possible to identify certain critical values of the parameter λ at which features at the corresponding scale go through a discontinuous change.

Using the ROF model with L^1 fidelity is a natural idea, and was introduced and studied in the context of image denoising and deblurring by previous authors [1, 3, 4, 16, 17, 18, 8]. Among these, Alliney and Nikolova's works are relevant to ours. Alliney's previous work involves the variational model (1.2) in only one space dimension; moreover, his results are restricted to the discrete versions of the energy. Nevertheless, many of his observations are directly relevant to our results (see, for instance, Proposition 4.2 that we quote from his work), and some of our results (for instance part of Theorem 5.2) can be thought of as continuum analogues of his results in arbitrary dimensions. In [16] Nikolova shows that for certain types of noise the total variation regularization with L^1 fidelity outperforms the standard model. And [17] contains many impressive numerical results that clearly demonstrate the advantages of using the L^1 norm for a fidelity term in some applications. In fact, the analysis presented in [16] applies more generally to fidelity terms that are, like the L^1 fidelity term and unlike the L^2 fidelity term, nondifferentiable at the origin. The techniques of Nikolova also allow her to study certain typical properties of minimizers to the ROF model and its variants with different types of fidelity terms. For example, among the results is a characterization of the staircasing effect. Moreover, she calls attention to the fact that, with L^1-type fidelity terms, the solution reconstructs the given image exactly at some pixels; this relates to the contrast preserving property we touched on above. However, unlike the focus of this paper, results in [16, 17] mostly concern discrete versions of the denoising energies and depend on the discretization size; continuum analogues are not treated. Our focus in this paper is squarely on the continuum energies so that we can study geometric properties of their minimizers independently of the discretization.

We conclude the introduction with an outline of the remaining sections. Section 2 introduces the notation that is used throughout the paper. Section 3 works out the solution to minimization problems (1.1) and (1.2) in the simple case when the observed image $f(x)$ is the characteristic function of a disk in two dimensions. This illustrates some of the results obtained in subsequent sections for more general types of images. Section 4 consists of a collection of simple but useful facts that follow immediately from the definitions of section 2; these are used in the following sections of the paper. Section 5 deals with properties of minimizers of energy (1.2). In particular, it considers the case where the observed image is the characteristic function of a bounded set. It recalls the known results for the standard ROF model in this case and uses them for comparison. Section 6 elaborates on the differences between the scale spaces generated by the two models given by (1.1) and (1.2); it shows that the model based on L^1 fidelity makes it possible to determine special values of the parameter λ completely from the given observed image. Finally, section 7 presents numerical experiments and gives some implementation details. The numerical results corroborate the overall picture suggested by the analytical results of the previous sections.

2. Notation. In this section we introduce notation that will be used throughout the paper to compare the original ROF model (1.1) with the modified one (1.2) that uses an L^1 fidelity term. First, we recall the standard definitions of total variation of a function and the perimeter of a set [11, 12]. The total variation of a function $u(x) \in L^1_{loc}(\mathbf{R}^N)$ is defined to be

$$\int_{\mathbf{R}^N} |\nabla u(x)| := \sup_{\substack{\phi \in C_c^1(\mathbf{R}^N;\mathbf{R}^N) \\ |\phi(x)| \le 1 \forall x \in \mathbf{R}^N}} - \int_{\mathbf{R}^N} u(x) \operatorname{div} \phi(x)\, dx.$$

The perimeter of a set $\Sigma \subset \mathbf{R}^N$ is defined in terms of the above definition to be

$$\operatorname{Per}(\Sigma) := \int_{\mathbf{R}^N} |\nabla \mathbf{1}_\Sigma(x)|.$$

For a given possibly noisy image $f(x) \in L^1(\mathbf{R}^N)$, we will denote the energy of the total variation model with L^1 fidelity as $E_1(u, \lambda)$:

$$E_1(u, \lambda) := \int_{\mathbf{R}^N} |\nabla u| + \lambda \int_{\mathbf{R}^N} |f - u|\, dx.$$

It will be compared, for $f \in L^1(\mathbf{R}^N) \cap L^2(\mathbf{R}^N)$, with the energy of the standard ROF model, which we denote by $E_2(u, \lambda)$:

$$E_2(u, \lambda) := \int_{\mathbf{R}^N} |\nabla u| + \lambda \int_{\mathbf{R}^N} (f - u)^2\, dx.$$

Of particular interest are the minimum values of these energies as a function of the parameter λ:

$$\mathcal{E}_1(\lambda) := \min_{u \in L^1(\mathbf{R}^N)} E_1(u, \lambda),$$
$$\mathcal{E}_2(\lambda) := \min_{u \in L^2(\mathbf{R}^N)} E_2(u, \lambda).$$

Minimizers of the standard ROF energy $E_2(\cdot, \lambda)$ for a fixed λ are unique; this is a consequence of the energy's strict convexity. Minimizers of the modified energy $E_1(\cdot, \lambda)$ need not be unique in general. We therefore introduce the following notation to denote the set of minimizers of $E_1(\cdot, \lambda)$ at a given $\lambda \ge 0$:

$$M(\lambda) := \Big\{ u \in L^1(\mathbf{R}^N) : E_1(u, \lambda) = \mathcal{E}_1(\lambda) \Big\}.$$

For any given $f(x) \in L^1(\mathbf{R}^N)$ and $\lambda \ge 0$, the set $M(\lambda)$ is nonempty: A standard argument shows the existence of minimizers. Because of nonuniqueness, $M(\lambda)$ can have several elements. Different elements of $M(\lambda)$ can stand at different distances from the observed image $f(x)$. This motivates the following notation:

$$\mu^+(\lambda) := \sup \Big\{ \|f - u\|_{L^1(\mathbf{R}^N)} : u \in M(\lambda) \Big\},$$
$$\mu^-(\lambda) := \inf \Big\{ \|f - u\|_{L^1(\mathbf{R}^N)} : u \in M(\lambda) \Big\}.$$

The values of the parameter λ at which $M(\lambda)$ contains elements whose distances to the given image $f(x)$ are different turn out to be special. We therefore adopt the following notation to denote this set of special λ values:

$$S(f) := \Big\{ \lambda \in \mathbf{R}^+ : \mu^-(\lambda) \ne \mu^+(\lambda) \Big\}.$$

To emphasize the dependence of $E_i(\cdot,\lambda)$, $\mathcal{E}_i(\lambda)$,$M(\lambda)$, and $\mu^{\pm}(\lambda)$ on the observed image $f(x)$ in addition to λ, we will write $E_i(\cdot,\lambda,f),\mathcal{E}_i(\lambda,f)$, $M(\lambda,f)$, and $\mu^{\pm}(\lambda,f)$ whenever necessary.

3. An example. In this section we consider a very simple but illustrative example. Namely, we work out explicitly the solution to the problem of minimizing the two dimensional version of $E_1(\cdot,\lambda)$ in the case when the observed image $f(x)$ is given by the characteristic function $\mathbf{1}_{B_r(0)}(x)$ of a disk $B_r(0)$ that is centered at the origin and with radius r. It is important to compare the result with the one for the standard ROF model, which—as we noted in the introduction—was calculated in [25].

We start by recalling the calculation of [25]. For $\lambda \geq 0$ and the observed image given by $f(x) = \mathbf{1}_{B_r(0)}(x)$, the unique minimizer $u_\lambda(x)$ of $E_2(\cdot,\lambda)$ is given by

$$u_\lambda(x) \equiv \begin{cases} 0 & \text{if } 0 \leq \lambda \leq \dfrac{1}{r}, \\ \left(1 - \dfrac{1}{\lambda r}\right)\mathbf{1}_{B_r(0)}(x) & \text{if } \lambda > \dfrac{1}{r}. \end{cases}$$

Turning now to the case of $E_1(\cdot,\lambda)$, one can reason (for example with the help of some of the results presented in sections 5 and 6 of this paper) that for each $\lambda \geq 0$ every minimizer has to be of the form $c\mathbf{1}_{B_r(0)}(x)$ for some constant $c \in [0,1]$. We therefore need to minimize the function

$$E_1(c\mathbf{1}_{B_r(0)}(x),\lambda) = 2\pi r c + \lambda\pi r^2|1-c|$$

over $c \in [0,1]$. We get

$$M(\lambda) = \begin{cases} \{0\} & \text{if } 0 \leq \lambda < \dfrac{2}{r}, \\ \left\{c\mathbf{1}_{B_r(0)}(x) : c \in [0,1]\right\} & \text{if } \lambda = \dfrac{2}{r}, \\ \left\{\mathbf{1}_{B_r(0)}(x)\right\} & \text{if } \lambda \geq \dfrac{2}{r}. \end{cases}$$

Thus, we see that the solution is unique for all except one special value of the parameter λ. The special value is related to radius of the disk; for more general images we would expect such special values of the parameter λ to be related to the geometric scale of distinct objects contained in the scene.

The difference between scale spaces generated by the standard ROF model and the one with L^1 fidelity is made abundantly clear by this simple example. When L^1 fidelity is used, unlike in the standard ROF model, the scale space is mostly constant; it only makes a sudden transition at a special value of the scale parameter. This difference can also be manifested by plotting the "fidelity of minimizer" as a function of the parameter λ for each model and comparing the qualitative properties. Figure 1 shows the plots obtained based on the minimizers calculated above.

This example brings out another elementary aspect of using an L^1 fidelity term with total variation regularization. Fix a $\lambda > 0$. Then the unique minimizer of $E_1(\cdot,\lambda)$ with the observed image $f(x) = \mathbf{1}_{B_r(0)}(x)$ is identically 0 if $r < \frac{2}{\lambda}$, but $\mathbf{1}_{B_r(0)}(x)$ if $r > \frac{2}{\lambda}$. Thus the dependence of the solution to the L^1 model on the observed image is not continuous with respect to, say, the L^1-norm. This is clearly related to the lack of uniqueness in solutions to the model, and is a price to pay for having solutions in

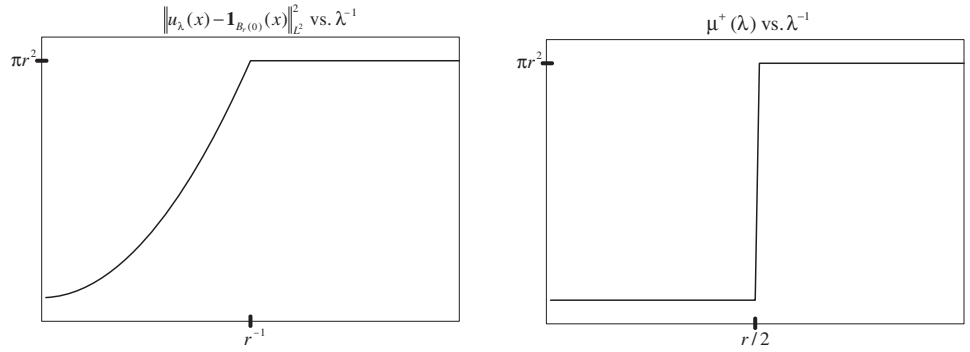


FIG. 1. *Left: Plot of* $\|u_\lambda(x) - f(x)\|_{L^2}^2$ *vs.* λ^{-1} *for the example of section* 3*, where* $u_\lambda(x)$ *denotes the unique minimizer of* $E_2(\cdot, \lambda)$*. Right: Plot of* $\mu^+(\lambda)$ *vs.* λ^{-1} *for the ROF model with* L^1 *fidelity, using the example of section* 3*.*

which features of interest maintain good contrast until they are completely eliminated. However, sections 6 and 7 explain some applications for which such a discontinuity can actually be desirable, and Proposition 6.4 shows that certain important features of the *scale space* are continuous as a function of observed signal.

4. Basic facts. In this section, we collect a number of elementary facts that follow immediately from the definitions introduced in the previous section. These results will be useful in the subsequent sections.

The following claim shows that the minimum energies $\mathcal{E}_i(\lambda)$ are well-behaved functions of the parameter λ.

CLAIM 1. *For any given observed image* $f(x) \in L^1(\mathbf{R}^N)$ *the function* $\mathcal{E}_1(\lambda)$*, and for any given observed image* $f(x) \in L^2(\mathbf{R}^N)$ *the function* $\mathcal{E}_2(\lambda)$*, satisfy the following properties:*

1. $\mathcal{E}_i(\lambda)$ *for* $i = 1, 2$ *are increasing and concave.*
2. $\mathcal{E}_i(0) = 0$ *for* $i = 1, 2$*.*
3. $0 \le \mathcal{E}_1(\lambda) \le \|f\|_{L^1}\lambda$ *and* $0 \le \mathcal{E}_2(\lambda) \le \|f\|_{L^2}^2\lambda$ *for all* $\lambda \in [0, \infty)$*.*
4. $\mathcal{E}_i(\lambda)$ *are Lipschitz continuous for* $i = 1, 2$*.*

Proof. $\mathcal{E}_i(\lambda)$ are defined as pointwise infima of a collection of linear functions that are increasing in λ; this makes them increasing and concave. Statements 2 and 3 follow from the trivial fact that $\mathcal{E}_i(\lambda) \le E_i(0, \lambda)$ for $i = 1, 2$. Statement 4 now follows from the first three. □

CLAIM 2. *The set* $M(\lambda)$ *is closed and convex.*

Proof. This follows from convexity of the energy $E_{\mathbf{1}}$. □

The following claim, which must be a well-known fact, shows that the fidelity of the minimizer to the original ROF model varies continuously as a function of λ. This should be contrasted with the results for the L^1 model that are obtained in the subsequent sections. We include its proof for completeness.

CLAIM 3. *Given* $f(x) \in L^2(\mathbf{R}^N)$*, for each* $\lambda \ge 0$ *let* $u_\lambda(x)$ *denote the unique minimizer of* $E_2(\cdot, \lambda)$*. Then the function* $\lambda \to \|f - u_\lambda\|_{L^2}$ *is continuous.*

Proof. Fix $\lambda_* \ge 0$ and let $u_{\lambda_*}(x)$ be the unique minimizer of $E_2(\cdot, \lambda_*)$. Let $\{\lambda_j\}_j^\infty \subset \mathbf{R}^+$ converge to λ_*. Consider the sequence of corresponding minimizers: $\{u_{\lambda_j}\}$. The obvious relation $E_2(u_{\lambda_j}, \lambda_j) \le E_2(0, \lambda_j) = \lambda_j\|f\|_{L^2}^2$ implies that the sequence has uniformly bounded total variation and L^2-norm. It also implies that $\|u_\lambda - f\|_{L^2} \le \|f\|_{L^2}$ for every $\lambda \ge 0$. Applying the standard compactness prop-

erty (for functions with uniformly bounded total variation) on compact sets, we can find a subsequence, also denoted $\{u_{\lambda_j}\}$, such that $u_{\lambda_j}(x) \to v(x) \in L^1_{loc}(\mathbf{R}^N)$ in L^1 on any bounded set. We may then pass to another subsequence to make sure that $u_{\lambda_j}(x) \to v(x)$ pointwise a.e. as well. Fatou's lemma then shows that $\|v - f\|_{L^2} \leq \liminf_{j\to\infty} \|u_{\lambda_j} - f\|_{L^2}$, so that in fact $v \in L^2(\mathbf{R}^N)$. Also, the standard lower semicontinuity result for total variation implies that $\int |\nabla v| \leq \liminf_{j\to\infty} \int |\nabla u_{\lambda_j}|$. Hence we get that $E_2(v, \lambda_*) \leq \liminf_{j\to\infty} E_2(u_{\lambda_j}, \lambda_j)$.

On the other hand, $E_2(u_{\lambda_*}, \lambda_*) \geq \limsup_{j\to\infty} E_2(u_{\lambda_j}, \lambda_j)$. To see this, suppose not. Then there is $\varepsilon > 0$ and arbitrarily large j such that $E_2(u_{\lambda_*}, \lambda_*) \leq E_2(u_{\lambda_j}, \lambda_j) - \varepsilon$. But also, $\lim_{j\to\infty} E_2(u_{\lambda_*}, \lambda_j) = E_2(u_{\lambda_*}, \lambda_*)$. These two statements mean $E_2(u_{\lambda_*}, \lambda_j) < E_2(u_{\lambda_j}, \lambda_j)$ for some large j, which is a contradiction, since u_{λ_j} are supposed to be minimizers of $E_2(\cdot, \lambda_j)$. This, along with the remarks of the previous paragraph, adds up to the following conclusion:

$$\limsup_{j\to\infty} E_2(u_{\lambda_j}, \lambda_j) \leq E_2(u_{\lambda_*}, \lambda_*) \leq E_2(v, \lambda_*) \leq \liminf_{j\to\infty} E_2(u_{\lambda_j}, \lambda_j).$$

We thus see that v is a minimizer of $E_2(\cdot, \lambda_*)$; by uniqueness of minimizers of $E_2(\cdot, \lambda_*)$, we get that $v = u_{\lambda_*}$.

If $\lambda_* = 0$, then $u_{\lambda_*} = 0$ and so $\|u_{\lambda_*} - f\|_{L^2} = \|f\|_{L^2}$. Recalling from above that $\|u_\lambda - f\|_{L^2} \leq \|f\|_{L^2}$ for all λ, we see that in this case

$$\limsup_{j\to\infty} \|u_{\lambda_j} - f\|_{L^2} \leq \|u_{\lambda_*} - f\|_{L^2} \leq \liminf_{j\to\infty} \|u_{\lambda_j} - f\|_{L^2},$$

which establishes continuity of the map in question at $\lambda = 0$.

If $\lambda_* > 0$, we reason as follows: We must once again have $\limsup_{j\to\infty} \|u_{\lambda_j} - f\|_{L^2} \leq \|u_{\lambda_*} - f\|_{L^2}$, which immediately leads to the conclusion of the claim. To see this, we suppose that it is false and proceed as we did in the previous paragraphs. There is then arbitrarily large j and an $\varepsilon > 0$ such that $\|u_{\lambda_*} - f\|_{L^2} \leq \|u_{\lambda_j} - f\|_{L^2} - \varepsilon$. But then

$$E_2(u_{\lambda_*}, \lambda_*) \leq \liminf_{j\to\infty} E_2(u_{\lambda_j}, \lambda_j) - \varepsilon\lambda_j.$$

Also, $E_2(u_{\lambda_*}, \lambda_j) \to E_2(u_{\lambda_*}, \lambda_*)$ as $j \to \infty$. These last two statements lead as before to the contradictory statement that $E_2(u_{\lambda_*}, \lambda_j) < E_2(u_{\lambda_j}, \lambda_j)$. ∎

We will see whether the analogue of Claim 3 holds for E_1. In that regard, we first make the following basic observation.

CLAIM 4. *Let $\lambda_2 > \lambda_1 \geq 0$, and assume that u_{λ_1} and u_{λ_2} are any two minimizers of $E_1(\cdot, \lambda_1)$ and $E_1(\cdot, \lambda_2)$, respectively. Then*

$$\|u_{\lambda_1} - f\|_{L^1(\mathbf{R}^N)} \geq \|u_{\lambda_2} - f\|_{L^1(\mathbf{R}^N)}.$$

Proof. Suppose $\|u_{\lambda_2} - f\|_{L^1} > \|u_{\lambda_1} - f\|_{L^1}$. Then, since $u_{\lambda_1} \in M(\lambda_1)$, we have $E_1(u_{\lambda_1}, \lambda_1) \leq E_1(u_{\lambda_2}, \lambda_1)$. We then have

$$\begin{aligned} E_1(u_{\lambda_1}, \lambda_2) &= E_1(u_{\lambda_1}, \lambda_1) + (\lambda_2 - \lambda_1)\|u_{\lambda_1} - f\|_{L^1} \\ &\leq E_1(u_{\lambda_2}, \lambda_1) + (\lambda_2 - \lambda_1)\|u_{\lambda_1} - f\|_{L^1} \\ &< E_1(u_{\lambda_2}, \lambda_1) + (\lambda_2 - \lambda_1)\|u_{\lambda_2} - f\|_{L^1} \\ &= E_1(u_{\lambda_2}, \lambda_2), \end{aligned}$$

which is a contradiction, since $u_{\lambda_2} \in M(\lambda_2)$ by hypothesis. ∎

COROLLARY 4.1. *The functions $\mu^{\pm}(\lambda)$ are decreasing. In fact,*

$$\mu^-(\lambda_1) \le \mu^+(\lambda_1) \le \mu^-(\lambda_2) \le \mu^+(\lambda_2)$$

whenever $\lambda_1 > \lambda_2 \ge 0$.

The functions $\mu^{\pm}(\lambda)$ are the analogue for E_1 of $\|u_\lambda - f\|_{L^2}$ in Claim 3. These functions in general can be discontinuous; in fact, their set of discontinuity is precisely $S(f)$ according to our notation. The corollary above allows us to make the following simple statement about the discontinuities of these functions.

CLAIM 5. *For any given $f \in L^1(\mathbf{R}^N)$, the set $S(f)$ is at most countable.*

Proof. If $\lambda \in S(f)$, then $\mu^-(\lambda) < \mu^+(\lambda)$. By the corollary above, at such a λ both μ^- and μ^+ have a jump discontinuity. The set of discontinuities of a monotone function are at most countable. □

Finally, for completeness let us state the following rather obvious fact about the asymptotic value of the functions $\mu^{\pm}(\lambda)$ as $\lambda \to \infty$.

CLAIM 6. *Given $f(x) \in L^1(\mathbf{R}^N)$, we have $\lim_{\lambda\to\infty} \mu^{\pm}(\lambda) = 0$.*

Proof. Given $\varepsilon > 0$, we can find $f_\varepsilon(x) \in BV(\mathbf{R}^N)$ such that $\|f_\varepsilon - f\|_{L^1} \le \frac{\varepsilon}{2}$. If $u_\lambda(x) \in M(\lambda)$ with $\mu^+(\lambda) = \|u_\lambda - f\|_{L^1}$, then

$$\mu^-(\lambda) \le \mu^+(\lambda) \le \frac{1}{\lambda} E_1(u_\lambda, \lambda) \le \frac{1}{\lambda} E_1(f_\varepsilon, \lambda) \le \frac{1}{\lambda} \int |\nabla f_\varepsilon| + \frac{\varepsilon}{2}.$$

Hence, for all large enough λ we have $\mu^{\pm}(\lambda) \le \varepsilon$. □

The following fact is taken directly from [3]. It says that any image $u_*(x)$ which arises as the solution to model (1.2) for some observed image $f(x)$ is in fact also the solution to model (1.2) with observed image $f(x)$ taken to be $u_*(x)$ itself, provided that the parameter λ is taken large enough. We include it as a good way to emphasize the difference of model (1.2) from (1.1) in regard to the loss of contrast in solutions.

PROPOSITION 4.2. *Let $\lambda_* \ge 0$, $f(x) \in L^1(\mathbf{R}^N)$, and $u_*(x) \in M(\lambda_*, f)$. Then for every $\lambda \ge \lambda_*$ we have $u_*(x) \in M(\lambda, u_*)$.*

Proof. For the proof of this claim, see [3].

5. Minimizers of E_1. In this section, we study the behavior of the ROF model with L^1 fidelity on simple images. Our motivation is twofold. First, studying the behavior of image denoising models on simple images is a first step towards understanding the type of images they can successfully process. Second, this type of question allows us to compare different models. In fact, we will stress the difference of these results from the analogous ones obtained for the standard ROF model by previous authors. In particular, our results will bolster the intuitive observation that the L^1 fidelity term leads to more geometric regularizations.

The following proposition constitutes our starting point. It shows that the ROF model with L^1 fidelity term almost decouples the level sets of the given image from each other; it almost becomes a geometry problem for each level set, independent of the level. This idea of writing total variation–based optimization problems in terms of level sets appears previously in the works [23, 24] of Strang, and is used to show the existence of binary solutions, as we do in Theorem 5.2.

PROPOSITION 5.1. *The energy $E_1(u, \lambda)$ can be rewritten as follows:*

$$E_1(u,\lambda) = \int_{-\infty}^{\infty} \mathrm{Per}\big(\{x : u(x) > \gamma\}\big) + \lambda \big|\{x : u(x) > \gamma\} \triangle \{x : f(x) > \gamma\}\big|\, d\gamma. \tag{5.1}$$

Proof. Recall the coarea formula for functions of bounded variation (see [12] or [11]):

$$\int_{\mathbf{R}^N} |\nabla u| = \int_{-\infty}^{\infty} \mathrm{Per}\big(\{x : u(x) > \gamma\}\big)\, d\gamma. \tag{5.2}$$

Also, there is the following "layer cake" formula:

$$\begin{aligned}
\int_{\mathbf{R}^N} |u-f|\,dx &= \int_{\{u>f\}} |u-f|\,dx + \int_{\{f>u\}} |u-f|\,dx \\
&= \int_{\{u>f\}} \int_{f(x)}^{u(x)} d\gamma\,dx + \int_{\{f>u\}} \int_{u(x)}^{f(x)} d\gamma\,dx \\
&= \int_{\mathbf{R}^N} \int_{\mathbf{R}} \mathbf{1}_{\{u>f\}}(x)\mathbf{1}_{[f(x),u(x))}(\gamma) + \mathbf{1}_{\{f>u\}}(x)\mathbf{1}_{[u(x),f(x))}(\gamma)\,d\gamma\,dx \\
&= \int_{\mathbf{R}} \int_{\mathbf{R}^N} \mathbf{1}_{\{u>f\}}(x)\mathbf{1}_{[f(x),u(x))}(\gamma) + \mathbf{1}_{\{f>u\}}(x)\mathbf{1}_{[u(x),f(x))}(\gamma)\,dx\,d\gamma,
\end{aligned}$$

where we simply changed the order of integration in the last step. But now we have

$$\mathbf{1}_{\{u>f\}}(x)\mathbf{1}_{[f(x),u(x))}(\gamma) = 1 \quad \text{iff } x \in \{u > f\} \cap \{u > \gamma\} \cap \{f > \gamma\}^c$$

and 0 otherwise, and

$$\mathbf{1}_{\{f>u\}}(x)\mathbf{1}_{[u(x),f(x))}(\gamma) = 1 \quad \text{iff } x \in \{f > u\} \cap \{u > \gamma\}^c \cap \{f > \gamma\}$$

and 0 otherwise. That means

$$\mathbf{1}_{\{u>f\}}(x)\mathbf{1}_{[f(x),u(x))}(\gamma) + \mathbf{1}_{\{f>u\}}(x)\mathbf{1}_{[u(x),f(x))}(\gamma) = \mathbf{1}_{\{u>\gamma\}\triangle\{f>\gamma\}}(x).$$

Therefore

$$\int_{\mathbf{R}^N} |u-f|\,dx = \int_{-\infty}^{\infty} |\{x : u(x) > \gamma\} \triangle \{x : f(x) > \gamma\}|\,d\gamma.$$

Putting these formulas together gives the one in the statement of the claim. ∎

We now explore some consequences of Proposition 5.1. First, we consider what happens when the observed image is binary. In other words, we assume that $f(x)$ is the characteristic function of a domain. We assume that the domain is bounded, but for now make no assumptions about the boundary of the domain.

THEOREM 5.2. *If the observed image $f(x)$ is the characteristic function of a bounded domain $\Omega \subset \mathbf{R}^N$, then for any $\lambda \geq 0$ there is a minimizer of $E_1(\cdot,\lambda)$ that is also the characteristic function of a (possibly different) domain. In other words, when the observed image is binary, then for each $\lambda \geq 0$ there is at least one $u(x) \in M(\lambda)$ which is also binary.*

In fact, if $u_\lambda(x) \in M(\lambda)$ is any minimizer of $E_1(\cdot,\lambda)$, then for almost every $\gamma \in [0,1]$ we have that the binary function

$$\mathbf{1}_{\{x:u_\lambda>\gamma\}}(x)$$

is also a minimizer of $E_1(\cdot,\lambda)$.

Proof. Let $f(x) := \mathbf{1}_\Omega(x)$, where Ω is a bounded domain in $\mathbf{R}^N$. It can be easily seen that any minimizer $u(x)$ of E_1 satisfies $u(x) \in [0,1]$ for almost every $x \in \mathbf{R}^N$. Formula (5.1) of Proposition 5.1 above becomes, in this case,

$$E_1(u,\lambda) = \int_0^1 \mathrm{Per}\left(\{x : u(x) > \gamma\}\right) + \lambda \left| \{x : u(x) > \gamma\} \triangle \Omega \right| d\gamma.$$

This suggests that we consider for each level set of $u(x)$ the following geometry problem:

$$\min_{\Sigma \subset \mathbf{R}^N} \Big(\mathrm{Per}(\Sigma) + \lambda \big| \Sigma \triangle \Omega \big| \Big). \tag{5.3}$$

Standard compactness and lower semicontinuity facts show the existence of minimizers; let $\Sigma_* \subset \mathbf{R}^N$ be one of them. Let $u_\lambda(x)$ be any minimizer of $E_1(\cdot,\lambda)$, i.e., $u_\lambda(x) \in M(\lambda)$. Set

$$\Sigma(\gamma) := \big\{ x : u_\lambda(x) > \gamma \big\}.$$

Then

$$\mathrm{Per}\big(\Sigma(\gamma)\big) + \lambda \big| \Sigma(\gamma) \triangle \Omega \big| \geq \mathrm{Per}(\Sigma_*) + \lambda \big| \Sigma_* \triangle \Omega \big| \tag{5.4}$$

for almost every $\gamma \in [0,\infty)$. This now immediately implies that

$$E_1(u_\lambda(x),\lambda) \geq E_1(\mathbf{1}_{\Sigma_*}(x),\lambda),$$

which means that $\mathbf{1}_{\Sigma_*}(x)$ is also a minimizer of $E(\cdot,\lambda)$.

Furthermore, since $u_\lambda(x)$ is a minimizer, the inequality of (5.4) is in fact an equality for almost every $\gamma \in [0,1]$. Thus, $\Sigma(\gamma)$ is a minimizer of the geometry problem (5.3), and $1_{\Sigma(\gamma)}(x)$ is a minimizer of $E_1(\cdot,\lambda)$ for almost every γ. ∎

Remark. A version of the first statement of Theorem 5.2 was obtained for the discrete analogue of model (1.2) in one space dimension by Alliney in [4]. ∎

Remark. The claim leaves open the possibility that for a given $\lambda \geq 0$ there might be a $u \in M(\lambda)$ that takes more than two values.

Remark. The conclusion of Theorem 5.2 is interesting because it establishes the equivalence of a nonconvex problem (the geometry problem of minimizing over only binary images, which is encountered in many applications such as improving the appearance of fax documents) to a convex problem (minimizing over all images). Indeed, it follows from the corollary that to obtain a solution to (5.3), one can first minimize $E_1(\cdot,\lambda)$, taking $f(x) = 1_\Omega(x)$ as the observed image, and then look at a level set of the solution obtained. The resulting algorithm would be very different from the standard level set method of Osher and Sethian [19, 20]. Whether this observation can be turned into a useful computational tool needs to be explored, but this question will not be pursued any further here.

Theorem 5.2 highlights an important *qualitative difference* of the L^1 model from the standard ROF model. In contrast to the content of these claims, it is easy to show that for certain types of binary images (even with smooth edge sets) the minimizer of the standard ROF model takes more than two values for every large enough choice of the parameter λ.

We do not know whether the following comparison principle holds for the geometry problem (5.3): If $\Omega_1 \subset \Omega_2$ and Σ_1, Σ_2 are minimizers of (5.3) with $\Omega = \Omega_1$ and

$\Omega = \Omega_2$, respectively, then do we necessarily have $\Sigma_1 \subset \Sigma_2$? If true, this would imply, in particular, uniqueness for solutions of (5.3). In any case, we can make the following statement.

COROLLARY 5.3. *If the observed image $f(x)$ is the characteristic function of a bounded convex domain $\Omega \subset \mathbf{R}^N$, then for almost every $\lambda \geq 0$ the minimizer of $E_1(\cdot, \lambda)$ is unique and is the characteristic function of a set contained in Ω.*

Proof. Let $\lambda \in [0, \infty) \setminus S(f)$, and let $u_\lambda(x) \in M(\lambda)$. We recall from the proof of Theorem 5.2 that, using the same notation as in that proof, the set $\Sigma(\gamma)$ minimizes the geometry problem (5.3) for almost every $\gamma \in [0, 1]$. Let $1 \geq \gamma_1 > \gamma_2 \geq 0$, and assume that $\Sigma(\gamma_1) \neq \Sigma(\gamma_2)$ both minimize the geometry problem. By definition, we have $\Sigma(\gamma_1) \subset \Sigma(\gamma_2)$. Furthermore, convexity of Ω implies that

$$\mathrm{Per}(\Sigma(\gamma_i) \cap \Omega) \leq \mathrm{Per}(\Sigma(\gamma_i)) \text{ for } i = 1, 2.$$

Since $\mathbf{1}_{\Sigma(\gamma_1)}(x)$ and $\mathbf{1}_{\Sigma(\gamma_2)}(x)$ are minimizers, it follows that $\Sigma(\gamma_1) \subset \Sigma(\gamma_2) \subseteq \Omega$. Hence, $|\Sigma(\gamma_1) \triangle \Omega| \neq |\Sigma(\gamma_2) \triangle \Omega|$. But then $\lambda \in S(f)$, which is a contradiction. We have thus reached the conclusion that if $\lambda \in [0, \infty) \setminus S(f)$, then any minimizer of $E_1(\cdot, \lambda)$ is necessarily binary (i.e., the characteristic function of a set). Now suppose that $u_1(x)$ and $u_2(x)$ are two binary minimizers of $E_1(\cdot, \lambda)$. By convexity of $E_1(\cdot, \lambda)$, we then have that $\frac{1}{2}(u_1(x) + u_2(x))$ is also a minimizer, and thus binary. But the average of two binary functions is binary only if the two functions are identical.

Thus, whenever $\lambda \in [0, \infty) \setminus S(f)$, the minimizer of $E_1(\cdot, \lambda)$ is unique and is binary: It is of the form $\mathbf{1}_\Sigma(x)$ for some set Σ. The argument above shows that $\Sigma \subseteq \Omega$. And Claim 5 says that $S(f)$ is at most countable and thus negligible. That proves the claim. ∎

As an aside, we note the following result about problem (5.3) that follows immediately from the previous corollary (perhaps it can be obtained also in a less roundabout way).

COROLLARY 5.4. *Let Ω be a bounded convex domain in $\mathbf{R}^N$. Then, for almost every $\lambda \geq 0$, the solution of problem* (5.3) *is unique.*

Proof. If Σ_1 and Σ_2 are solutions to (5.3), then $\mathbf{1}_{\Sigma_1}(x)$ and $\mathbf{1}_{\Sigma_2}(x)$ are minimizers of $E_1(\cdot, \lambda)$ with the observed image given by $f(x) = \mathbf{1}_\Omega(x)$. Conditions on Ω imply that Corollary 5.3 applies so that $\Sigma_1 = \Sigma_2$. That proves the claim. ∎

We will next consider some simple images $f(x)$ for which the minimizer of $E_1(\cdot, \lambda)$ turns out to be precisely the image $f(x)$ itself for every large enough λ. In section 1, we recalled a result from Meyer's lecture notes [15] which says that for the standard ROF model given by $E_2(\cdot, \lambda)$ the only such image is $f(x) := 0$. For E_1, however, there are many such images, as shown by Proposition 4.2, which we quoted in section 4 from [3]. The following lemma will be instrumental in establishing whether certain simple observed images $f(x)$ have this property.

LEMMA 5.5. *Given an observed image $f(x) \in BV(\mathbf{R}^N)$, assume that there is a vector field $\phi(x)$ with the following properties:*

1. *$\phi(x) \in C_c^1(\mathbf{R}^N; \mathbf{R}^N)$,*
2. *$|\phi(x)| \leq 1$ for all $x \in \mathbf{R}^N$,*
3. *$\int_{\mathbf{R}^N} f(x) \operatorname{div} \phi(x)\, dx = \int_{\mathbf{R}^N} |\nabla f|$.*

Then there exists a threshold $\lambda_* \geq 0$ such that $M(\lambda) = \{f(x)\}$ for all $\lambda > \lambda_*$. In other words, the unique minimizer of $E_1(\cdot, \lambda)$ is given by the observed image $f(x)$.

Proof. Set $\lambda_* := \max_{x \in \mathbf{R}^N} |\operatorname{div} \phi(x)|$. Take any $\lambda > \lambda_*$. Then, given any

$u(x) \in BV(\mathbf{R}^N)$, we have

$$\begin{aligned} E_1(u,\lambda) &= \int |\nabla u| + \lambda \int |u-f|\,dx \\ &\geq \int u \operatorname{div}\phi\,dx + \lambda \int |u-f|\,dx \\ &= \int f \operatorname{div}\phi\,dx + \lambda \int |u-f|\,dx + \int (u-f)\operatorname{div}\phi\,dx \\ &\geq E_1(f,\lambda) + \Big(\lambda - \max_{x\in\mathbf{R}^N} |\operatorname{div}\phi(x)|\Big) \int |u-f|\,dx. \end{aligned}$$

Since $\lambda > \lambda_* := \max|\operatorname{div}\phi(x)|$, the last inequality shows that $E_1(u,\lambda) > E_1(f,\lambda)$ unless $u \equiv f$. Since u is a minimizer, it must in fact be the case that $u \equiv f$. □

Lemma 5.5 can now be applied, for example, to binary images to obtain an important class of exact solutions. This requires making some smoothness assumption about the interface between then two values of the binary function.

THEOREM 5.6. *Let $\Omega \subset \mathbf{R}^N$ be a bounded domain with C^2 boundary. Let the observed image $f(x)$ be given by $f(x) = \mathbf{1}_\Omega(x)$. Then there exists a threshold $\lambda_* \geq 0$ such that whenever $\lambda > \lambda_*$, the unique minimizer of $E_1(\cdot,\lambda)$ is the observed image $f(x) = \mathbf{1}_\Omega(x)$ itself.*

Proof. Since the boundary $\partial\Omega$ of the bounded domain Ω is assumed to be C^2, the outward unit normal vector field $n(x) : \partial\Omega \to \mathbf{S}^{N-1}$ of $\partial\Omega$ can be extended in a C^1 manner to a tubular neighborhood of $\partial\Omega$, so that one gets a vector field $\phi(x) \in C_c^1(\mathbf{R}^N;\mathbf{R}^N)$ such that $\phi(x)\big|_{x\in\partial\Omega} = n(x)$, and $|\phi(x)| \leq 1$ for all $x \in \mathbf{R}^N$. But then

$$\begin{aligned} \int_{\mathbf{R}^N} f \operatorname{div}\phi\,dx &= \int_\Omega \operatorname{div}\phi(x)\,dx = \int_{\partial\Omega} \phi(x)\cdot n(x)\,d\sigma \\ &= \operatorname{Per}(\partial\Omega) = \int_{\mathbf{R}^N} |\nabla f|\,dx. \end{aligned}$$

Hence, the vector field $\phi(x)$ satisfies all the requirements of Lemma 5.5, from which the conclusion of the present claim follows. □

At this point it is worth recalling the behavior of the standard ROF model on binary images of the form $f(x) = 1_\Omega(x)$. As we noted above, simple considerations show that the minimizer of the standard ROF model almost never turns out to be $u(x) = f(x) = 1_\Omega(x)$. A related question is whether the solution $u(x)$ has at least the correct "set of edges"; see [10]. In case Ω is a ball, one can calculate the minimizer explicitly [25]; it turns out to be $u(x) = c\mathbf{1}_\Omega(x)$, where $c = 1 - \frac{\operatorname{Per}(\Omega)}{2\lambda|\Omega|}$. In particular, $u(x)$ has the same set of edges as $f(x)$. The results of [5] generalize the results of [25] but also show that the class of binary images that have this weaker property (i.e., images for which the solution to the standard ROF model turns out to be a constant multiple of the observed image) is still rather limited; for example, there are smooth but nonconvex shapes that lack this property.

Remark. Theorem 5.6 can easily be extended to images of a more general form. Indeed, if the level sets $\{x : f(x) = \gamma\}$ of the given image $f(x)$ are smooth and vary smoothly with respect to γ, the same conclusion holds. We also see, among other things, that such an image $f(x)$ cannot have strict local extrema, for at a strict local extremum the level sets shrink to a point. Moreover, there are also binary images that lack this property (i.e., which are not exactly recovered for any $\lambda \geq 0$, no matter how

large). In fact, a repetition of some of the arguments of Meyer given in his lecture notes [15] on the standard ROF model show that the characteristic function of, say, a square cannot arise as the solution to the ROF model with L^1 fidelity either, no matter what the observed image $f(x) \in L^1$ is, and no matter how large the parameter λ is chosen.

Remark. A discrete version of Theorem 5.6 is proved in [17] for denoising models with nonsmooth (including L^1) fidelity terms and smooth regularization terms. In those results, unlike ours, the threshold value for the parameter λ necessarily involves the grid size.

The last few claims dealt with the behavior of the L^1 fidelity-based model for large values of the parameter λ. Next, we consider what happens when $\lambda \geq 0$ is small enough. The following claim is a very simple application of the isoperimetric inequality.

PROPOSITION 5.7. *Let $R > 0$. Then there exists a threshold $\lambda_* = \lambda_*(R, N)$ such that if $f \in L^1(\mathbf{R}^N)$ with $\operatorname{supp}(f) \subset B_R(0)$, then $M(\lambda) = \{0\}$ for any $\lambda < \lambda_*$. In other words, the unique minimizer of $E_1(\cdot, \lambda)$ is given by $u(x) \equiv 0$.*

Proof. Let $C = C(N)$ be the isoperimetric constant

$$\int_{\mathbf{R}^N} |\nabla u| \geq C(N) \|u\|_{L^{\frac{N}{N-1}}(\mathbf{R}^N)} \quad \text{for all } u \in BV(\mathbf{R}^N).$$

Then we set

$$\lambda_*(R, N) := \frac{C(N)}{R\omega_N^{\frac{1}{N}}},$$

where ω_N is the volume of the unit ball in $\mathbf{R}^N$. Take a $\lambda > \lambda_*$ and let $u(x) \in M(\lambda)$. Then $E_1(u, \lambda) \leq E_1(0, \lambda)$. By the isoperimetric inequality, that means

$$C(N)\|u\|_{L^{\frac{N}{N-1}}(\mathbf{R}^N)} + \lambda\|u - f\|_{L^1(\mathbf{R}^N)} \leq \lambda\|f\|_{L^1(\mathbf{R}^N)} = \lambda\|f\|_{L^1(B_R(0))}.$$

We apply Holder's inequality to the first term on the left-hand side after splitting it into integrations over $B_R(0)$ and $B_R^c(0)$. That gives

$$\frac{C(N)}{R\omega_N^{\frac{1}{N}}} \|u\|_{L^1(B_R(0))} + \lambda\|u - f\|_{L^1(B_R(0))} + C(N)\|u\|_{L^{\frac{N}{N-1}}(B_R^c(0))} \leq \lambda\|f\|_{L^1(B_R(0))},$$

which shows that if $\lambda < C(N)/R\omega_N^{\frac{1}{N}} = \lambda_*$, then

$$\|u\|_{L^1(B_R(0))} = \|u\|_{L^{\frac{N}{N-1}}(B_R^c(0))} = 0.$$

In other words, $u \equiv 0$. □

Remark. This behavior of the L^1 model is to be expected, based on its contrast invariance, as we have already noted in the introduction. It differs from the behavior at small λ values of the standard ROF model, which, according to [15], entails not just the support of a given compactly supported image $f(x)$ but its $\|\cdot\|_*$-norm.

6. Scale space and the set $S(f)$. The set $S(f)$ of discontinuities of the functions $\mu^{\pm}$ play a distinguished role in the scale space generated by varying the parameter λ in the L^1 model. As the value of λ is gradually decreased, minimizers of the image models become coarser as small scale objects in the image merge to form larger

scale structures. Intuitively, for the L^1 model we can expect the values of $\lambda \in S(f)$ to correspond to scales of distinct objects that make up the image. These are the values of λ at which the scale space makes a rapid and drastic transition.

We would first like to prove that the set $S(f)$ is nonempty for the kind of images we have been considering in the previous sections, namely images of the form $f(x) = \mathbf{1}_\Omega(x)$, where Ω is a bounded domain. Our arguments are based on verifying this claim for the special case where the given image is the characteristic function of a ball, and then generalizing the result to $f(x) = \mathbf{1}_\Omega(x)$ by comparing Ω with a ball that is contained in Ω.

LEMMA 6.1. *Let Ω be a bounded domain in $\mathbf{R}^2$, and assume that $B_R(p) \subset \Omega$. Consider the observed image given by $f(x) = \mathbf{1}_\Omega(x)$. Then for any $\lambda \geq 0$ and $r \in (0, R)$ we have*

$$E_1(\mathbf{1}_{B_r(p)}(x), \lambda) > \min\Big\{E_1(0, \lambda), E_1(\mathbf{1}_{B_R(p)}(x), \lambda)\Big\}.$$

Proof. Since $B_r(p) \subset B_R(p) \subset \Omega$ for each $r \in (0, R)$, we have

$$\|\mathbf{1}_\Omega(x) - \mathbf{1}_{B_r(p)}(x)\|_{L^1(\mathbf{R}^2)} = |\Omega| - \pi r^2.$$

That means

$$E_1(\mathbf{1}_{B_r(p)}(x), \lambda) = \lambda\big(|\Omega| - \pi r^2\big) + 2\pi r.$$

Considering $E_1(\mathbf{1}_{B_r(p)}(x), \lambda)$ as a function of r, we see that it achieves its minimum on $[0, R]$ strictly at the end points of the interval. ∎

In order to show that $\mu^{\pm}(\lambda)$ is a discontinuous function, we will show that its range omits a full interval of values but does include certain values on either side of that interval. The next claim exhibits such an omitted interval.

LEMMA 6.2. *Let Ω be a bounded domain in $\mathbf{R}^2$, and let $B_R(0) \subset \Omega$. Consider the observed image given by $f(x) = \mathbf{1}_\Omega(x)$. There is no $\lambda \in \mathbf{R}^+$ such that*

$$|\Omega| - \pi R^2 < \mu^+(\lambda) < |\Omega|.$$

Proof. Suppose there is a $\lambda \geq 0$ such that $|\Omega| - \pi R^2 < \mu^+(\lambda) < |\Omega|$. There exists $u(x)$ such that $u(x) \in M(\lambda)$ and $\|u - f\|_{L^1(\mathbf{R}^N)} = \mu^+(\lambda)$. As before, let $\Sigma(\gamma) := \{x : u(x) > \gamma\}$. By Proposition 5.1, we have $\mathbf{1}_{\Sigma(\gamma)}(x) \in M(\lambda)$ for almost every $\gamma \in (0, 1)$. Therefore, for almost every γ we have

$$\|\mathbf{1}_{\Sigma(\gamma)}(x) - f\|_{L^1(\mathbf{R}^2)} < |\Omega|$$

(otherwise $\mu^+(\lambda) \geq |\Omega|$). It also *cannot be the case* that $|\Sigma(\gamma) \triangle \Omega| \leq |\Omega| - \pi R^2$ for almost every $\gamma \in (0, 1)$ since we know that

$$\int_0^1 |\Sigma(\gamma) \triangle \Omega|\, d\gamma = \|u - f\|_{L^1(\mathbf{R}^2)} = \mu^+(\lambda) > |\Omega| - \pi R^2.$$

Thus, there exists $\gamma_* \in (0, 1)$ such that

$$\mathbf{1}_{\Sigma(\gamma_*)}(x) \in M(\lambda) \quad \text{and} \quad |\Omega| - \pi R^2 < \|\mathbf{1}_{\Sigma(\gamma_*)}(x) - f(x)\|_{L^1(\mathbf{R}^2)} < |\Omega|.$$

Case 1. $|\Sigma(\gamma_*)| \geq \pi R^2$. But then $\mathrm{Per}(B_R(0)) \leq \mathrm{Per}(\Sigma(\gamma_*))$, and

$$|\Omega \triangle B_R(0)| = |\Omega| - \pi R^2 < \|\mathbf{1}_{\Sigma(\gamma_*)}(x) - f(x)\|_{L^1}.$$

Hence, $E_1(\mathbf{1}_{B_R(0)}(x), \lambda) < E_1(\mathbf{1}_{\Sigma(\gamma_*)}(x), \lambda)$. This is a contradiction, since $\mathbf{1}_{\Sigma(\gamma)}(x)$ was supposed to be a minimizer.

Case 2. $|\Sigma(\gamma_*)| < \pi R^2$. In this case, take $r = \frac{1}{\sqrt{\pi}}|\Sigma(\gamma_*)|^{\frac{1}{2}}$. Since $r \in (0, R)$, we have that $B_r(0) \subset \Omega$. This implies

$$\|\mathbf{1}_{B_r(0)}(x) - f(x)\|_{L^1(\mathbf{R}^2)} \leq \|\mathbf{1}_{\Sigma(\gamma_*)}(x) - f(x)\|_{L^1(\mathbf{R}^2)}.$$

Moreover, as before, $\mathrm{Per}(B_R(0)) \leq \mathrm{Per}(\Sigma(\gamma_*))$. Therefore,

$$E_1(\mathbf{1}_{B_r(0)}(x), \lambda) \leq E_1(\mathbf{1}_{\Sigma(\gamma_*)}(x), \lambda) = E_1(u(x), \lambda).$$

On the other hand, by Lemma 6.1 we have

$$E_1(\mathbf{1}_{B_r(0)}(x), \lambda) > \min\Big\{E_1(0, \lambda), E_1(\mathbf{1}_{B_R(0)}(x), \lambda)\Big\}.$$

This is a contradiction, since $u(x) \in M(\lambda)$.

THEOREM 6.3. *Let Ω be a nonempty bounded domain in $\mathbf{R}^2$. Consider the observed image given by $f(x) = \mathbf{1}_\Omega(x)$. Then the functions $\mu^{\pm}(\lambda)$ are discontinuous.*

Proof. By Proposition 5.7, we have that $\mu^+(\lambda) = \|f\|_{L^1} = |\Omega|$ for all small enough λ. On the other hand, by Claim 6 we have that $\mu^{\pm}(\lambda) \to 0$ as $\lambda \to \infty$. However, by Lemma 6.2, there is a range of values near $|\Omega|$ that the function μ^+ cannot take. It therefore has to be discontinuous. Discontinuity of μ^- follows from that of μ^+ via Claim 4. ∎

Remark. This should be contrasted with the situation for the standard total variation model (with L^2 fidelity), which is explained in Claim 3.

We thus see that the scale spaces generated by the two models, the standard ROF model and the one with L^1 fidelity, are very different. With the standard ROF model, pronounced objects of distinct scale with sharp edges in the image gradually lose their contrast and merge with their neighbors as the parameter λ is lowered. With the L^1 model, such objects maintain their contrast with respect to their neighbors—however, their boundaries might be gradually smoothed out. This goes on until a critical value of λ is reached—one that belongs to the set $S(f)$, at which point the object suddenly merges with a neighboring one.

At this point, it is also worth comparing the scale space generated by the L^1 model with that generated by anisotropic diffusion via motion by mean curvature of level sets. The two are drastically different. This can be seen most easily in the case when $f(x)$ is the characteristic function of a disk. The scale space generated by motion by curvature consists of a family of concentric disks shrinking gradually to a point. Hence the same feature, i.e., the original disk, appears at many intermediate scales, albeit in different sizes. On the other hand, the scale space generated by the total variation model with L^1 fidelity term consists of either the original disk or the constant background at any given scale.

Finally, we return to the topic of continuous dependence on the observed signal for the L^1 model. Despite our remarks in section 3, we show in the next claim that the fidelity of minimizer versus λ graph depends on the observed image continuously.

PROPOSITION 6.4. *Let $\{f_j(x)\}_{j=1}^{\infty}$ be a sequence in $L^1(\mathbf{R}^N)$ that converges to $f(x)$ in the L^1-norm. Then, for almost all $\lambda \geq 0$, $\mu^{\pm}(\lambda, f_j)$ converges to $\mu^{\pm}(\lambda, f)$ as $j \to \infty$.*

Proof. Let $\mathcal{S} := S(f) \cup (\cup_{j=1}^{\infty} S(f_j))$. According to Claim 5, $S(f)$ and each $S(f_j)$ are countable. Therefore, $\mathcal{S}$ is countable and thus negligible. Fix $\lambda \in [0, \infty) \setminus \mathcal{S}$. For

each j, take $u_j \in M(\lambda, f_j)$. The sequence $\{u_j\}_{j=1}^{\infty}$ is bounded in total variation norm and hence is precompact in L^1 on compact sets. Passing to a subsequence if necessary, we may assume that $u_j \to u_\infty$ pointwise a.e. as $j \to \infty$.

We must have $u_\infty \in M(\lambda, f)$. To see this, assume otherwise. $M(\lambda, f)$ is nonempty, so take a $u \in M(\lambda, f)$. By lower semicontinuity we have

$$E_1(u, \lambda, f) < E_1(u_\infty, \lambda, f) \le \liminf_{j\to\infty} E_1(u_j, \lambda, f_j).$$

However, $E_1(u, \lambda, f_j) \to E_1(u, \lambda, f)$ as $j \to \infty$. Therefore, for large enough j, we get $E_1(u, \lambda, f_j) < E_1(u_j, \lambda, f_j)$. This gives a contradiction, since $u_j \in M(\lambda, f_j)$.

Now that we know $u_\infty \in M(\lambda, f)$, recall next that $\lambda \notin \mathcal{S}$. Therefore,

$$\mu^{\pm}(\lambda, f) = \|u_\infty - f\|_{L^1} = \lim_{j\to\infty} \|u_j - f_j\|_{L^1} = \lim_{j\to\infty} \mu^{\pm}(\lambda, f_j).$$

That proves the claim. ∎

7. Computation. In this section, we show numerical examples that bring out unique features of the total variation–based denoising model with L^1 fidelity term. We also give some details on the numerical schemes used to obtain these results.

Our computations are based on gradient descent schemes for decreasing the energies involved. The nondifferentiability of the terms involved in the energies calls for some sort of regularization. The regularized versions of energies $E_1(\cdot, \lambda)$ and $E_2(\cdot, \lambda)$ used in our numerical experiments are the following:

$$E_1^{\varepsilon,\delta}(u, \lambda) := \int_{\mathbf{R}^N} \sqrt{|\nabla u|^2 + \varepsilon} + \lambda \int_{\mathbf{R}^N} \sqrt{(f-u)^2 + \delta}\, dx,$$
$$E_2^{\varepsilon}(u, \lambda) := \int_{\mathbf{R}^N} \sqrt{|\nabla u|^2 + \varepsilon} + \lambda \int_{\mathbf{R}^N} (f-u)^2\, dx.$$

This type of approximation to total variation–based models is very standard. The discrete versions of these energies lead to the following equally standard explicit gradient descent schemes in two space dimensions:

$$\frac{u_{i,j}^{n+1} - u_{i,j}^n}{\delta t} = D_x^- \left(\frac{D_x^+ u_{i,j}^n}{\sqrt{(D_x^+ u_{i,j}^n)^2 + (D_y^+ u_{i,j}^n)^2 + \varepsilon}} \right) + D_y^- \left(\frac{D_y^+ u_{i,j}^n}{\sqrt{(D_x^+ u_{i,j}^n)^2 + (D_y^+ u_{i,j}^n)^2 + \varepsilon}} \right) + \lambda \frac{(f - u_{i,j}^n)}{\left((f - u_{i,j}^n)^2 + \delta\right)^{\alpha}},$$

where $\alpha = \frac{1}{2}$ for $E_1^{\varepsilon,\delta}$ and $\alpha = 0$ for E_2^{ε}. Here, D^+ and D^- denote forward and backward difference quotients, respectively, in the direction of their subscript.

We note that efficient numerical minimization of energies considered in this work is a topic unto itself; no doubt there are better ways to do it than the gradient descent approach taken and the specific choice of scheme made above. In particular, it is better to use algorithms that do not need to regularize the nondifferentiable terms appearing in the energy. Such an algorithm is presented by Alliney in [2] with applications to one dimensional signals in the context of an objective functional with mixed l^1,l^2 norms. Also, Chambolle recently developed an efficient algorithm for minimizing the standard ROF model for images without regularizing the total variation term

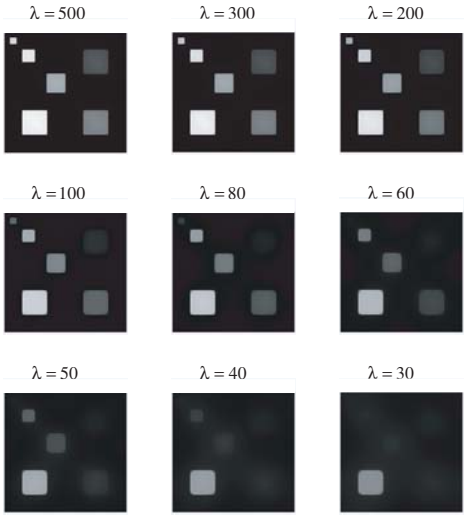


FIG. 2. *Example of scale space generated by the standard total variation model. Compare with the same example for the model with L^1 fidelity, shown in Figure* 3.

[7]. Further alternative numerical approaches to total variation–based models can be found in [6, 9]. Whether these algorithms can be adapted to our setting is a very interesting question that will be explored elsewhere.

An important point that we need to clarify is the following. Although, as we already noted several times, the energy $E_1(\cdot, \lambda)$ is not strictly convex and its minimizers in general lack uniqueness, for any given $\delta > 0$ the approximate energy $E_1^{\varepsilon,\delta}(\cdot, \lambda)$ *is* strictly convex so that its minimizers enjoy uniqueness. It is these minimizers that we have computed. Moreover, it is a very routine matter to verify that a sequence of minimizers of $E_1^{\varepsilon,\delta}(\cdot, \lambda)$ converges to the set of minimizers $M(\lambda)$ of $E_1(\cdot, \lambda)$ as $\varepsilon, \delta \to 0^+$. The analogous convergence statement is, of course, true also for a sequence of minimizers of $E_2^{\varepsilon}(\cdot, \lambda)$.

Figures 2 and 3 compare the scale spaces generated by the standard total variation model and the one with L^1 fidelity on a synthetic image. This experiment makes the more geometric nature of the L^1 model abundantly clear. The observed image consists of squares of various sizes and gray levels. In the scale space generated by the standard total variation model, the squares gradually lose their contrast (while at the same time their geometries get regularized) and gradually disappear. Moreover, some large squares with low contrast against the background—namely the square near the upper right corner—disappear before some smaller squares that have higher contrast against the background—namely the two intermediate sized squares along the diagonal. On the other hand, in the scale space generated by the model with L^1 fidelity, the squares get processed only in terms of their geometry: They preserve their contrast very well until all of a sudden they disappear. (They should, in fact, preserve their contrast perfectly, but because our numerical scheme regularizes the L^1 fidelity term to make it differentiable, in practice there is some loss of contrast.) In principle, the contrast of the squares plays no role in determining the order in which they are removed; that order is determined completely in terms of the geometry of the features.

Figure 4 shows the graph of the fidelity of the minimizer versus λ for the standard

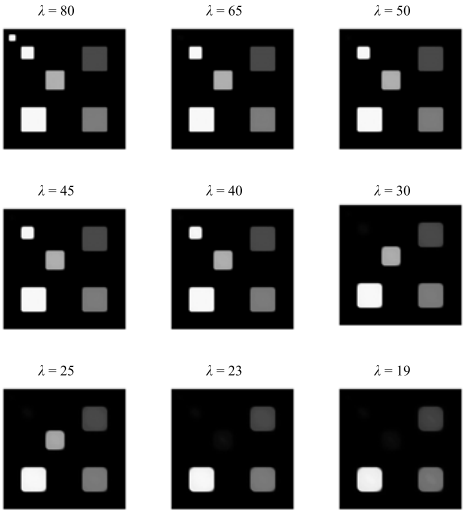


FIG. 3. *Example of scale space generated by the total variation model with L^1 fidelity. Compare with Figure 2.*

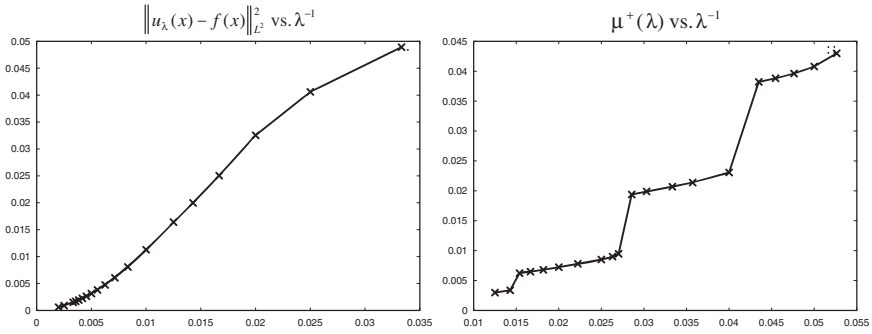


FIG. 4. *Plot of the fidelity of minimizer (i.e., $\|u_\lambda(x) - f(x)\|_{L^2}^2$) versus λ^{-1} for the standard ROF model (top graph) and of the fidelity of minimizer (i.e., $\|u_\lambda(x) - f(x)\|_{L^1}$) versus λ^{-1} for the ROF model with L^1 fidelity (bottom graph).*

total variation model and for the model with L^1 fidelity. An important ambiguity that we need to resolve is how the nonuniqueness of minimizers of $E_1(\cdot, \lambda)$ affects the fidelity-versus-λ plot for $E_1(\cdot, \lambda)$. To answer this question, recall that the fidelity of various minimizers of $E_1(\cdot, \lambda)$ differs from each other at only countably many values of λ. In particular, all ways of obtaining the second graph in Figure 4 yield plots that are identical up to a set of measure 0. Hence, there is no ambiguity in the results shown.

Discontinuities in the minimizer's fidelity-versus-λ graph for the L^1 model correspond to distinguished values of the parameter λ. As can be seen from the results, these are the values of λ at which a drastic change in the scale space takes place. Namely, at such values of λ one of the "features" (squares in this example) gets eliminated. There is no such distinguished value of λ in the plot for the standard ROF model at which the graph becomes discontinuous (as shown both by our theoretical results and by the numerical example shown). However, the graph in that case might have kinks, which are of course harder to detect than discontinuities. Thus, unlike

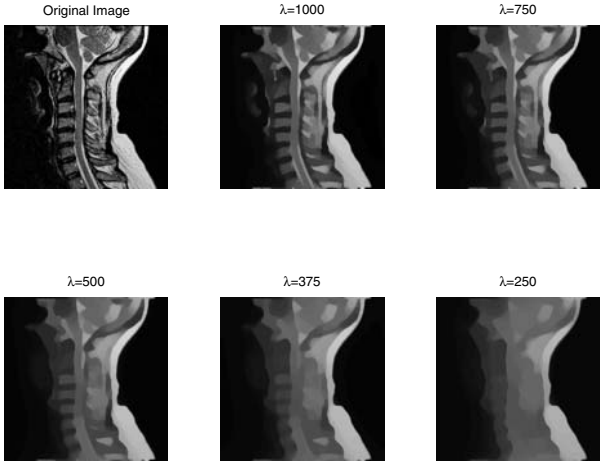


FIG. 5. *Scale space generated by the standard ROF model.*

the standard total variation model, the model with L^1 fidelity thus suggests a method for *data-driven parameter selection*.

For those familiar with the notion of an L-curve [13, 14] (which is a technique for choosing regularization parameters in ill-posed inverse problems), let us point out that from the point of view of this paper there is no apparent useful connection between the fidelity-versus-λ graph and the L-curve. According to the L-curve method, to determine a distinguished value of the regularization parameter λ, one should find the corner (point of maximum curvature) in the $\int |\nabla u_\lambda|$ versus $\|u_\lambda - f\|_{L^1}$ graph. However, for instance in the case of the example of section 3 (i.e., with $f(x) = \mathbf{1}_{B_r(0)}(x)$), the curvature of this graph is easily seen to be independent of the radius r; thus, the L-curve method does not yield any scale information.

The special values of parameter λ obtained from the fidelity-of-minimizer graph via the L^1 model can be used in many ways. For example, denoising models are sometimes used for generating multiscale decomposition of images, as in [26]. In such applications, it is necessary to select a *schedule* for the parameter λ a priori. In [26], this schedule is chosen in the form $\lambda = 2^j \lambda_0$, with $j = 1, 2, 3, \ldots$, and the initial value λ_0 is arbitrarily chosen by the user. The L^1 scale space suggests a more natural data-driven way to select these parameters using the discontinuities in the fidelity-of-minimizers graph. Moreover, even if one opts to use a λ-schedule of the form used in [26], the theoretical results and preliminary numerical examples of this paper suggest that one might obtain a much cleaner decomposition using the ROF model with L^1 fidelity in place of the standard ROF model. All these ideas pertaining to multiscale decomposition of images using the L^1 fidelity-based model will be explored elsewhere.

Finally, Figures 5 and 6 illustrate the differences between the standard ROF model and the one with L^1 fidelity on a real medical image. In this example also, one can see that the small scale features in the observed image, such as these indicated by the arrow on the lower-left-hand-side image of Figure 6, maintain their contrast much better in the L^1 fidelity model than in the standard ROF model, even as the parameter λ is gradually decreased to very low values.

8. Conclusion. We have considered the total variation–based image denoising model of Rudin, Osher, and Fatemi with the L^1-norm as the fidelity term. Our

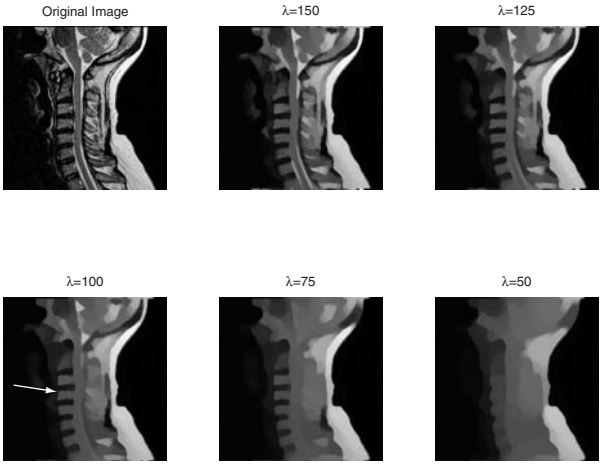


FIG. 6. *Scale space generated by the ROF model with L^1 fidelity term.*

results highlight that this modification leads to many interesting qualitative differences in the behavior of the modified model from the standard one. These differences have important consequences for image denoising. They also suggest interesting new research directions into applications to data-driven parameter selection and multiscale image decomposition.

Acknowledgments. The authors would like to thank Antonin Chambolle, Mila Nikolova, Stanley Osher, and Luminita Vese for helpful discussions.

REFERENCES

[1] S. ALLINEY, *Digital filters as absolute norm regularizers,* IEEE Trans. Signal Process., 40 (1992), pp. 1548–1562.

[2] S. ALLINEY AND S. A. RUZINSKY, *An algorithm for the minimization of mixed l_1 and l_2 norms with applications to Bayesian estimation,* IEEE Trans. Signal Process., 42 (1994), pp. 618–627.

[3] S. ALLINEY, *Recursive median filters of increasing order: A variational approach,* IEEE Trans. Signal Process., 44 (1996), pp. 1346–1354.

[4] S. ALLINEY, *A property of the minimum vectors of a regularizing functional defined by means of the absolute norm,* IEEE Trans. Signal Process., 45 (1997), pp. 913–917.

[5] G. BELLETTINI, V. CASELLES, AND M. NOVAGA, *Total variation flow in $\mathbf{R}^N$*, J. Differential Equations, 184 (2002), pp. 475–525.

[6] F. CATTÉ, F. DIBOS, AND G. KOEPFLER, *A morphological scheme for mean curvature motion and applications to anisotropic diffusion and motion of level sets,* SIAM J. Numer. Anal., 32 (1995), pp. 1895–1909.

[7] A. CHAMBOLLE, *An algorithm for total variation minimization and applications,* J. Math. Imaging and Vision, 20 (2004), pp. 89–97.

[8] E. CHEON AND A. PARANJPYE, *Noise Removal Project by Total Variation Minimization*, Math 199 Project Report, S. Osher and L. Vese, advisors, UCLA Mathematics Department, Los Angeles, 2002, http://www.math.ucla.edu/∼lvese/MATH199/index.html.

[9] F. DIBOS AND G. KOEPFLER, *Global total variation minimization,* SIAM J. Numer. Anal., 37 (2000), pp. 646–664.

[10] D. C. DOBSON AND F. SANTOSA, *Recovery of blocky images from noisy and blurred data*, SIAM J. Appl. Math., 56 (1996), pp. 1181–1198.

[11] L. C. EVANS AND R. F. GARIEPY, *Measure Theory and Fine Properties of Functions,* Stud. Adv. Math., CRC Press, Boca Raton, FL, 1992.

[12] E. GIUSTI, *Minimal Surfaces and Functions of Bounded Variation,* Monogr. Math. 80, Birkhäuser-Verlag, Basel, 1984.

[13] P. C. Hansen, *Analysis of discrete ill-posed problems by means of the L-curve*, SIAM Rev., 34 (1992), pp. 561–580.
[14] P. C. Hansen and D. P. O'Leary, *The use of the L-curve in the regularization of discrete ill-posed problems,* SIAM J. Sci. Comput., 14 (1993), pp. 1487–1503.
[15] Y. Meyer, *Oscillating patters in image processing and nonlinear evolution equations*, AMS University Lecture Series 22, AMS, Providence, RI, 2002.
[16] M. Nikolova, *Minimizers of cost-functions involving nonsmooth data-fidelity terms. Application to the processing of outliers*, SIAM J. Numer. Anal., 40 (2002), pp. 965–994.
[17] M. Nikolova, *A variational approach to remove outliers and impulse noise*, J. Math. Imaging and Vision, 20 (2004), pp. 99–120.
[18] M. Nikolova, *Weakly constrained minimization. Application to the estimation of images and signals involving constant regions*, J. Math. Imaging and Vision, 21 (2004), pp. 155–175.
[19] S. Osher and R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces,* Appl. Math. Sci. 153, Springer-Verlag, New York, 2003.
[20] J. Sethian and S. Osher, *Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulations,* J. Comput. Phys., 79 (1988), pp. 12–49.
[21] S. Osher, A. Solé, and L. Vese, *Image decomposition and restoration using total variation minimization and the H^{-1} norm*, Multiscale Model. Simul., 1 (2003), pp. 349–370.
[22] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.
[23] G. Strang, *L^1 and L^∞ approximation of vector fields in the plane*, in Nonlinear Partial Differential Equations in Applied Science (Tokyo, 1982), North–Holland Math. Stud. 81, North–Holland, Amsterdam, 1983, pp. 273–288.
[24] G. Strang, *Maximal flow through a domain,* Math. Programming, 26 (1983), pp. 123–143.
[25] D. Strong and T. F. Chan, *Edge-preserving and scale-dependent properties of total variation regularization*, Inverse Problems, 19 (2003), pp. S165–S187.
[26] E. Tadmor, S. Nezzar, and L. Vese, *A multiscale image representation using hierarchical (BV, L^2) decompositions,* Multiscale Model. Simul., 2 (2004), pp. 554–579.
[27] L. Vese and S. Osher, *Modeling textures with total variation minimization and oscillating patterns in image processing*, J. Sci. Comput., 19 (2003), pp. 553–572.

SIAM J. APPL. MATH.
Vol. 65, No. 6, pp. 1839–1854

CHANNEL FORMATION IN GELS*

N. G. COGAN† AND JAMES P. KEENER‡

Abstract. We derive and give an analysis of a model of gel dynamics based on a two-phase description of the gel, where one phase consists of networked polymer and the second phase is the fluid solvent. It is found that for the gel to maintain an edge in a poor solvent, the function describing the osmotic pressure must be of a particular form. The model is used to study the behavior of a gel forced by a pressure gradient to move between two flat plates. The distribution of the network phase under these conditions is found to be nonuniform and dependent on the pressure gradient. There is a range of pressure gradients for which the network has regions of high and low volume fraction separated by a sharp boundary, indicative of a channel. We provide the bifurcation analysis of how these novel, singularly perturbed, channeled solutions occur.

Key words. gel, model, viscoelasticity, osmotic pressure, biofilm

AMS subject classifications. 74D99, 74G10, 76T99

DOI. 10.1137/040605515

1. Introduction.

There are numerous biological and biotechnological examples where the structure and dynamics of polymer gels regulates the local environment. Biological examples include maintenance of structural integrity in biofilms [8], cellular cytoplasm [3], force generators in myxobacteria [14], and chemical diffusion and adsorption mediation in biofilm clusters [12]. Gel patches and ingestible pills used to regulate the diffusion and adsorption of drugs are examples of bioengineered gels. Quantifying the role of the polymer gel in such diverse systems requires understanding the effect of the physical and chemical structure of the polymers on the material properties of the system.

Gels are composed of a polymer network and a fluid solvent. This composition endows gels with properties different than those of viscous materials for two primary reasons. First, the polymeric structure induces viscoelastic behavior. Second, the chemical structure of the polymer induces force, causing gel swelling and deswelling. In this paper we first introduce a two-phase description of gel dynamics that emphasizes these two important differences between gels and Newtonian fluids. The behavior of a pressure driven gel between two flat plates is analyzed in a manner similar to the standard Poiseulle flow problem. Results from this analysis indicate that the steady-state network profile depends on the pressure gradient in a relatively complicated manner. There is an intermediate range of pressure gradients for which the majority of the network is compressed and located near the plates, creating a channeled region which is relatively free of polymer. This channeled solution arises via a novel bifurcation mechanism from a nearly uniform network distribution by forming a deep, narrow channel.

2. A model of gel dynamics.

Gels consist of two materials, networked polymer and fluid solvent, where the network encapsulates the solvent. The polymer

*Received by the editors March 23, 2004; accepted for publication (in revised form) December 27, 2004; published electronically August 3, 2005. This work was supported in part by NSF-FRG grant DMS 0139926.

http://www.siam.org/journals/siap/65-6/60551.html

†Department of Mathematics, Tulane University, New Orleans, LA (cogan@math.tulane.edu).

‡Department of Mathematics, University of Utah, Salt Lake City, UT (keener@math.utah.edu).

network can be formed by several different interactions between the polymers themselves including covalent bonding, coulombic bonding, hydrogen bonding, and physical entanglement.

In response to external conditions, gel networks absorb or expel solvent, causing swelling or contraction, respectively. Thus the structure of the gel depends on the temperature, solvent composition, pH, hydrostatic pressure, and ionic concentrations. The potential which is responsible for the swelling properties of the gel is referred to as osmotic or swelling pressure.

Forces due to osmotic pressure are not the only forces acting on the polymer network. Deformation of the gel induces forces due to the elastic nature of the polymer network. The elasticity is caused by both the elasticity of the polymers themselves and polymer interactions. That is, a single polymer acts as a spring for small deformations, while entanglement and cross-linking cause the network to resist deformations. The behavior is in general not well described by a simple linear relationship between displacement (strain) and stress primarily because the deformations are typically large.

Because the cross-links may be broken, a strain imposed on the gel and held induces a stress which dissipates, a process referred to as relaxation. Further, if a fixed stress is imposed on the gel, the gel will continue to displace, which is referred to as creep. The two behaviors of creep and relaxation indicate that gels are viscoelastic materials; therefore the constitutive relationship between stress and strain is typically more complicated than for viscous materials.

Here we assume that a gel is composed of two immiscible materials, polymer network and fluid solvent. The resulting model is similar to other models [3, 6, 9, 11, 13] that describe the gel as a two-phase material. The primary variation among models in the literature results from the treatment of the viscoelastic stress and the swelling pressure. In this study, we will neglect the relaxation of the network and assume that the gel is composed of an elastic solid (network) embedded in a viscous fluid (solvent). The swelling pressure is specified to ensure that physically reasonable swelling/deswelling is reflected in the deformation process.

In the following sections we describe a general model of gel dynamics and specify the forms of the viscoelastic stress and osmotic pressure used in this investigation. The resulting model is then used to study the distribution of the polymer network when the gel is forced to move between two flat plates by a pressure gradient.

2.1. Model derivation. We consider a region of space that contains networked polymer and solvent, where the volume fraction of network, θ_n, and the volume fraction of solvent, θ_s, sum to one. The network is assumed to act as a constant density viscoelastic material, while the solvent acts as a Newtonian fluid of much less viscosity than the networked material. The velocities of network and solvent are denoted $\vec{U}_n$ and $\vec{U}_s$, respectively.

The equation describing the momentum of the polymer network is given by the balance of four forces that act on the network. Surface forces are given by $\nabla \cdot (\theta_n \sigma_n)$, where σ_n is the network stress tensor. We assume that $\sigma_n = \sigma_v + \sigma_e$, where the viscous and elastic stresses are denoted σ_v and σ_e, respectively. The viscous stress tensor is proportional to the velocity gradient, $\sigma_v = \frac{\eta}{2}(\nabla \vec{U}_n + \nabla \vec{U}_n^T)$. The non-Newtonian stress tensor is given by constitutive relations which depend on the material and flow regimes [1]. Here we take the elastic stress to be proportional to the elastic strain, which is determined by the displacement gradient. We do not allow for creep or relaxation of stress. Thus, we are describing the deformation process of the moving gel. Since the displacements are not small, we use a finite strain tensor. The displacement of a fluid

particle relative to fixed Eulerian coordinates is determined by

$$\vec{x}' = \vec{x} + \vec{D}(\vec{x}, t),$$

where $\vec{x}'$ denotes the past position of the fluid particle and the components of the vector $\vec{D}$ are the displacements.

Following the development given in [1], the stress is related to the strain through

$$\sigma_e = \gamma \mathbf{C}, \tag{1}$$

where $\mathbf{C}$ is the relative Cauchy strain tensor

$$\mathbf{C}(\vec{x}, t)_{i,j} = \mathbf{F}_{ji}\mathbf{F}_{ij} - \delta_{ij},$$

with $\mathbf{F}_{ij} = \frac{\partial x_i'}{\partial x_j} = \frac{\partial D_i}{\partial x_j} + \delta_{ij}$ the deformation gradient tensor and $\delta_{ij} = 0$ if $i \neq j$, $\delta_{ii} = 1$.

We must also specify equations describing the change in displacements due to advection. The time derivative is measured in convected coordinates (i.e., relative to a fixed coordinate system). We assume that the gel is an elastic solid with rest position at which there is no network strain, while displacements from rest induce a strain on the network. Relaxation of the network has been ignored since we are primarily interested in coupling between elastic stress and network motion. Thus

$$\frac{\partial}{\partial t}\vec{D} + \nabla \cdot (\vec{D}\vec{U}_n) = \vec{U}_n. \tag{2}$$

The motion of the solvent influences the network through frictional drag, which we model by $h_f\theta_n\theta_s(\vec{U}_n - \vec{U}_s)$, where $\vec{U}_n$ and $\vec{U}_s$ are the network and solvent velocities and h_f is the constant coefficient of friction.

The third force is induced by the chemically active nature of the polymers within the gel. To model this force, we assume that there exists an osmotic pressure, $\Psi(\theta)$, gradients of which induce force on the polymers. Additional description of this term is provided below.

The final force that is included is due to hydrostatic pressure, P. Balancing all these forces yields

$$\begin{aligned}\nabla \cdot (\theta_n\sigma_n) - h_f\theta_n\theta_s(\vec{U}_n - \vec{U}_s) \\ - \nabla\Psi(\theta_n) - \theta_n\nabla P = 0.\end{aligned} \tag{3}$$

The equation governing the solvent momentum is derived in a similar manner. However, the fluid is chemically passive so there is no osmotic force on the solvent and the stress is Newtonian. Force balance yields

$$\nabla \cdot (\theta_s\sigma_s) + h_f\theta_n\theta_s(\vec{U}_n - \vec{U}_s) - \theta_s\nabla P = 0, \tag{4}$$

where $\sigma_s = \frac{\eta_s}{2}(\nabla\vec{U}_s + \nabla\vec{U}_s^T)$.

Notice that (3) and (4) are very similar to the Stokes equation for incompressible flows at zero Reynolds number. In particular, by neglecting the inertial terms, we are assuming that the system responds instantaneously to applied forces.

The redistribution of the polymer network is governed by the conservation equation

$$\frac{\partial}{\partial t}\theta_n + \nabla \cdot (\theta_n\vec{U}_n) = 0, \tag{5}$$

and a similar equation governs the conservation of solvent, namely,

$$\frac{\partial}{\partial t}\theta_s + \nabla \cdot (\theta_s \vec{U}_s) = 0. \tag{6}$$

Assuming that $\theta_n + \theta_s = 1$, we combine (5) and (6) to conclude that the divergence of the average flow, $\theta_n \vec{U}_n + \theta_s \vec{U}_s$, is zero; i.e.,

$$\nabla \cdot (\theta_n \vec{U}_n + \theta_s \vec{U}_s) = 0. \tag{7}$$

Equations (2), (3), (4), (5), and (7) govern the gel dynamics, subject to boundary conditions which depend on the specific problem. Throughout this paper, it will be useful to allow for diffusive smoothing of the network. This can be motivated physically by the fact that there is probably a small amount of polymeric diffusion within the gel. It is also useful from a mathematical perspective because it guarantees that solutions are smooth, even if there are sharp transitions. This modification yields the equation

$$\frac{\partial}{\partial t}\theta_n + \nabla \cdot (\theta_n \vec{U}_n) = \epsilon \nabla^2 \theta_n \tag{8}$$

for the redistribution of polymer network, and

$$\nabla \cdot (\theta_n \vec{U}_n + \theta_s \vec{U}_s) = \epsilon \nabla^2 \theta_n \tag{9}$$

for the incompressibility condition.

2.2. Osmotic pressure. Although there are many models of gel dynamics in the literature which include terms representing osmotic pressure [2, 3, 6, 7, 10, 11, 13], there is little agreement on either the definition or the derivation of this term. The treatment of this term varies from qualitative [3, 6] to quantitative [13]. In [7, 10, 11] a specific functional form of the osmotic pressure is not given. In fact, there has been little investigation of the dynamic behavior using different forms of the swelling pressure. Therefore our first task is to determine a model of swelling pressure which reflects some experimental results. Specifically, in many experiments a blob of gel is suspended in a solvent, causing the gel to swell. The amount of swelling is a measure of the effectiveness of the solvent. In general, the gel does not completely dissolve; instead, the blob swells a certain amount and then persists with a lower volume fraction, maintaining a distinct interface between the gel and the surrounding solvent.

We wish to determine what choice of Ψ, if any, allows for the existence of an edge between the gel and the surrounding solvent. To do so, we examine the solution of a simplified one-dimensional model of network redistribution due to swelling pressure alone. In the absence of elastic restoring force ($\sigma_e = 0$), network motion is governed by the balance of forces due to viscous stress, osmotic pressure, and hydrostatic pressure. Considering the steady-state problem implies that $\vec{U}_s = 0$ (from (5)). Using (4) to eliminate P from (3), the time independent one-dimensional equations governing network distribution are

$$\eta \frac{d}{dx}\left(\theta_n \frac{dV_n}{dx}\right) = h_f \theta_n V_n + \frac{d}{dx}\Psi(\theta_n), \tag{10}$$

$$\frac{d}{dx}(\theta_n V_n) = \epsilon \frac{d^2 \theta_n}{dx^2}, \tag{11}$$

where V_n is the x-component of $\vec{U}_n$. To be physically relevant, the solution should persist in the limit $\epsilon \to 0$.

The boundary conditions for this system are that $V_n = 0$ and there is no network flux $\epsilon \frac{d\theta_n}{dx} = \theta_n V_n$ at $x = 0$ and $x = L$, where L is the length of the one-dimensional spatial domain. This second condition allows us to integrate (11) and then substitute the result into (10), and also integrate this to find the second order system of equations

$$\eta \frac{dV_n}{dx} = \epsilon h_f + \frac{\Psi(\theta_n)}{\theta_n} + \frac{k}{\theta_n}, \tag{12}$$

$$\epsilon \frac{d\theta_n}{dx} = \theta_n V_n. \tag{13}$$

This is a singularly perturbed system. We want there to be solutions $\theta_n = 0$ and $\theta_n = \theta_{ref}$ which exist in the limit $\epsilon \to 0$ and which also can be connected by a transition layer. For $\theta_n = \theta_{ref}$ to be a solution, it must be that $k + \Psi(\theta_{ref}) = 0$, and for $\theta_n = 0$ to be a solution, it must be that

$$\lim_{\theta_n \to 0} \frac{\Psi(\theta_n)}{\theta_n} + \frac{k}{\theta_n} = 0. \tag{14}$$

It follows that

$$\Psi(\theta_n) = -k + \theta_n^2 f(\theta_n), \tag{15}$$

where $f(\theta_{ref}) = 0$. Of course, we can take $k = 0$, since only the gradient of Ψ appears in the governing equations.

Now we seek a transition layer that connects the two solutions $\theta_n = 0$ and $\theta_n = \theta_{ref}$. In this transition layer it must be that (ignoring the term ϵh_f)

$$\frac{\eta}{\epsilon} \frac{dV_n}{d\theta_n} = \frac{f(\theta_n)}{V_n}, \tag{16}$$

from which it follows that

$$\frac{\eta}{\epsilon} V_n^2 = -\int_{\theta_n}^{\theta_{ref}} f(\theta) d\theta, \tag{17}$$

implying that $f(\theta_n) < 0$ on the interval $0 \le \theta_n \le \theta_{ref}$. In the special case that $f(\theta_n) = \gamma_{os}(\theta_n - \theta_{ref})$, we find that

$$V_n = \pm \sqrt{\frac{\gamma_{os}\epsilon}{\eta}} (\theta_n - \theta_{ref}), \tag{18}$$

with transition layer trajectory satisfying

$$\frac{d\theta_n}{dx} = \pm \sqrt{\frac{\gamma_{os}}{\epsilon\eta}} \theta_n (\theta_n - \theta_{ref}), \tag{19}$$

a hyperbolic tangent solution with boundary layer width the order of $\sqrt{\epsilon}$.

It follows that, for a gel to hold an edge, Ψ must be of the form (up to an arbitrary additive constant) $\Psi(\theta_n) = \theta_n^2 f(\theta_n)$ with $f(\theta_{ref}) = 0$ and $f(\theta_n) \le 0$ for $0 < \theta_n < \theta_{ref}$. A specific example of such a function that we use throughout the rest of this paper is $\Psi(\theta_n) = \gamma_{os}\theta_n^2(\theta_n - \theta_{ref})$.

Having determined the form of the osmotic pressure that allows transition layers between $\theta_n = 0$ and $\theta_n = \theta_{ref}$, we now wish to determine the stability of steady solutions. Notice that any constant $\theta_n = \theta_0$ is a solution of (3), (4), (8), and (9) with $\vec{U}_n = \vec{U}_s = 0$ (provided $\sigma_e = 0$). To study its stability, we linearize the governing equations about this uniform solution (setting $\vec{U}_n = u$, $\vec{U}_s = v$, $\theta_n = \theta_0 + \phi$, $\theta_s = 1 - \theta_0 - \phi$.), to find

$$\frac{\partial \phi}{\partial t} + \nabla \cdot (u\theta_0) = \epsilon \nabla^2 \phi, \tag{20}$$

$$\nabla \cdot (u\theta_0 + v(1-\theta_0)) = \epsilon \nabla^2 \phi, \tag{21}$$

$$\frac{1}{2}\nabla \cdot (\theta_0 \eta(\nabla u + \nabla u^T)) - h_f \theta_0 (u - v) - \Psi'(\theta_0)\nabla\phi = 0. \tag{22}$$

To find the dispersion relation for this problem, we try a solution of the form $\phi = A(t)e^{i\omega\cdot x}$, $u = B(t)e^{i\omega\cdot x}$, $v = C(t)e^{i\omega\cdot x}$, and obtain equations for A, B, and C:

$$\frac{dA}{dt} + i\omega(B\theta_0) = -\omega^2 \epsilon A, \tag{23}$$

$$i\omega(B\theta_0 + C(1-\theta_0)) = -\omega^2 \epsilon A, \tag{24}$$

$$-\omega^2 \theta_0 \eta B - h_f \theta_0 (B - C) - \Psi'(\theta_0) i\omega A = 0. \tag{25}$$

We solve for B and C and substitute into (23) to find

$$C = \frac{i\omega\epsilon A - B\theta_0}{1-\theta_0}, \tag{26}$$

$$B = \frac{\epsilon h_f \theta_0 - (1-\theta_0)\Psi'(\theta_0)}{\omega^2 \theta_0 \eta (1-\theta_0) + h_f \theta_0} i\omega A, \tag{27}$$

$$\frac{dA}{dt} = \omega^2 \frac{\epsilon h_f \theta_0 - (1-\theta_0)\Psi'(\theta_0)}{\omega^2 \eta (1-\theta_0) + h_f} A - \omega^2 \epsilon A. \tag{28}$$

In the limit $\epsilon \to 0$ this is

$$\frac{dA}{dt} = -\omega^2 \frac{(1-\theta_0)\Psi'(\theta_0)}{\omega^2 \eta (1-\theta_0) + h_f} A. \tag{29}$$

Clearly, this is stable if $\psi'(\theta_0) > 0$ and unstable if $\psi'(\theta_0) < 0$.

Thus, for the function $\Psi(\theta_n) = \gamma_{os}\theta_n^2(\theta_n - \theta_{ref})$, a uniform gel with $\theta_n < \gamma_{os}\frac{2}{3}\theta_{ref}$ is unstable, while a uniform gel with $\theta_n > \gamma_{os}\frac{2}{3}\theta_{ref}$ is stable. Thus, very low density gels are not stable and will tend to form deswelled spatially localized aggregates.

3. Channeling behavior. We now turn to a simple problem illustrating one difference between gel dynamics and Newtonian fluid dynamics. We consider the deformation of the network component of a gel which is forced to move between two flat plates due to a constant imposed pressure drop. We make one further assumption that the viscosity of the system is dominated by the network viscosity ($\theta_n \eta \gg \theta_s \eta_s$), and thus, following [13], we neglect the solvent viscosity. Notice that although the solvent is inviscid, the frictional interaction between the solvent and the network still renders the entire system viscous.

The motion is assumed to be two-dimensional, where x, y and $\vec{U}_* = (V_*, W_*)$ denote the horizontal and vertical coordinates and velocities, respectively. For Newtonian fluids the steady-state x independent velocity profile is parabolic in y for all pressure drops. This is not the case for the gel-Poiseulle flow. Instead, the steady-state profile of the network volume fraction undergoes a large change as the magnitude of the pressure gradient varies.

To demonstrate this, we seek a solution of (3)–(7) that is the analogue of Poiseulle flow—the horizontally independent steady velocity profile for a fluid forced between two flat plates by a pressure drop.

Under the assumption that D_1 and D_2 are independent of x, the elements of the deformation gradient tensor $\mathbf{F}_{ij}$ are

$$\begin{aligned} \frac{\partial x'}{\partial x} &= 1, \\ \frac{\partial x'}{\partial y} &= \frac{\partial D_1}{\partial y}, \\ \frac{\partial y'}{\partial x} &= 0, \\ \frac{\partial y'}{\partial y} &= 1 + \frac{\partial D_2}{\partial y}, \end{aligned}$$

and the stress tensor becomes

$$\sigma_e = \gamma \begin{bmatrix} 0 & \frac{\partial D_1}{\partial y} \\ \frac{\partial D_1}{\partial y} & \left(\frac{\partial D_1}{\partial y}\right)^2 + 2\frac{\partial D_2}{\partial y} + \left(\frac{\partial D_2}{\partial y}\right)^2 \end{bmatrix}.$$

We change from vector to component notation here, so that the following simplifications are more apparent. In component form the steady-state equations for the gel-Poiseulle flow are

$$\eta \frac{\partial}{\partial y}\left(\theta_n \frac{\partial}{\partial y} V_n\right) - \frac{\partial P}{\partial x} + \gamma \frac{\partial}{\partial y}\left(\theta_n \frac{\partial}{\partial y} D_1\right) = 0, \tag{30}$$

$$\begin{aligned} \eta \frac{\partial}{\partial y}\left(\theta_n \frac{\partial}{\partial y} W_n\right) - \frac{\partial}{\partial y}\Psi(\theta_n) - \frac{\partial P}{\partial y} \\ + \gamma \frac{\partial}{\partial y}\left(\theta_n \left(\left(\frac{\partial D_1}{\partial y}\right)^2 + 2\frac{\partial D_2}{\partial y} + \left(\frac{\partial D_2}{\partial y}\right)^2\right)\right) = 0, \end{aligned} \tag{31}$$

$$h_f \theta_n (V_n - V_s) - \frac{\partial P}{\partial x} = 0, \tag{32}$$

$$h_f \theta_n (W_n - W_s) - \frac{\partial P}{\partial y} = 0, \tag{33}$$

$$\frac{\partial}{\partial y}(\theta_n W_n + (1 - \theta_n) W_s) = \epsilon \frac{\partial^2 \theta_n}{\partial y^2}, \tag{34}$$

$$\frac{\partial}{\partial y}(\theta_n W_n) = \epsilon \frac{\partial^2 \theta_n}{\partial y^2}, \tag{35}$$

$$\frac{\partial}{\partial y}(D_1 W_n) = V_n, \tag{36}$$

$$\frac{\partial}{\partial y}(D_2 W_n) = W_n. \tag{37}$$

Here we allow for network diffusion, with diffusion coefficient ϵ, but our goal is to solve the system in the limit $\epsilon \to 0$.

The distance between the two plates is taken to be L; hence the domain of the problem consists of an infinite strip $(-\infty < x < \infty) \times (0 < y < L)$. The boundary conditions are $D_1 = D_2 = 0$ and $\epsilon \frac{\partial \theta_n}{\partial y} = \theta_n W_n$ at $y = 0, L$, implying that there is neither network displacement nor network flux at the boundary.

We can simplify these equations substantially. Integrating (35) and solving for the vertical velocity of the network, we find

$$W_n = \epsilon \frac{\frac{\partial \theta_n}{\partial y} + c_1}{\theta_n}, \tag{38}$$

which, combined with (34), yields

$$(1 - \theta_n) W_s = c_2. \tag{39}$$

These can be used in (33) to solve for the $\frac{\partial P}{\partial y}$ as

$$\frac{\partial P}{\partial y} = h_f \theta_n \left(W_n - \frac{c_2}{1 - \theta_n} \right). \tag{40}$$

The boundary conditions imply that $c_1 = c_2 = 0$; hence $W_s = 0$ and $\frac{\partial P}{\partial y} = h_f \epsilon \frac{\partial \theta_n}{\partial y}$. Because $V_n = 0$ at steady state, and because the equations are independent of x, $\frac{\partial P}{\partial x}$ is independent of x. That is, (32) implies that $P = Gx + \hat{P}(y)$.

Integrating (30) and solving for $\frac{\partial D_1}{\partial y}$ yields

$$\frac{\partial D_1}{\partial y} = \frac{Gy + a}{\gamma \theta_n}. \tag{41}$$

We specify a by assuming that the steady-state profiles are symmetric about the center line, $y = \frac{1}{2}$. We also relate the vertical displacements to the network volume fraction using the Jacobian of the transformation

$$\hat{\theta}_n = \theta_n \left(1 + \frac{\partial D_2}{\partial y} \right),$$

where $\hat{\theta}_n$ is the original homogeneous unstressed distribution of the network.

Finally, (31) reduces to an ordinary differential equation (ODE) relating the volume fraction of the network to y and parameters G, γ, h_f, etc.:

$$\begin{aligned} &\epsilon \eta \frac{d}{dy} \left(\theta_n \frac{d}{dy} \left(\frac{\frac{d\theta_n}{dy}}{\theta_n} \right) \right) - \epsilon h_f \frac{d\theta_n}{dy} \\ &- \frac{d\Psi}{dy} + \gamma \frac{d}{dy} \left(\theta_n \left(\left(\frac{Gy - GL/2}{\gamma \theta_n} \right)^2 + \left(\frac{\hat{\theta}_n}{\theta_n} \right)^2 - 1 \right) \right) = 0. \end{aligned} \tag{42}$$

We nondimensionalize (42) by defining the nondimensional variable $y^* = \frac{y}{L}$ and the nondimensional parameters $\epsilon^* = \frac{\eta}{L^2\gamma}\epsilon$ and $h_f^* = \frac{L^2 h_f}{\eta}$, $G^* = \frac{L}{\gamma}G$; substituting these into (42); and dropping the $*$-notation. Integrating once yields

(43)
$$\epsilon\left(\theta_n \frac{d}{dy}\left(\frac{\frac{d\theta_n}{dy}}{\theta_n}\right)\right) - \epsilon h_f \theta_n - \frac{1}{\gamma}\Psi(\theta_n) + \left(\theta_n\left(G^2\left(\frac{y-1/2}{\theta_n}\right)^2 + \left(\frac{\hat{\theta}_n}{\theta_n}\right)^2 - 1\right)\right) = k,$$

which must be solved subject to the constraint that mass is conserved,

$$\int_0^1 \theta_n \, dy = \hat{\theta}_n. \tag{44}$$

Although simpler than the original system, there remains quite a lot of interesting structure in (43). In particular, (43) is a second order ODE which is singular in the limit $\epsilon \to 0$. In the following section, we describe the singular perturbation analysis of this problem, revealing the existence of channels, i.e., solutions with sharp interior transition layers.

3.1. The channeling bifurcation. In this section we analyze the bifurcation structure of channels by examining the solutions of (43) in the singular limit $\epsilon \to 0$. We assume that the initially uniform gel at $\hat{\theta}_n$ is stable so that $\Psi'(\hat{\theta}_n) > 0$.

The reduced equation ($\epsilon = 0$) is an algebraic equation relating the network volume fraction to the location between the plates. The steady state network profile is given by the solution of the algebraic equation

$$H(y, \theta_n) = G^2\left(y - \frac{1}{2}\right)^2 + h(\theta_n) = 0, \tag{45}$$

where

$$h(\theta_n) = -\theta_n\Psi(\theta_n) + \theta_n\Psi(\hat{\theta}_n) + \hat{\theta}_n^2 - \theta_n^2 - k\theta_n, \tag{46}$$

and

$$\Psi(\theta_n) = \kappa\theta_n^2(\theta_n - \theta_{ref}), \tag{47}$$

where $\kappa = \frac{\gamma_{os}}{\gamma}$ represents the strength of osmosis compared to the elastic modulus. Here k has been redefined so that $H(\hat{\theta}_n) = 0$ when $k = 0$. The solution profile $\theta_n(y)$ must also satisfy the integral constraint 44.

By virtue of the form of Ψ, the gel is capable of supporting an edge. In many hydrogels, the polymer network is of very low density and is highly charged [5]. This suggests that the strength of the osmotic pressure is large compared to the magnitude of the elastic modulus, so that κ is large.

This problem can be viewed as a nonlinear eigenvalue problem: "For each value of G determine the value(s) of k for which the solution of (45) satisfies the integral constraint (44)." However, it turns out that it is easier to view the problem as follows: "For each value of k find the value(s) of G for which the solution of (45) satisfies the integral constraint (44)." We explain below why this is the case.

First we make some observations about the function $h(\theta)$. Because $h(0) = \hat{\theta}_n^2 > 0$, and $h(\theta) < 0$ for large θ, $h(\theta)$ always has at least one positive and one negative root.

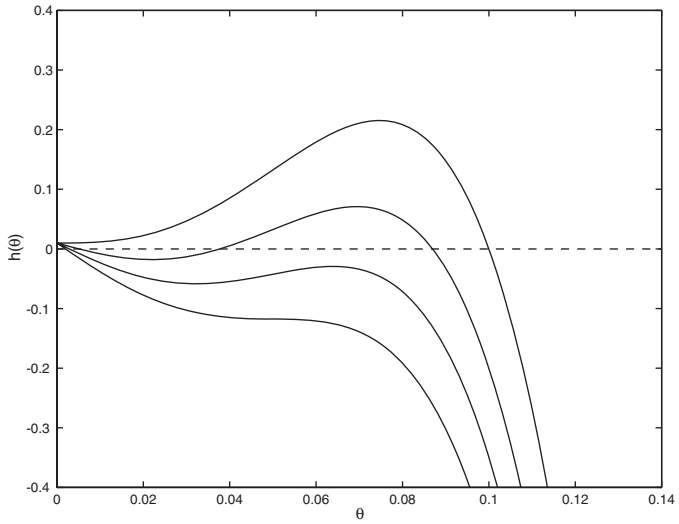


FIG. 1. *Plot of $h(\theta)$ as a function of θ for $k = 0, 2, 3.5, 5$. Other parameter values are $\kappa = 20{,}000$ and $\theta_{ref} = \hat{\theta}_n = 0.1$.*

Since $h(\theta)$ is a quartic polynomial in θ, there can be as many as three positive roots of $H(y, \theta_n) = 0$, depending on the value of k. To see this, in Figure 1 are shown four different plots of $h(\theta)$ for four values of $k = 0, 2, 3.5, 5$ (top to bottom). If $3\kappa\theta_{ref}^2 > 8$, the function $h(\theta)$ has two positive inflection points. Thus, if κ is sufficiently small, the function $h(\theta)$ is monotone for positive θ, whereas, if κ is sufficiently large, it is possible that $h(\theta)$ is nonmonotone.

We seek solutions of (45) that are of the form $\theta_n = \Theta_n(y)$. However, because $h(\theta)$ need not be monotone, such solutions do not always exist. However, it is always possible to write the solution implicitly for y as function of θ_n,

$$y = \frac{1}{2} \pm \frac{1}{G}\sqrt{-h(\theta_n)}. \tag{48}$$

Thus, one can visualize solutions by turning the plots in Figure 1 "on their side." If the resulting solution is single-valued, there is little more to do. If the resulting solution is multivalued, then one must determine which pieces of the multivalued function should be used to construct a single-valued function.

To construct admissible single-valued solutions from multivalued ones, we look for interior transition layers that connect different branches of the multivalued solution. Suppose that at $y = y_0$, $H(y_0, \theta_n) = 0$ has three positive roots, $\theta_- \le \theta_0 \le \theta_+$. We introduce an inner scaling of (43) defining $Y = \frac{y - y_0}{\epsilon^{1/2}}$. Substituting this into (43) and retaining the leading order terms in ϵ, we obtain

$$\frac{d}{dY}\left(\frac{\frac{d\theta_n}{dY}}{\theta_n}\right) + \frac{H(y_0, \theta_n)}{\theta_n^2} = 0. \tag{49}$$

With the change of variable $w = \ln(\theta_n)$, we can rewrite this as

$$\frac{d^2 w}{dY^2} + F(w, y_0) = 0, \tag{50}$$

where $F(w, y_0) = H(y_0, e^w)e^{-2w}$. Clearly, the function $F(w, y_0)$ has three roots, $w_\pm = \ln(\theta_\pm), w_0 = \ln(\theta_0)$. It is well known [4] that there is a solution to the inner-layer (50) that provides a transition between w_- and w_+ if

$$\int_{w_-}^{w_+} F(w, y_0)dw = 0. \tag{51}$$

Inverting the transformation yields an equivalent requirement in the variable θ_n. An interior layer providing a transition between θ_- and θ_+ can be fit at $y = y_0$ if

$$\int_{\theta_-}^{\theta_+} \frac{H(y_0, \theta_n)}{\theta_n^3} d\theta_n = 0. \tag{52}$$

There is another interior layer solution that can be used to construct solutions. If $y_0 = \frac{1}{2}$ and there are three positive roots of $H(\frac{1}{2}, \theta) = 0$, and if

$$\int_{\theta_-}^{\theta_+} \frac{H(\frac{1}{2}, \theta)}{\theta^3} d\theta < 0, \tag{53}$$

then there is a homoclinic orbit of (49) that approaches θ_+ asymptotically as $Y \to \pm\infty$ and has as its minimal value $\theta = \theta^*$, where $\theta_- < \theta^* < \theta_0$ and

$$\int_{\theta^*}^{\theta_+} \frac{H(\frac{1}{2}, \theta)}{\theta^3} d\theta = 0. \tag{54}$$

The first integral for this trajectory is

$$\frac{1}{2}\left(\frac{d\theta_n}{dY}\right)^2 - \theta_n^2 \int_{\theta_n}^{\theta_+} \frac{H(\frac{1}{2}, \theta)}{\theta^3} d\theta = 0. \tag{55}$$

Now we use this information to construct all the possible single-valued solutions. To do this we pick a value of k, determine the possible single-valued solutions, and then find the appropriate value of G (and y_0 if any) for this solution. There are three different types of solutions.

If $h(\theta) = 0$ has only one positive root and if and $h(\theta)$ is monotone decreasing for θ larger than this root, then the solution of $H(y, \theta_n) = 0$ is unique, for any value of G, as seen in Figure 1 for the curves with $k = 2$ and for $k = 5$. In Figure 2 the solution profiles $\theta_n(y)$ for $k = 5$ are shown for three different values of G.

Since G acts as a y-axis scale factor for these profiles, it is apparent that $\int_0^1 \theta_n(y)dy$ is a monotone decreasing function of G. Thus, there is a unique value of G for which $\int_0^1 \theta_n(y)dy = \hat{\theta}_n$. For the profiles shown in Figure 2, this unique value of G is 4.022.

Similarly, for small values of k, unique solutions can be obtained. For example, Figure 3 shows the solution profile for $k = 2$. Again, since the y-axis for this profile is scaled by G, the unique value of G for which $\int_0^1 \theta_n(y)dy = \hat{\theta}_n$ is easily determined. For the profile in Figure 3, this value is $G = 1.74$.

If the function $h(\theta_n)$ is not monotone decreasing, then there is the possibility of nonunique solutions of $H(y, \theta_n) = 0$. If a (positive) level x can be found so that

$$\int_{\theta_-}^{\theta_+} \frac{x + h(\theta_n)}{\theta_n^3} d\theta_n = 0, \tag{56}$$

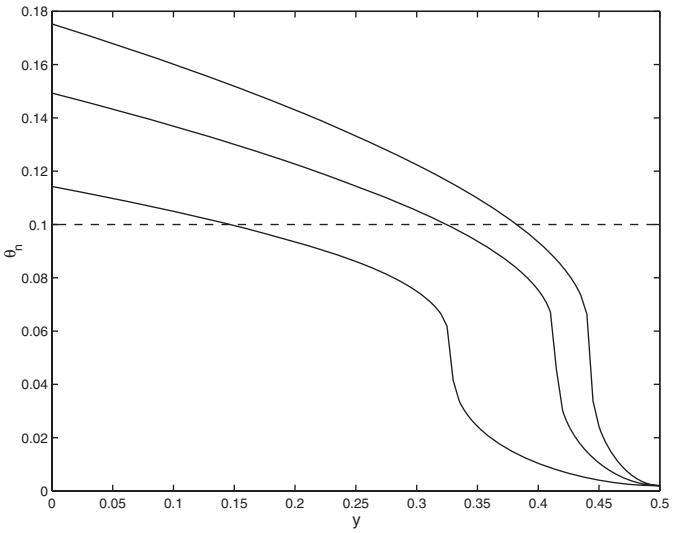


FIG. 2. *Plot of* $\theta_n(y)$ *for* $k = 5$ *and* $G = 2, 4.022, 6$. *Other parameter values are* $\kappa = 20{,}000$ *and* $\theta_{ref} = \hat{\theta}_n = 0.1$.

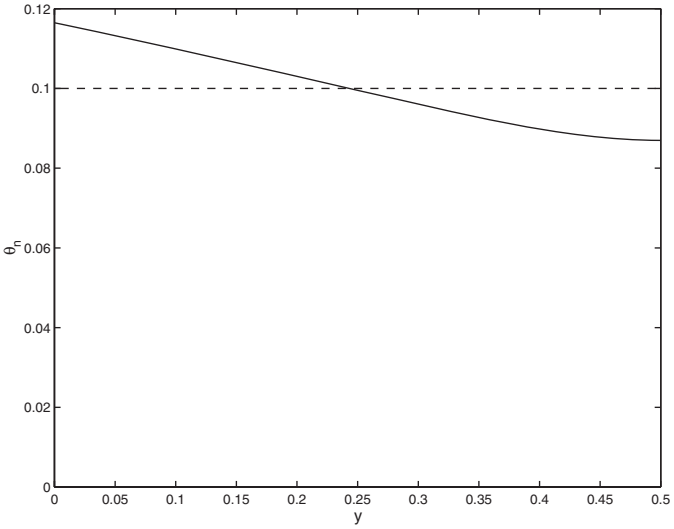


FIG. 3. *Plot of* $\theta_n(y)$ *for* $k = 2$ *and* $G = 1.74$. *Other parameter values are* $\kappa = 20{,}000$ *and* $\theta_{ref} = \hat{\theta}_n = 0.1$.

then a boundary layer can be inserted into the profile at $y_0 = \frac{1}{2} \pm \frac{\sqrt{x}}{G}$, and this boundary layer can be used to connect the largest solution of $H(y_0, \theta_n) = 0$ with the smallest. A plot of a profile that results is shown in Figure 4.

Notice that for this value of k (=2), there are three possible solution profiles, one with no interior layer (shown in Figure 3), one with a boundary layer (shown in Figure 4), and one with a symmetric boundary layer located at $y_0 = \frac{1}{2}$. In the limit that $\epsilon \to 0$, the third of these looks identical to the profile shown in Figure 3, with the exception that θ_n is discontinuous at $y = \frac{1}{2}$, with $\theta_n(\frac{1}{2}) = \theta^*$ defined in (54). The

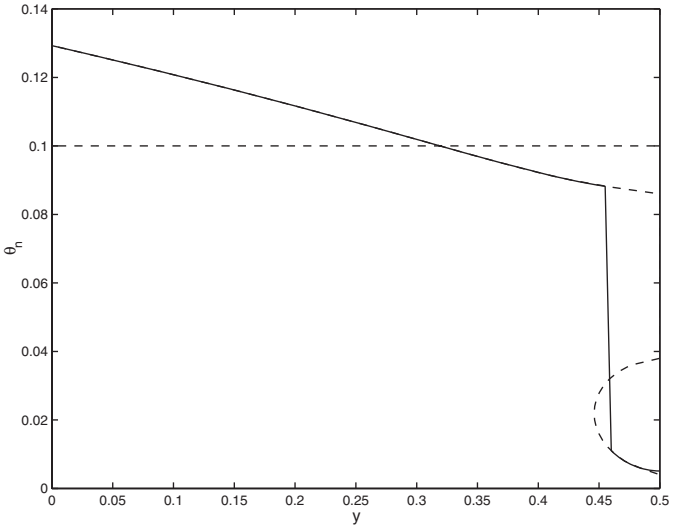


FIG. 4. *Plot of $\theta_n(y)$ for $k = 2$ and $G = 2.47$ with a boundary layer inserted at $y_0 = 0.46$. The dashed curves show all possible solutions of $H(y, \theta_n) = 0$. Other parameter values are $\kappa = 20{,}000$ and $\theta_{ref} = \hat{\theta}_n = 0.1$.*

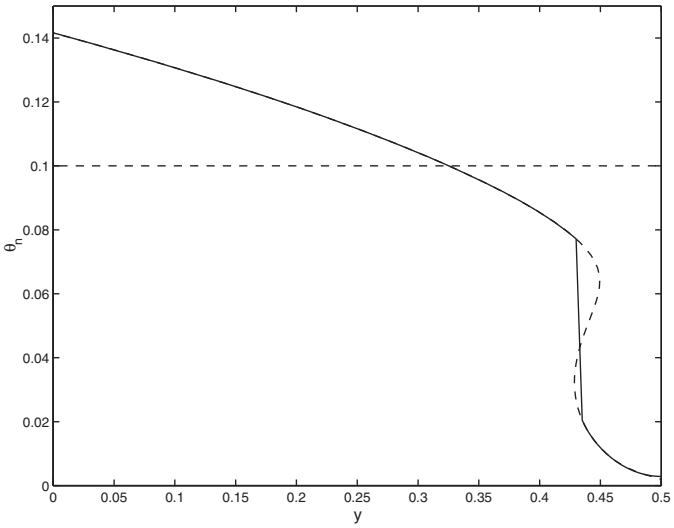


FIG. 5. *Plot of $\theta_n(y)$ for $k = 3.5$ and $G = 3.39$ with a boundary layer inserted at $y_0 = 0.43$. The dashed curves show all possible solutions of $H(y, \theta_n) = 0$. Other parameter values are $\kappa = 20{,}000$ and $\theta_{ref} = \hat{\theta}_n = 0.1$.*

value of G for the first and third solution profiles is the same, but it differs from the value of G for the second transition-layer profile. For some values of k, the boundary layer profile is the only possible solution. A profile of this type occurs for $k = 3.5$ and is shown in Figure 5.

In this way, for each value of k we determine all possible solutions and their corresponding values of G. A plot of the relationship between k and G is shown in Figure 6. Here we see two curves. The lower curve that extends from $k = 0$ to

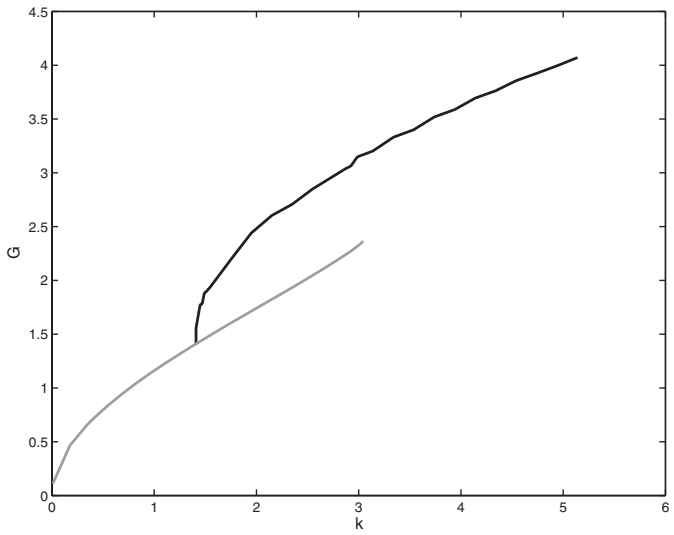


FIG. 6. *Plot of G versus k for the solutions of the gel-flow problem. For this plot, $\kappa = 20{,}000$ and $\theta_{ref} = \hat{\theta}_n = 0.1$.*

$k = 2.9$ corresponds to solutions like those shown in Figure 3 , with no boundary layer. The upper curve that extends from the lower curve at about $k = 1.4$ (with $y_0 = \frac{1}{2}$) corresponds to channeled solutions, with a boundary layer as shown in Figures 4 and 5 for $k < 5$. These solutions merge smoothly into non–boundary layer solutions, such as those shown in Figure 2, as k increases.

The nature of the bifurcation structure of these solutions is not apparent from Figure 6. This is because, for values of k larger than the merger point, the lower branch corresponds to two different solutions. The easier way to visualize this difference is seen in Figure 7, where $\theta_n(\frac{1}{2})$ is plotted as a function of G. Here, the upper solution branch corresponds to those solutions with no boundary layers, the lower solution branch corresponds to those with interior transition layers, and the middle branch (shown dashed) gives the solutions with a symmetric boundary layer at $y = \frac{1}{2}$. In the limit $\epsilon \to 0$, this boundary layer has no thickness and so has no influence on the integral of θ_n. Here we see that the solution is an S-shaped curve, and the bifurcations are via limit points.

The physically significant feature of these curves is that for some values of G there are two physically realizable solutions, a boundary layer, or channeled, solution and a nonchanneled solution. The solution with a boundary layer at $y = \frac{1}{2}$ is transitional between the two and is interesting for mathematical reasons but is unstable and hence not physically realized. Thus, the solution of the gel-flow problem is not unique and exhibits hysteretic behavior, with a hysteresis loop between channeled and nonchanneled solutions, governed by the pressure gradient G.

The behavior of the fluid flow through these two different solution types is understandably different, as the channeled solution permits a higher flux for the same cost. This is illustrated by Figure 8, where the flux of solvent,

$$J = \int_0^1 V_s dy = \frac{1}{h_f} \int_0^1 \frac{1 - \theta_n}{\theta_n} dy, \tag{57}$$

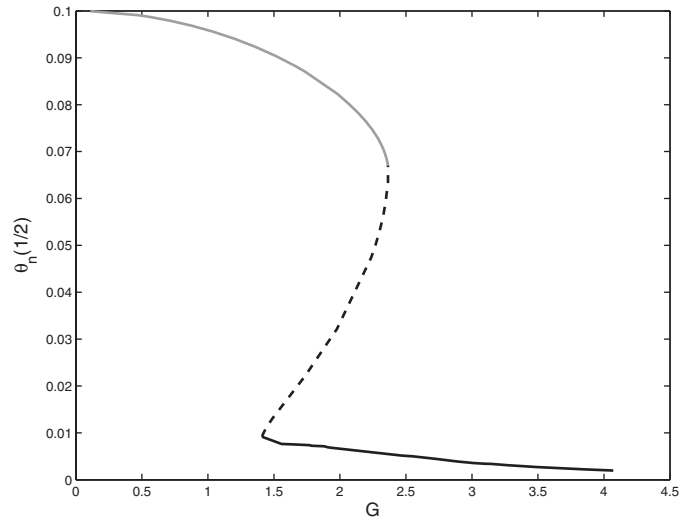


FIG. 7. *Plot of* $\theta_n(\frac{1}{2})$ *versus* G *for the solutions of the gel-flow problem. For this plot,* $\kappa = 20{,}000$ *and* $\theta_{ref} = \hat{\theta}_n = 0.1$.

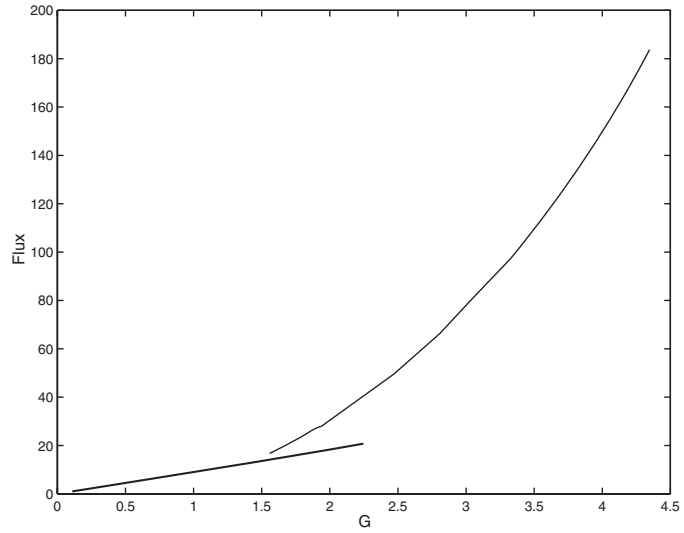


FIG. 8. *Plot of solvent flux as a function of* G. *For this plot,* $\kappa = 20{,}000$ *and* $\theta_{ref} = \hat{\theta}_n = 0.1$.

is plotted as a function of G for the two different solution types. Not surprisingly, if two solutions are possible for the same value of G, the boundary layer solution permits a larger solvent flux than the non–boundary layer solution.

4. Discussion. From this analysis we can deduce the physical mechanism that underlies the formation of channels in a gel. If the osmotic force is sufficiently strong compared to the elastic restoring force, then under a sufficiently high pressure gradient, it is energetically favorable to compress the gel near the wall and swell the gel in the interior, thereby forming a low-resistance channel.

This same conclusion is correct for all gels for which there are two stable gel

concentrations. That is, if $\Psi(\theta_n)$ is such that $\Psi'(\theta_n) < 0$ for $0 \leq \theta_* < \theta_n < \theta^* < 1$ and is positive elsewhere, then if the uniform gel distribution has $\hat{\theta}_n > \theta^*$ and if the osmotic force is sufficiently strong compared to the elastic force, channels will form under sufficiently high pressure gradient flows. This follows from the analysis of the previous section, which relied entirely upon the generic "cubic" shape of the function $\Psi(\theta_n)$ and not on its details. Any function $\Psi(\theta_n)$ with similar structure will lead to the same bifurcation channeling behavior.

REFERENCES

[1] R. B. Bird, R. C. Armstrong, and O. Hassager, *Dynamics of Polymeric Liquids*, Vol. 1, John Wiley and Sons, New York, 1987.

[2] N. G. Cogan and J. P. Keener, *The role of the biofilm matrix in structural development*, Math. Medicine and Biol., 21 (2004), pp. 147–166.

[3] X. He and M. Dembo, *On the mechanics of the first cleavage division of the sea urchin egg*, Exper. Cell Res., 233 (1997), pp. 252–273.

[4] J. P. Keener *Principles of Applied Mathematics: Transformation and Approximation*, Addison–Wesley, Reading, MA, 1988.

[5] A. Kumar and R. K. Gupta, *Fundamentals of Polymers*, chaps. 8 and 9, McGraw–Hill, New York, 1998.

[6] S. R. Lubkin and T. Jackson, *Multiphase mechanics of capsule formation in tumors*, J. Biomech. Eng., 124 (2002), pp. 237–243.

[7] S. T. Milner, *Dynamical theory of concentration fluctuations in polymer-solutions under shear*, Phys. Rev. E, 48 (1993), pp. 3674–3691.

[8] M. Strathmann, T. Griebe, and H.-C. Flemming, *Agarose hydrogels and EPS models*, Water Sci. Technol., 43 (2001), pp. 169–175.

[9] H. Tanaka, *Viscoelastic model of phase separation*, Phys. Rev. E, 56 (1997), pp. 4451–4462.

[10] H. Tanaka, *Viscoelastic phase separation*, J. Phys. Condens. Matter, 12 (2000), pp. R207–R264.

[11] T. Tomari and M. Doi, *Hysteresis and incubation in the dynamics of volume transitions of spherical gels*, Macromolecules, 28 (1995), pp. 8334–8343.

[12] J. Wingender, T. R. Neu, and H.-C. Flemming, *Microbial Extracellular Polymeric Substances. Characterization, Structure and Function*, Springer, New York, 1999.

[13] C. Wolgemuth, E. Hoiczyk, D. Kaiser, and G. Oster, *How myxobacteria glide*, Current Biol., 12 (2002), pp. 369–377.

[14] C. W. Wolgemuth, E. Hoiczyk, and G. Oster, *How gliding bacteria glide*, Biophys. J., 82 (2002), p. 1956.

SIAM J. Appl. Math.
Vol. 65, No. 6, pp. 1855–1869

FORCE DENSITY FUNCTION RELATIONSHIPS IN 2-D GRANULAR MEDIA*

ROBERT C. YOUNGQUIST†, PHILIP T. METZGER†, AND KELLY N. KILTS†

Abstract. An integral transform relationship is developed to convert between two important probability density functions (distributions) used in the study of contact forces in granular physics. Developing this transform has now made it possible to compare and relate various theoretical approaches with one another and with the experimental data, despite the fact that one may predict the Cartesian probability density and another the force magnitude probability density. Also, the transforms identify which functional forms are relevant to describing the probability density observed in nature, and so the modified Bessel function of the second kind has been identified as the relevant form for the Cartesian probability density corresponding to exponential forms in the force magnitude distribution. Furthermore, it is shown that this transform pair supplies a mathematical framework sufficient for describing the evolution of the force magnitude distribution under shearing. Apart from the choice of several coefficients, whose evolution of values must be explained in the physics, this framework successfully reproduces the features of the distribution that are taken to be an indicator of jamming and unjamming in a granular packing.

Key words. granular physics, probability density functions, Fourier transforms

AMS subject classifications. 60K40, 82D30

DOI. 10.1137/04061221X

1. Introduction.

A central topic within modern granular physics research is the study of intergranular force probability densities [1, 2, 3, 4, 5, 6]. The goal being to develop theory—from simplest assumptions—predicting the force density functions seen in simulations and experimental measurements. However, this goal is complicated by the differing forms for the force density functions presented in the literature, two of which are often treated as fundamental.

The first of these force density functions is dependent upon the magnitude and angle of the contact forces between grains. It predicts force chains, the onset of granular jamming [7], strain hardening [1], and fracture of the individual grains. It has several odd and unexplained features, such as a finite, but nonzero, probability at zero force, followed by a probability increasing to a peak near the average value of the force. It has attracted scientific attention because its exponential tail at high forces and power law at weak forces are reminiscent of the Maxwell–Boltzman distribution of velocities from statistical mechanics. However, its overall form is unlike any of the known statistical mechanics distributions, and there have been numerous attempts to derive it [8, 9, 10, 11, 12, 13].

The second type of distribution that has been treated as fundamental in the physics literature is a force density function dependent upon the Cartesian components of the contact force vectors. This type of distribution is appealing because it relates to

*Received by the editors July 25, 2004; accepted for publication (in revised form) January 19, 2005; published electronically August 3, 2005.

http://www.siam.org/journals/siap/65-6/61221.html

†The KSC Applied Physics Laboratory (YA-C3-E), Kennedy Space Center, FL 32899 (Robert.C.Youngquist@nasa.gov, Philip.T.Metzger@nasa.gov, Kelly.N.Kilts@nasa.gov).

the following conservation law: If external forces are ignored and if the arrangement of grains is static, then the total Cartesian force perpendicular to a plane cutting through the medium is conserved for any translation of the plane. Consequently, granular thermodynamic theories have been developed based on this conservation of the total Cartesian forces [14, 15].

The problem addressed by this paper is to clearly define these two types of force density functions and then to establish relationships between them. At first glance this seems straightforward, and it would be if in the literature these densities were handled as functions of two dimensions for two-dimensional media and three dimensions for three-dimensional media, but this is usually not the case. Probability force distributions are typically expressed using one of the following variables: the force magnitude, the force angle, or a selected Cartesian component. Because of this contraction to a single variable, pertinent information is not available, and converting between the two types of densities is not possible. Yet this conversion is important. Analytical theories usually favor one distribution or the other, and empirical investigations typically collect only one type of distribution. Without a well established conversion it is not possible to compare the competing theories and their data against one another.

The paper is organized as follows. Random variables are chosen such that the two two-dimensional probability force densities can be defined and the probability theory relations between them shown. In particular, an integral relation is developed whereby the two-dimensional polar force probability density can be converted to a Cartesian force density. From probability theory alone this relationship cannot be inverted, but the integral relation can be recast as a set of Fourier transforms. Doing this allows an inverse relation to be found such that if the one-dimensional Cartesian force density function is known for all rotations of the axes, then the two-dimensional polar force probability function can be found. This inversion allows, for the isotropic case, the two force distributions to be treated as a transform pair. The properties of this transform relationship are discussed and significant solutions provided. An example is provided along with data generated from a Monte Carlo process. Then, an example of a pair of force density functions for the anisotropic case is given and compared with published results.

2. Two-dimensional force probability density functions. Suppose that a large number of grains are placed randomly into a two-dimensional container and that the edges of the container push the grains against each other. These grains are not idealized and may be compressible, have arbitrary shape, be attracted or repulsed by each other, or have frictional contacts. The only restrictions on the grains are that they be static and that at each grain-to-grain contact a force vector be identifiable. Thus, the development presented here is applicable to a wide range of granular media and could be extended to other discrete, static media such as intertwined fibers, foams, glass, and emulsions. Also, the contact forces between the grains and walls may be included in the density functions provided below, as long as identifiable force vectors exist and both force vectors, the grain on the wall and the wall on the grain, are counted in the statistics.

At each grain-to-grain or grain-to-wall contact there are two force vectors, of equal magnitude and opposite direction, according to Newton's third law, as shown in Figure 2.1. Assign the random variable F, where $0 \leq F < \infty$, to the magnitude of these force vectors and the angle θ, where $0 \leq \theta < 2\pi$, to the angle between them and the x-axis. A two-dimensional "polar" force probability density can then be defined in terms of the random variables, F and θ, and expressed as $P_{F,\theta}(F, \theta)$, describing

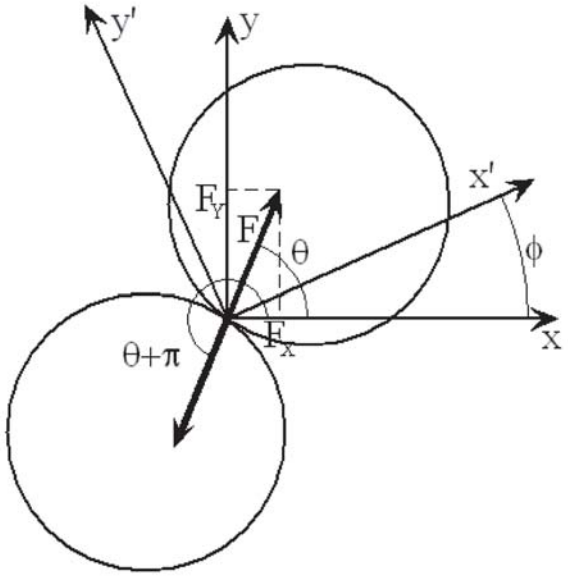


FIG. 2.1. *A sketch showing intergrain forces and the associated random variables.*

the probability of finding a granular contact force with magnitude F and angle θ. An immediate attribute of this function is that

$$P_{F,\theta}(F,\theta) = P_{F,\theta}(F,\theta+\pi), \tag{2.1}$$

a result of there being equal and opposite forces at each granular contact. (It is implied in the definition of $P_{F,\theta}(F,\theta)$ that the angle θ is modulo 2π.)

A one-dimensional force magnitude probability density function, $P_F(F)$, can be defined by integrating over all values of θ, as shown in (2.2) below. This equation also shows that the symmetry relation given by (2.1) allows the θ integration to occur over any contiguous arc of length π radians:

$$P_F(F) \equiv \int_0^{2\pi} P_{F,\theta}(F,\theta)d\theta = 2\int_0^{\pi} P_{F,\theta}(F,\theta)d\theta = 2\int_{\phi-\pi/2}^{\phi+\pi/2} P_{F,\theta}(F,\theta)\,\mathrm{d}\theta. \tag{2.2}$$

By integrating over all values of the force magnitude variable, a one-dimensional force angle probability density function, $P_\theta(\theta)$, is defined by

$$P_\theta(\theta) \equiv \int_0^{\infty} P_{F,\theta}(F,\theta)\,\mathrm{d}F. \tag{2.3}$$

From (2.1) and (2.3) the equal and opposite force result, i.e., $P_\theta(\theta) = P_\theta(\theta+\pi)$, can be shown. For frictionless grains (often studied in the physics literature) the direction of the force vector is normal to the contacting surfaces of the grains. In that special case the density function of (2.3) is identical to the distribution of contact angles, referred to in the literature as the fabric of the granular material [16].

Not unexpectedly, there exists an alternative pair of random variables, $F_x = F\cos(\theta)$ and $F_y = F\sin(\theta)$, that can be used to describe a Cartesian force probability density function, $P_{F_x,F_y}(F_x,F_y)$, which describes the probability of finding a granular

contact force with Cartesian components F_x and F_y, where $0 \leq F_x,\ F_y < \infty$. The polar and Cartesian force probability densities are related through their corresponding Jacobians, yielding

$$P_{F,\theta}(F,\theta) = P_{F_x,F_y}(F\cos(\theta), F\sin(\theta))F \tag{2.4}$$

and

$$P_{F_x,F_y}(F_x,F_y) = \frac{P_{F,\theta}((F_x^2+F_y^2)^{1/2}, \arctan(F_y/F_x))}{(F_x^2+F_y^2)^{1/2}}. \tag{2.5}$$

In (2.5), care should be taken to ensure that the arctangent returns an angle in the proper quadrant. Using (2.1) and (2.4), a Cartesian force density analogue for the existence of equal and opposite forces is found,

$$P_{F_x,F_y}(F_x,F_y) = P_{F_x,F_y}(-F_x,-F_y). \tag{2.6}$$

Also, by integrating over either of the two Cartesian random variables, a one-dimensional Cartesian force density can be defined. Without loss of generality, integrating over F_y yields $P_{F_x}(F_x)$, which, as shown in (2.7) below, can also be found by integrating the polar two-dimensional density function expression from (2.5),

$$\begin{aligned} P_{F_x}(F_x) &= \int_{-\infty}^{\infty} P_{F_x,F_y}(F_x,F_y)\,\mathrm{d}F_y \\ &= \int_{-\infty}^{\infty} \frac{P_{F,\theta}((F_x^2+F_y^2)^{1/2}, \arctan(F_y/F_x))}{(F_x^2+F_y^2)^{1/2}}\,\mathrm{d}F_y. \end{aligned} \tag{2.7}$$

Changing variables from F_y to θ via $\tan\theta = F_y/F_x$ yields the following integral relationship between the polar force density function and the one-dimensional Cartesian force density:

$$\begin{aligned} P_{F_x}(F_x) &= \left(\int_0^{\pi/2} + \int_{3\pi/2}^{2\pi}\right) P_{F,\theta}(F_x\sec\theta,\theta)\sec\theta\,\mathrm{d}\theta \qquad \text{if } F_x \geq 0, \\ P_{F_x}(F_x) &= \int_{\pi/2}^{3\pi/2} P_{F,\theta}(F_x\sec\theta,\theta)\sec\theta\,\mathrm{d}\theta \qquad \text{if } F_x \leq 0. \end{aligned} \tag{2.8}$$

Equations (2.6) and (2.7) imply that only one of the integrals in (2.8) needs to be calculated.

A more general form for the Cartesian density function can be defined that will prove useful in the analysis that follows. Figure 2.1 shows a coordinate system rotated by angle ϕ, yielding new random variables $F_{x'}$ and $F_{y'}$. Recalling that rotation does not stretch space, the Jacobian of the transformation between the random variables $F_{x'}, F_{y'}$ and F_x, F_y is unity, so their respective probability density functions are equal; i.e., $P_{F_{x'},F_{y'}}(F_{x'},F_{y'}) = P_{F_x,F_y}(F_x(F_{x'},F_{y'}), F_y(F_{x'},F_{y'}))$. Using this result, a one-dimensional Cartesian force density along the x'-axis, $P_{F_{x'}}(F_{x'})$, can be defined as

$$\begin{aligned} P_{F_{x'}}(F_{x'}) &= \int_{-\infty}^{\infty} P_{F_{x'},F_{y'}}(F_{x'},F_{y'})\,\mathrm{d}F_{y'} \\ &= \int_{-\infty}^{\infty} \frac{P_{F,\theta}((F_{x'}^2+F_{y'}^2)^{1/2}, \arctan(F_{y'}/F_{x'})+\phi)}{(F_{x'}^2+F_{y'}^2)^{1/2}}\,\mathrm{d}F_{y'}, \end{aligned} \tag{2.9}$$

similar to the result in (2.7). Now, changing variables from $F_{y'}$ to θ via $\arctan(F_{y'}/F_{x'})+\phi=\theta$ yields a result similar to that shown in (2.8) except that the additional ϕ angle appears in the integrand and in the limits of integration. The variable transformation yields

$$P_{F_{x'}}(F_{x'}) = \left(\int_{\phi}^{\phi+\pi/2} + \int_{\phi+3\pi/2}^{\phi+2\pi}\right) P_{F,\theta}(F_{x'}\sec(\theta-\phi),\theta)\sec(\theta-\phi)\,\mathrm{d}\theta \quad (\text{if } F_{x'}\geq 0)$$

$$= \int_{\phi+\pi/2}^{\phi+3\pi/2} P_{F,\theta}(F_{x'}\sec(\theta-\phi),\theta)\sec(\theta-\phi)\,\mathrm{d}\theta \qquad (\text{if } F_{x'}\leq 0). \tag{2.10}$$

Even though the density function $P_{F_{x'}}(F_{x'})$ is well defined by the above equations, it suffers from two shortcomings. First, it is an explicit function of the angle ϕ, and this should be reflected in the notation chosen. Second, if the angle ϕ is allowed to range from zero to 2π radians, there is an ambiguity in the choice of how to represent the domain of this function. Specifically, choosing a rotation angle ϕ and a random variable $F_{x'}$ is identical to rotating by $\phi+\pi$ radians and choosing a value for the random variable of $-F_{x'}$. We choose to resolve the second issue by allowing the angle ϕ to range from zero to 2π radians and defining a new random variable F_ϕ, which is equal to $F_{x'}$ but is always positive, i.e., $0\leq F_\phi<\infty$. The first issue above can then be resolved by adopting the notation $P_{F_\phi}(F_\phi,\phi)$ for the density function associated with this new random variable. The single variable subscript indicates that this is a one-dimensional density function, but the two-dimensional domain shows that it is an explicit function of the two variables F_ϕ and ϕ. Thus the function $P_{F_\phi}(F_\phi,\phi)$ has the same "polar" domain as the density function $P_{F,\theta}(F,\theta)$. It can be explicitly expressed using the first integral expression above—where $F_{x'}>0$—and merging the integrals by using the 2π periodic nature of the variable θ:

$$P_{F_\phi}(F_\phi,\phi) = 2\int_{\phi-\pi/2}^{\phi+\pi/2} P_{F,\theta}(F_\phi\sec(\theta-\phi),\theta)\sec(\theta-\phi)\,\mathrm{d}\theta, \tag{2.11}$$

where a factor of 2 has been added for normalization. Also, note that the function $P_{F_\phi}(F_\phi,\phi)$, in order to be a density function, must be normalized at each angle ϕ. In other words, the total number of forces does not change with the choice of ϕ, so the integral of $P_{F_\phi}(F_\phi,\phi)$ over all F_ϕ must always equal 1.

Equation (2.11) is a general integral relation allowing integration of the "polar" two-dimensional probability force density function to yield any desired Cartesian projection density function. This is a useful relationship within granular physics research in that it allows the calculation of a Cartesian density function from a force magnitude density function, but it is not a surprising result. Once the proper definitions are made, the derivation is straightforward. The more difficult result is to perform the inverse operation, namely, to find the polar form from the Cartesian form, but this is not possible within the realm of probability theory because the function $P_{F_\phi}(F_\phi,\phi)$ is not a two-dimensional probability density function. Even so, the inverse operation can be performed as demonstrated in the next section.

Finally, by using the variable transformation $\theta'=\theta+\pi$, it can be shown that

$$P_{F_\phi}(F_\phi,\phi+\pi) = P_{F_\phi}(F_\phi,\phi). \tag{2.12}$$

3. Fourier transform representation of force density integrals. In this section it will be shown that (2.11) above can be represented as a set of Fourier

transforms. Accomplishing this allows (2.11) to be inverted by utilizing the Fourier transform inversion properties. This approach is similar to the mathematics used in tomography [17], but the development presented here is distinct for two reasons. First, the symmetry relations of (2.1) and (2.3) provide simplification that does not occur in tomography. Second, the emphasis in tomography is on generating three-dimensional functions from a set of two-dimensional images, while in the present development the goal is to obtain two- or, in a future work, three-dimensional representations of the force density functions from one-dimensional Cartesian projections.

As a result of the symmetries obtained above in (2.1) and (2.12), we will require only restricted forms of the Fourier transforms. For example, we require only the Fourier cosine transform instead of the full exponential form. The Fourier cosine transform of a function $f(x)$ is expressed as

$$\mathcal{F}_c[f(x);u] = \sqrt{\frac{2}{\pi}} \int_0^\infty f(x)\cos(ux)\ \mathrm{d}x,$$

which has the same form as the inverse cosine transform [18]. The Fourier transform representations are written out in detail in order to resolve notational and normalization variations that appear in the literature. In addition to this one-dimensional transform, we will require a two-dimensional Fourier transform. In Cartesian form this is written as

$$\mathcal{F}_{2D}[f(x,y);(u,v)] = \frac{1}{2\pi} \int_{-\infty}^\infty \int_{-\infty}^\infty f(x,y)\exp\left(i(ux+vy)\right)\ \mathrm{d}x\ \mathrm{d}y, \tag{3.1}$$

but since the density functions are in polar form, this transform will need to be expressed in polar form also. Using the change of variables $x = F\cos(\theta)$, $y = F\sin(\theta)$, $u = G\cos(\phi)$, $v = G\sin(\phi)$, the following form for the two-dimensional, polar, Fourier transform is found,

$$\begin{aligned}
&\mathcal{F}_{2D}[f(F,\theta);(G,\phi)] \\
&= \frac{1}{2\pi} \int_0^\infty \int_0^{2\pi} f(F,\theta)\exp\left(iFG(\cos(\theta)\cos(\phi)+\sin(\theta)\sin(\phi))\right)F\ \mathrm{d}F\ \mathrm{d}\theta \\
&= \frac{1}{2\pi} \int_0^\infty \int_0^{2\pi} f(F,\theta)\exp\left(iFG\cos(\theta-\phi)\right)F\ \mathrm{d}F\ \mathrm{d}\theta,
\end{aligned}$$

but we will only be transforming functions where $f(F,\theta) = f(F,\theta+\pi)$. Using this symmetry simplifies this to a two-dimensional form of the Fourier cosine transform,

$$\mathcal{F}_{2D,c}[f(F,\theta);(G,\phi)] = \frac{1}{2\pi} \int_0^\infty \int_0^{2\pi} f(F,\theta)\cos(FG\cos(\theta-\phi))F\ \mathrm{d}F\ \mathrm{d}\theta. \tag{3.2}$$

Having established the above notation, now consider the following lemma which is the key result of this paper.

LEMMA 1. *The projected force density function is given by*

$$P_{F_\phi}(F_\phi,\phi) = 2\sqrt{2\pi}\mathcal{F}_c\left[\mathcal{F}_{2D,c}\left[\frac{P_{F,\theta}(F,\theta)}{F};(G,\phi)\right];F_\phi\right]. \tag{3.3}$$

Proof. Note that

$$
\begin{aligned}
&2\sqrt{2\pi}\mathcal{F}_c\left[\mathcal{F}_{2D,c}\left[\frac{P_{F,\theta}(F,\theta)}{F};(G,\phi)\right];F_\phi\right]\\
&\quad=2\sqrt{2\pi}\sqrt{\frac{2}{\pi}}\int_0^\infty\left[\frac{1}{2\pi}\int_0^\infty\int_0^{2\pi}P_{F,\theta}(F,\theta)\cos(FG\cos(\theta-\phi))\,\mathrm{d}F\,\mathrm{d}\theta\right]\cos(GF_\phi)\,\mathrm{d}G\\
&\quad=\frac{2}{\pi}\int_0^{2\pi}\int_0^\infty P_{F,\theta}(F,\theta)\left(\int_0^\infty\cos(FG\cos(\theta-\phi))\cos(GF_\phi)\,\mathrm{d}G\right)\,\mathrm{d}F\mathrm{d}\theta\\
&\quad=\int_0^{2\pi}\int_0^\infty P_{F,\theta}(F,\theta)\left(\delta(F\cos(\theta-\phi)-F_\phi)\right)\mathrm{d}F\mathrm{d}\theta.
\end{aligned}
$$

Changing the argument of the Dirac delta function, $\delta(x)$, is done using

$$\delta(F\cos(\theta-\phi)-F_\phi)=\frac{\delta(F-F_\phi\sec(\theta-\phi))}{|\cos(\theta-\phi)|}.$$

Thus integration with respect to F is well defined. The limits of integration with respect to θ can be compacted and shifted so that $\cos(\theta-\phi)$ is always positive, removing the need for the absolute value and introducing a factor of two. Then

$$
\begin{aligned}
&2\sqrt{2\pi}\mathcal{F}_c\left[\mathcal{F}_{2D,c}\left[\frac{P_{F,\theta}(F,\theta)}{F};(G,\phi)\right];F_\phi\right]\\
&=2\int_{\phi-\pi/2}^{\phi+\pi/2}P_{F,\theta}(F_\phi\sec(\theta-\phi),\theta)\sec(\theta-\phi)\,\mathrm{d}\theta\\
&=P_{F_\phi}(F_\phi,\phi).\qquad\square
\end{aligned}
$$

The benefit of this form is that (3.3) can be immediately inverted because each of the Fourier transforms is its own inverse, yielding the second key result of this paper, as follows.

LEMMA 2. *The two-dimensional polar probability force density may be expressed as a function of the projected Cartesian force density by*

$$P_{F,\theta}(F,\theta)=\frac{F}{2\sqrt{2\pi}}\,\mathcal{F}_{2D,c}[\mathcal{F}_c[P_{F_\phi}(F_\phi,\phi);(G,\phi)];(F,\theta)].\tag{3.4}$$

Equivalently,

$$
\begin{aligned}
&P_{F,\theta}(F,\theta)\\
&=\frac{F}{(2\pi)^2}\int_0^\infty\int_0^{2\pi}\int_0^\infty P_{F_\phi}(F_\phi,\phi)\cos(F_\phi G)\cos(FG\cos(\phi-\theta))G\,\mathrm{d}G\,\mathrm{d}\phi\,\mathrm{d}F_\phi.
\end{aligned}\tag{3.5}
$$

This integral can be integrated immediately with respect to the variable G, but this leads to an awkward functional form that is not amenable to solution using the standard look-up tables. Instead, we choose to leave it in this form, allowing the order of integration to be carried out on a case-by-case basis. This integral can be further integrated with respect to either F or θ to yield the single variable force density functions, demonstrating that knowledge of the projected Cartesian force density function can yield an explicit form for the force magnitude density.

Furthermore, for a frictionless packing, in which the force angles are the same as the contact angles, as discussed above, one may integrate out F from (3.5) according

to (2.3) to obtain the fabric of the packing. This result is striking because it begins with a knowledge of only $P_{F_\phi}(F_\phi, \phi)$, which is a set of contact force distributions that on the surface appear to have no contact angle information. This is the first time that force distributions, alone, have been directly related to the fabric of a packing, and this may be important to developing a theory of granular rheology in which changes in the fabric and the forces are coupled. Also, if the empirical studies can determine a simple ϕ-dependence for $P_{F_\phi}(F_\phi, \phi)$, then it may be possible to discern the fabric of a frictionless packing simply by sampling P_{F_ϕ} at only a few orientations of ϕ, or maybe only along the principle stress axes.

4. Force density integrals for isotropic material. In this section the two integral equations derived above, (2.11) and (3.5), are simplified for the isotropic case, yielding a pair of integral transform equations. Some of the properties of this transform pair are presented, and a list of useful solutions is shown. We start with the following definition.

DEFINITION 3. *An isotropic medium is one in which* $P_{F,\theta}(F, \theta) = P_F(F)/(2\pi)$.

Thus, an isotropic medium implies a force density with no angular dependence. As an immediate consequence of this definition, (2.11) can be simplified to show that an isotropic material also has no ϕ-dependence on its Cartesian projected force density,

$$P_{F_\phi}(F_\phi, \phi) = \frac{1}{\pi} \int_{-\pi/2}^{\pi/2} P_F(F_\phi \sec(\theta)) \sec(\theta)\, \mathrm{d}\theta. \tag{4.1}$$

Thus, for the rest of this section we will use the notation $P_{F_\phi}(F_\phi)$ for the Cartesian projected force density, since the function is no longer dependent upon the angle ϕ.

Equation (4.1) can be put into a more useful form by changing variables from the angle θ back to the force magnitude using $F = F_\phi \sec(\theta)$. This yields the integral equation

$$P_{F_\phi}(F_\phi) = \frac{2}{\pi} \int_{F_\phi}^{\infty} \frac{P_F(F)}{(F^2 - F_\phi^2)^{1/2}}\, \mathrm{d}F. \tag{4.2}$$

Equation (3.5) can be simplified as well. Since neither of the density functions has angular dependence, the ϕ integration on the right-hand side can be performed, yielding the result

$$P_F(F) = F \int_0^{\infty} \int_0^{\infty} P_{F_\phi}(F_\phi) \cos(F_\phi G) J_0(FG) G\, \mathrm{d}F_\phi\, \mathrm{d}G, \tag{4.3}$$

where J_0 is the zero-order Bessel function [18, eqn. 3.715.18], where we choose to perform the integration with respect to F_ϕ before G.

Equations (4.2) and (4.3) are inverse transform relations relating the force magnitude probability density, $P_F(F)$, to the projected Cartesian force probability density, $P_{F_\phi}(F_\phi)$, for isotropic two-dimensional granular materials. Such materials are of current theoretical and experimental interest, and these equations apply to any such material as long as an identifiable force can be assigned to the grain-to-grain, and, if included, grain-to-wall contacts. Consequently, a wide variety of force distributions are expected to result from current research, making it beneficial to discuss some of the properties of this transform pair as well as to present some of the more significant solution pairs. In the discussion below we will use the symbol $\leftrightarrow$ to link transform pairs, and we will not always normalize the pairs.

Equations (4.2) and (4.3) are clearly linear in that if $P_{F_\phi}(F_\phi) \leftrightarrow P_F(F)$ and $R_{F_\phi}(F_\phi) \leftrightarrow R_F(F)$ are two sets of solutions, then

$$aP_{F_\phi}(F_\phi) + bR_{F_\phi}(F_\phi) \leftrightarrow aP_F(F) + bR_F(F) \tag{4.4}$$

is a solution pair, where a and b are arbitrary constants. Another method for generating new solutions is to note that if $P_{F_\phi}(F_\phi) \leftrightarrow P_F(F)$ is a solution pair, then

$$F_\phi \frac{\partial P_{F_\phi}(F_\phi)}{\partial F_\phi} \leftrightarrow F \frac{\partial P_F(F)}{\partial F} \tag{4.5}$$

is also a solution pair, as a result of properties of the transform. By combining this result with the linearity result, it can be shown that

$$F_\phi^2 \frac{\partial^2 P_{F_\phi}(F_\phi)}{\partial F_\phi^2} \leftrightarrow F^2 \frac{\partial^2 P_F(F)}{\partial F^2} \tag{4.6}$$

is also a solution pair and more generally that

$$F_\phi^n \frac{\partial^n P_{F_\phi}(F_\phi)}{\partial F_\phi^n} \leftrightarrow F^n \frac{\partial^n P_F(F)}{\partial F^n} \tag{4.7}$$

is a solution pair. This result is very useful for obtaining sets of similar solutions when trying to fit experimental or simulation data.

A set of Bessel function integral equations [18, eqns. 6.592.10, 6.592.12–15], provides a method for obtaining useful solution pairs. Let Z_ν represent any of the Bessel functions, first kind J_ν, second kind Y_ν, third kind H_ν, or modified second kind K_ν. Then after appropriate changes of variable the referenced integral equations can be put into the form of (4.2), yielding the following transform pair:

$$\sqrt{\frac{2}{\pi\alpha}} \frac{Z_{\nu-1/2}(\alpha F_\phi)}{F_\phi^{\nu-1/2}} \leftrightarrow \frac{Z_\nu(\alpha F)}{F^{\nu-1}}. \tag{4.8}$$

Using this result and the differential generation result of (4.7), the solution pairs listed in Table 4.1 can be found. Not unexpectedly, since the transforms derived above are between polar and Cartesian representations, the solution pairs are often a Cartesian function (i.e., a sine, cosine, or exponential) and a polar function (i.e., a Bessel function). Probability densities are positive functions with finite total integrals, so the exponential and modified Bessel function pairs are especially useful. Since the transform relationships are linear, it is worthwhile to show the cosine and sine solutions (both as $P_F(F)$ and $P_{F_\phi}(F_\phi)$) so that, if desired, the Fourier components of a solution can be considered.

Other solution pairs to (4.2) and (4.3) exist and can be found in the standard integral tables, but many of them cannot be normalized. Table 4.2 shows five normalized solution pairs involving modified Bessel functions and Gaussians, which may be useful in granular physics applications.

To close this section a normalized pair of solutions is shown that has been fit to the results of a discrete element model (DEM) of a granular packing. DEM is a well-established technique which solves Newton's laws numerically grain-by-grain. It was implemented in the commercially available software package "Particle Flow Code in Two Dimensions" (PFC2D) by HCItasca (http://www.hcitasca.com/pfc.html). In

TABLE 4.1
Solution pairs for Cartesian and force magnitude functions in isotropic packings generated from (4.8).

	$\mathbf{P}_{\mathbf{F}_\phi}(\mathbf{F}_\phi)$	$\mathbf{P}_{\mathbf{F}}(\mathbf{F})$
1.	$\frac{2\alpha}{\pi}K_0(\alpha F_\phi)$	$\alpha\exp(-\alpha F)$
2.	$\frac{2\alpha^2}{\pi}F_\phi K_1(\alpha F_\phi)$	$\alpha^2 F\exp(-\alpha F)$
3.	$(-F_\phi)^n\frac{2\alpha}{\pi}\frac{\partial^n}{\partial F_\phi^n}(K_0(\alpha F_\phi))$	$\alpha^{n+1}F^n\exp(-\alpha F)$
4.	$\frac{2\alpha^3}{\pi}F_\phi^2 K_2(\alpha F_\phi)$	$\alpha^2 F(1+\alpha F)\exp(-\alpha F)$
5.	$-Y_0(\alpha F_\phi)$	$\cos(\alpha F)$
6.	$J_0(\alpha F_\phi)$	$\sin(\alpha F)$
7.	$\alpha\exp(-\alpha F_\phi)$	$\alpha^2 F K_0(\alpha F)$
8.	$\alpha^2 F_\phi\exp(-\alpha F_\phi)$	$\alpha^2 F(\alpha F K_1(\alpha F)-K_0(\alpha F))$
9.	$\alpha^{n+1}F_\phi^n\exp(-\alpha F_\phi)$	$\alpha^2(-F)^n\frac{\partial^n}{\partial F^n}(FK_0(\alpha F))$
10.	$\frac{2}{\pi}\cos(\alpha F_\phi)$	$\alpha F J_0(\alpha F)$
11.	$\frac{2}{\pi}\sin(\alpha F_\phi)$	$\alpha F Y_0(\alpha F)$

TABLE 4.2
More solution pairs for Cartesian and force magnitude distributions in isotropic packings.

	$\mathbf{P}_{\mathbf{F}_\phi}(\mathbf{F}_\phi)$	$\mathbf{P}_{\mathbf{F}}(\mathbf{F})$
1.	$2\sqrt{\alpha/\pi}\exp(-\alpha F_\phi^2)$	$2\alpha F\exp(-\alpha F^2)$
2.	$2\alpha F_\phi\exp(-\alpha F_\phi^2)$	$2\alpha F(2\alpha F^2-1)\exp(-\alpha F^2)$
3.	$2\sqrt{\alpha/\pi^3}\exp(-\alpha F_\phi^2/2)K_0(\alpha F_\phi^2/2)$	$2\sqrt{\alpha/\pi}\exp(-\alpha F^2)$
5.	$(2\alpha/\pi)K_0^2(\alpha F_\phi/2)$	$(2\alpha/\pi)K_0(\alpha F)$
6.	$(2\alpha^2 F_\phi/\pi^2)K_0(\alpha F_\phi/2)K_1(\alpha F_\phi/2)$	$(2\alpha^2 F/\pi)K_1(\alpha F)$

this case, the grains are round, hard, frictionless disks with a linear spring contact law implemented at the disk contacts. The disk diameters were distributed uniformly over the range 1.0 to 1.5 length units and were deposited randomly into a rectangular cell with flat, rigid, frictionless walls. Their diameters were then increased by rescaling until the disks jammed together across the span of the container. At each step of the incremental rescaling the grains were allowed to push each other around within the cell until every grain obtained static equilibrium. This incrementally dynamic

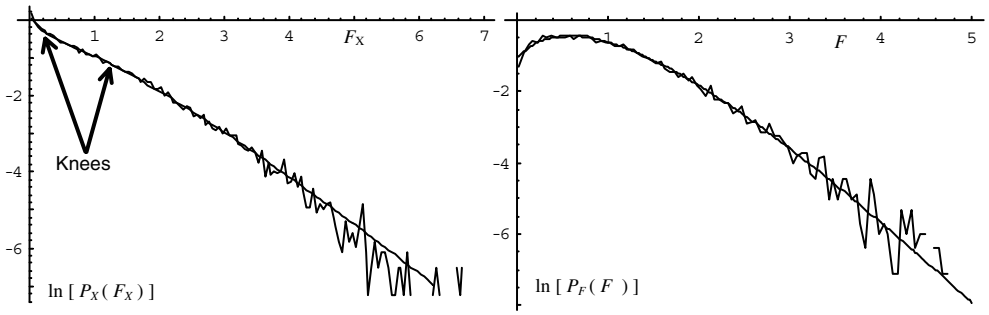


FIG. 4.1. *(Left) Graph of the Cartesian distribution from a DEM with an analytical fit. (Right) Graph of the force magnitude distribution from a DEM with an analytical fit.*

rescaling method naturally produces a disordered but statistically homogeneous and isotropic fabric within the bulk of the packing. Over 55,000 grain forces in the bulk of the packing were calculated, and the Cartesian force density function and the force magnitude density function were found. Grains within four grain diameters of the walls were excluded from these statistics. A fit to the Cartesian force density function was then found using a three-term modified Bessel function summation, as shown below. The solution pairs given above were then used to determine the force magnitude density function. The expansion coefficients can be chosen so that both distributions provide a good fit to the empirical data, as seen in Figure 4.1, thus demonstrating the success of the transformation developed in this paper. The plotted density functions are

$$P_{F_\phi}(F_\phi) = C\ \left[11F_\phi^2 K_2\left(\frac{\pi}{2}F_\phi\right) - 2F_\phi K_1\left(\frac{\pi}{2}F_\phi\right) + 3K_0\left(\frac{\pi}{2}F_\phi\right)\right] \tag{4.9}$$

and

$$P_F(F) = C\ \left[\left(\frac{11\pi}{2}\right)F^2 + (11-\pi)F + \frac{3\pi}{2}\right]\exp\left(-\frac{\pi}{2}F\right), \tag{4.10}$$

where the normalization constant $C = \pi^2/(132 - 4\pi + 3\pi^2)$.

It should be helpful to the field of granular research that the modified Bessel function of the second kind has been identified as the naturally occurring form for the Cartesian distribution, corresponding to the exponential forms of the polar distributions. The two "knees" in the curve that are visible in the Cartesian distribution of Figure 4.1 seem to be indicated in the Cartesian distributions of Bagi [14], as well, although this identification has not been previously made.

5. An example of an anisotropic solution pair. In most real world cases, granular media are subjected to greater total compressional force along one axis than the other, for example, in a gravity-dominated situation. Consequently, the anisotropic case is of interest, although it is significantly more complicated. Furthermore, the physics of jamming and unjamming have emerged as possibly the key concepts in granular media. The anisotropic case is relevant to this because shear stress (an aspect of anisotropy in the stress state) is one of the three ways to unjam granular media, as represented by the jamming phase diagram [19]. It has been proposed that the evolution of $P_F(F)$—from having a peak, as shown in Figure 4.1, to being monotonically decreasing—serves as an indicator of unjamming [7]. Numerical

simulations have indeed shown that this evolution occurs when the material is unjammed through stress anisotropy [1]. Therefore, an explanation for the evolution of $P_F(F)$ is central to the aims of granular physics. In this section a reasonable form is selected for the density function $P_{F_\phi}(F_\phi, \phi)$ for a sample anisotropic case and then the integrals evaluated to yield $P_F(F)$. This solution is discussed and compared against published data [1]. The purpose is to demonstrate that this mathematical framework is a *sufficient* framework to include the evolution of $P_F(F)$.

The reason we start with $P_{F_\phi}(F_\phi, \phi)$ instead of $P_{F,\theta}(F, \theta)$ in this demonstration is because the Cartesian distribution is the one associated with the force conservation law, and it is through that conservation law that the anisotropy is injected into the problem. It is well known that the normal components (diagonal elements) in the stress tensor scale according to $\sigma_{xx} = a + b\cos(2\phi)$ as the coordinate system is rotated through angle ϕ. Hence, the quantity of conserved force normal to the layers of a granular material will also scale according to this form when the layer is oriented at angle ϕ. The values of a and b are determined by the forces applied along the principal stress axes of the system. Based on the successful fits presented at the end of the previous section, we choose to let $P_{F_\phi}(F_\phi, \phi)$ be represented as a sum of the first three modified Bessel functions, but where this explicit angular dependence is added.

$$P_{F_\phi}(F_\phi, \phi) = \sum_{n=0}^{2} a_n \left(\frac{a-b}{a+b}\right)^{n-1} (a+b\cos(2\phi))^{n+1} F_\phi^n K_n((a+b\cos(2\phi))F_\phi). \tag{5.1}$$

The parameter b determines the amount of variation in force with angle, equaling zero for the isotropic case and approaching a for extreme anisotropy. Thus the force density is shifted towards higher forces along the y-axis with b nonzero. The $(a+b\cos(2\phi))^{n+1}$ factor has been included to normalize the distribution at every particular value of ϕ, as required from the discussion above. The $(a+b)/(a-b)$ factor is conjectural but it, or something similar to it, is necessary. Without this weighting, the resultant force density functions change only minimally with increasing anisotropy and do not agree with published literature. Including it yields results that correspond to dynamic simulations [1], as seen in Figure 5.1, but whose basis is unclear. The point is that the framework developed above allows choices to be made for the Cartesian form of the force density functions, which can then be converted to force magnitude or force angle density functions (i.e., fabric) and compared to published data.

The force magnitude density function is found by using this form for $P_{F_\phi}(F_\phi, \phi)$ in (3.5) and integrating with respect to θ (see (2.2)), yielding

$$\begin{aligned} P_F(F) = \frac{F}{(2\pi)^2} \sum_{n=0}^{2} a_n \left(\frac{a-b}{a+b}\right)^{n-1} \int_0^{2\pi}\int_0^{2\pi}\int_0^{\infty}\int_0^{\infty} (a+b\cos(2\phi))^{n+1} F_\phi^n \\ \cdot K_n((a+b\cos(2\phi))F_\phi) \cos(F_\phi G)\cos(FG\cos(\phi-\theta)) G \,\mathrm{d}G\,\mathrm{d}F_\phi\,\mathrm{d}\phi\,\mathrm{d}\theta. \end{aligned}$$

The integration with respect to θ can be performed immediately, yielding $2\pi J_0(FG)$, and the integration with respect to F_ϕ can be performed via [18, eqn. 6.699.12]. Then the integration with respect to G can be performed via [18, eqn. 6.565.4], yielding the partial result

$$\begin{aligned} P_F(F) = \sum_{n=0}^{2} \frac{a_n F^{n+1/2}}{4} \sqrt{\frac{2}{\pi}} \left(\frac{a-b}{a+b}\right)^{n-1} \\ \cdot \int_0^{2\pi} (a+b\cos(2\phi))^{n+3/2} K_{n-1/2}(F(a+b\cos(2\phi)))\,\mathrm{d}\phi. \end{aligned}$$

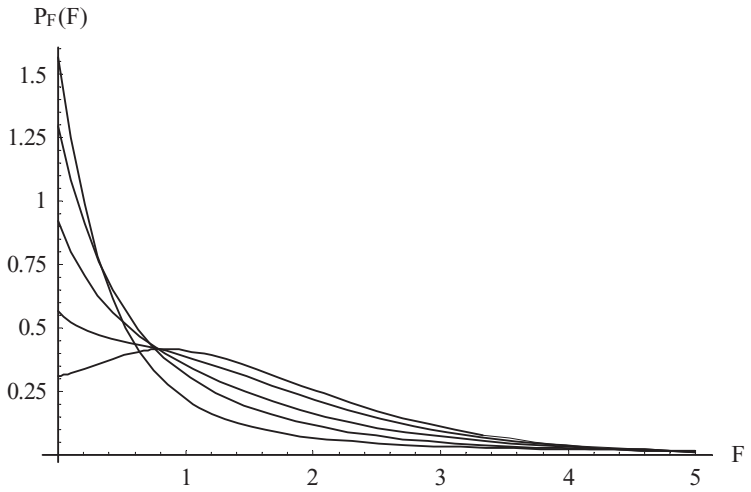


FIG. 5.1. *Evolution of force magnitude distribution with increasing anisotropy. The curve with the lowest probability density for $F = 0$ is the $b = 0$ isotropic case. The other curves, in order, correspond to $b = 0.3, 0.6, 0.9,$ and 1.57.*

Using the identities for the half-order modified Bessel functions, the integrations with respect to ϕ can be made, yielding the result

$$\begin{aligned} P_F(F) = \frac{\pi}{2}\ \exp(-aF)\Bigg(& a_0\left(\frac{a+b}{a-b}\right)(aI_0(bF) - bI_1(bF)) \\ &+ a_1((a^2+b^2)FI_0(bF) - (b+2abF)I_1(bF)) \\ &+ a_2\left(\frac{a-b}{a+b}\right)(F(a^2+2b^2+a^3F+3ab^2F)I_0(bF) \\ &- (3b+5abF+3a^2bF^2+b^3F^2)I_1(bF))\Bigg), \end{aligned} \tag{5.2}$$

where the exponential dependence on the force is expected from the isotropic case, but the Bessel function dependence on the parameter b is novel. ($I(x)$ is the modified Bessel function of the first kind.) Using the values $a_0 = 3\pi^2/4$, $a_1 = -\pi$, $a_2 = 11$, and $a = \pi/2$, Figure 5.1 shows a plot of (5.2) for various degrees of anisotropy (these plots have been normalized).

For $b = 0$, the isotropic case, the plot is identical to that shown in Figure 4.1, and (5.2) reduces (with the addition of a normalization term) to (4.10), but as b increases, the shape of the curve changes, slowly moving towards a pure exponential. This is in agreement with published simulation data, where the force magnitude density function evolves in a similar fashion with increasing anisotropy [1]. This demonstrates that the mathematical framework can produce this evolution naturally; nothing more exotic than the relative weighting of the Bessel terms need be invoked to produce it.

6. Summary and conclusions. It is possible within the straightforward techniques of probability theory to convert from $P_{F,\theta}(F,\theta)$ to $P_{F_\phi}(F_\phi,\phi)$. Unfortunately, those techniques cannot provide a conversion in the opposite direction. However, we may recognize that the conversion in the forward direction is equivalent to the composition of Fourier cosine transforms of the function. Since these transforms have their

own well-defined inverses, the conversion from $P_{F_\phi}(F_\phi, \phi)$ to $P_{F,\theta}(F, \theta)$ can likewise be expressed.

This inverse conversion is interesting for several reasons. First, it allows theoretical models that only predict Cartesian force distributions (such as the q model [15]) to be directly compared against the force magnitude distributions, which have been more important to granular physics. Second, the inverse conversion indicates a previously unrecognized relationship between the Cartesian force component distributions $P_{F_\phi}(F_\phi, \phi)$ and the fabric of the granular packing, which may be important in future theoretical developments. Third, for the special case of isotropic granular packings in which the ϕ and θ dependencies may be eliminated, the transform pair reduces to a simple form that can be solved for a wide range of functions. This indicates which functional forms for the Cartesian components correspond to particular functional forms for the force magnitudes. Since it is well known that the distribution of the latter has an exponential tail, the corresponding form for $P_{F_X}(F_X)$ ought to be modified Bessel functions of the second kind. Expansions in a series of such functions (of increasing order) display two characteristic "knees" when graphed, and indeed it turns out that such knees are observed in the empirical Cartesian distributions. Thus, the natural form for $P_{F_X}(F_X)$ appears to have been identified, and this should provide insight into the physical mechanisms that produce the distributions. Fourth, treating these modified Bessel functions with increasing anisotropic stress naturally produces an evolution of $P_F(F)$ that depends upon the choice of coefficients in the series expansion. Prior research has associated this evolution with the occurrence of jamming and unjamming in granular packings, and so the inverse transform indicates that jamming may be described as an increase in weighting of the zeroth-order modified Bessel function. This insight should be helpful to explain the physics of jamming and unjamming, which are important concepts in granular physics. Because this inverse conversion identifies these relationships and natural functional forms for granular force distributions, its derivation should be a helpful contribution in future research into the physics of granular jamming and unjamming.

Acknowledgment. The authors gratefully acknowledge William Miles of Stetson University for review and editing of this paper.

REFERENCES

[1] S. J. Antony, *Evolution of force distribution in three-dimensional granular media*, Phys. Rev. E (3), 63 (2002), article 011302.

[2] D. L. Blair, N. W. Muggenburg, A. H. Marshall, H. M. Jaeger, and S. R. Nagel, *Force distribution in three-dimensional granular assemblies: Effects of packing order and inter-particle friction*, Phys. Rev. E (3), 63 (2001), article 041304.

[3] D. M. Mueth, H. M. Jaeger, and S. R. Nagel, *Force distribution in a granular medium*, Phys. Rev. E (3), 57 (1998), pp. 3164–3169.

[4] F. Radjai, M. Jean, J. Moreau, and S. Roux, *Force distributions in dense two-dimensional granular systems*, Phys. Rev. Lett., 77 (1996), pp. 274–277.

[5] F. Radjai, S. Roux, and J. Moreau, *Contact forces in a granular packing*, Chaos, 9 (1999), pp. 544–550.

[6] O. Tsoungui, D. Vallet, and J. Charmet, *Experimental study of the force distributions inside 2-D granular systems*, NATO Sci. Ser. E Appl. Sci., 350 (1998), pp. 149–154.

[7] C. S. O'hern, S. A. Langer, A. J. Liu, and S. R. Nagel, *Force distributions near jamming and glass transitions*, Phys. Rev. Lett., 86 (2001), pp. 111–114.

[8] S. F. Edwards and D. V. Griner, *Statistical mechanics of granular materials: Stress propagation and distribution of contact forces*, Granular Matter, 4 (2003), pp. 147–153.

[9] N. P. Kruyt and L. Rothenburg, *Probability density functions of contact forces for cohesionless frictional granular materials*, Internat. J. Solids Structures, 39 (2002), pp. 571–583.

[10] N. P. Kruyt, *Contact forces in anisotropic frictional granular materials*, Internat. J. Solids Structures, 40 (2003), pp. 3537–3556.
[11] P. Metzger, *Comment on "Mechanical analog of temperature for the description of force distribution in static granular packings"*, Phys. Rev. E (5), 69 (2004), article 053301.
[12] P. Metzger, *Granular contact force density of states and entropy in a modified Edwards ensemble*, Phys. Rev. E (5), 70 (2004), article 051303.
[13] A. Ngan, *Mechanical analog of temperature for the description of force distribution in static granular packings*, Phys. Rev. E (3), 68 (2003), article 011301.
[14] K. Bagi, *Analysis of micro-variables through entropy principle*, in Powders & Grains 97, R. P. Behringer and J. T. Jenkins, eds., Balkema, Rotterdam, The Netherlands, 1997, pp. 251–254.
[15] S. N. Coppersmith, C. H. Liu, S. Majumdar, O. Narayan, and T. A. Witten, *Model for force fluctuations in bead packs*, Phys. Rev. E (3), 53 (1996), pp. 4673–4685.
[16] J. Rajchenbach, *Stress transmission through textured packings*, Phys. Rev. E (3), 63 (2001), article 041301.
[17] F. Natterer, *The Mathematics of Computerized Tomography*, John Wiley & Sons, New York, 1986.
[18] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed., Academic Press, New York, 2000.
[19] A. J. Liu and S. R. Nagel, *Nonlinear dynamics: Jamming is not just cool any more*, Nature, 396 (1998), pp. 21–22.

SIAM J. APPL. MATH.
Vol. 65, No. 6, pp. 1870–1887

A SCHOOL-ORIENTED, AGE-STRUCTURED EPIDEMIC MODEL*

VIGGO ANDREASEN† AND THOMAS FROMMELT†

Abstract. A model of childhood epidemics focusing on the impact of the school-year is presented. At the onset of the epidemic season, a new cohort of susceptible students enter the school, all other age-classes advance one year, while the oldest age-group leaves the mixing pool. If the susceptible pool is sufficiently large at the onset of the season, an epidemic will arise and run to its conclusion prior to the end of the school-year. The system is expressed in terms of a discrete dynamical system giving the changes in the age-dependent immunity structure on a year-to-year basis. If disease transmission is independent of age, the system settles at epidemics of constant size in each season. If disease transmission is age-dependent, more complicated dynamics may occur, including multiple stable states and chaos.

Key words. epidemic model, seasonal forcing, discrete map, closed epidemic, age-structure

AMS subject classifications. 92D30, 34A37

DOI. 10.1137/040610684

1. Introduction.

Disease transmission in schools played a central role in maintaining the regularly recurring epidemics of childhood diseases during the prevaccination era. In particular the summer break with low transmission, the annual infusion of a new cohort of (susceptible) first-year students, and the progression of students through the school system contributed substantially to the external forcing and the well-known two-year cycle in measles epidemics [42, 18, 38, 3]. The aim of this paper is to include this pulsed forcing in a mathematically tractable, age-structured epidemic model.

From an analytical viewpoint, the discontinuous nature of student admission and class-progression makes it rather awkward to incorporate the phenomenon into an epidemic model describing disease transmission dynamics in terms of the flow of hosts from the susceptible through the infected to the recovered class (SIR-model). Consequently the previous models of childhood epidemics with annual updates of the host structure have not been amenable to analytic methods. However, these "realistic age-structured" models (RAS-models) reflect quite accurately the prevaccination measles epidemics in England [42, 10, 23, 28], although recent results suggest that a detailed description of the seasonal variation in contact rates—rather than age-structure—is essential for matching the observed pattern of childhood diseases in England [16, 8, 18, 29, 41]. To simplify the analysis most mathematical studies of seasonally driven childhood epidemics have assumed that transmission strength varies sinusoidally with a period of one year, neglecting the details of the school-year [13, 31, 7, 33, 37]. Recently Billings and Schwartz [9] have provided a detailed description of the Poincaré return map associated with a forced SIR-model and showed how noise may lead to stochastic chaotic dynamics in such models.

*Received by the editors June 28, 2004; accepted for publication (in revised form) January 10, 2005; published electronically August 3, 2005. Supported by National Institutes of Health (NIH) award 1 RO1 GM607929 to S. A. Levin and by grant 51–00–0392 from the Danish Natural Science Research Council.

http://www.siam.org/journals/siap/65-6/61068.html

†Department of Mathematics and Physics, Roskilde University, DK-4000 Roskilde, Denmark (viggo@ruc.dk, frommet@ruc.dk).

In Schenzle's original model [42] as well as in other RAS-models, the host population is divided into cohorts, where new susceptible hosts are recruited continuously into the youngest age-class, age-classes in the school-age are updated annually, and hosts in adult age-classes continuously move through a fixed number of classes (one class in Schenzle's final formulation) and all mortality is concentrated in the adult age-classes. Thus the demographic structure is a discretization of the "von Forester model," which is often used in conjunction with epidemic models (see [26]). However, the original partial differential equation model as well as the age-group formulation of Tudor [44] describe a continuous flow through the age-classes, in contrast to the discrete annual progression through the school-system. See [25] for a recent review.

In this paper we will take an approach that allows us to account for the summer break and the annual infusion of new susceptible hosts into the mixing pool. In order to obtain an analytically tractable model, we will neglect the fine-scale variation in the school-year, assuming that the year can be divided into a (long) period of high disease transmission and a period where no transmission takes place. The basic idea is to completely separate the time scale of the epidemic from that of the school demographics by assuming that the duration of an epidemic is short compared to the length of the school-year. Thus each school-year starts with the introduction of a new cohort of susceptible hosts, one-year progression of all cohorts, and removal of the oldest age-class. We will assume that if the susceptible pool at the beginning of the school-year is large enough to support an epidemic, then an epidemic will occur that year, and this epidemic will run to its conclusion before the end of the school-year. This allows us to describe the epidemics only in terms of their size, i.e., the fraction of susceptible hosts that get infected during the season in question. Once the size of the epidemic is known, we can update the age-dependent immunity structure at the end of the school year, and we are ready to start the next season.

Clearly the model will not account for pathogen persistence over the summer period with low disease activity. Therefore the model cannot describe the exact inoculum at the onset of the subsequent season, and consequently the timing of the epidemic within the season cannot be determined. We shall return to these issues in the discussion.

Models of annual epidemics combined with interseasonal updating of the host population was first used by Gillespie [20] in his description of disease-induced natural selection in an (annual) insect population, and by May [32] and Dwyer et al. [15], who described disease-induced regulation of an insect population. Recently Andreasen and coworkers [6, 11] used the same modeling approach to describe the transmission dynamics of influenza under drift.

In the next section we derive a map connecting the age-distribution of susceptible hosts at the onset of one epidemic season to the age-distribution at the onset of the subsequent season. The dynamics of the season-to-season map can be quite complicated, and only a partial analysis is presented. In section 3 we show that the season-to-season model has a unique equilibrium. The case of age-independent susceptibility is studied in section 4, and in particular we shall show that when disease transmission is completely age-independent, the endemic equilibrium is locally asymptotically stable. In section 5 we give examples where the size of the annual epidemic changes in an irregular manner, with some years totally lacking an epidemic, indicating that the dynamics of the model may in fact be quite complicated.

2. The model. The separation of time scales naturally breaks the dynamics into two steps. The first step describes the demographic and school class dynamics

of the population, which is updated once a year at the beginning of the school-year, while the second step will describe disease transmission during a school-year in the closed population.

Reflecting the structure of school-classes, the population is divided into m cohorts, one for each year from introduction into the mixing-pool until removal from the pool, and we denote by N_k^T the number of hosts who are in the kth cohort during season T. At the onset of an epidemic season, M new (susceptible) hosts are introduced into the first cohort: $N_1^{T+1}(\textit{season-start}) = M$; cohorts $k = 1, \ldots, m-1$ move one age-class up: $N_k^T(\textit{season-end}) \to N_{k+1}^{T+1}(\textit{season-start})$, while the oldest cohort is removed. We will assume that the population is closed in the sense that all new hosts enter through the first cohort at the onset of a season and that hosts leave only by the removal of the oldest cohort at end of the season. Thus, provided that the population is at demographic equilibrium, all cohorts are of the same size throughout the epidemic season. This corresponds to an extreme case of type I mortality [34, 2]. Since the population size remains constant, it turns out to be convenient to describe the population structure in terms of n_k^T, the fraction of the host population that is in cohort k during season T. We note that $n_k = 1/m$ for all k.

To determine disease transmission dynamics during the season, we assume that the epidemic can be described by an SIR-epidemic model. Thus we subdivide the kth cohort of the host population into susceptible, infectious, and immune hosts and denote by s_k, i_k, and r_k the fraction of the total population in age-class k that is currently susceptible, infected, and recovered, respectively. During the season the dynamics become

$$\frac{ds_k}{dt} = -\sigma_k \Lambda s_k, \tag{2.1}$$

$$\frac{di_k}{dt} = \sigma_k \Lambda s_k - \nu i_k, \tag{2.2}$$

$$\frac{dr_k}{dt} = \nu i_k, \tag{2.3}$$

where $\Lambda = c \sum \tau_j i_j$ gives the force of infection that would be experienced by a totally susceptible host; ν denotes the rate of recovery, while c is the contact rate. To account for age-dependence in disease transmission, each age-class is assigned two factors: σ_k and τ_k, denoting, respectively, the susceptibility and infectivity of hosts in age-class k relative to the maximal values. Thus by definition we have $0 \le \sigma_k, \tau_k \le 1$. This description of age-dependent transmission corresponds to the "proportionate mixing" assumption in the continuous age-structured epidemic model [14]. Although it is straightforward to include more complicated mixing patterns in (2.1)–(2.3), such mixing patterns are not easily amenable to the analysis we shall apply below.

Initial conditions for the seasonal dynamics are determined by the demographic model, so that at the onset of the season, $s_k(0)$ and $r_k(0)$ are known. Since the epidemic from the previous season has run to its conclusion, we have $s_k(0) + r_k(0) = n_k = 1/m$ and $i_k(0) \ll 1$. At the beginning of the epidemic season a few infectious individuals are introduced into the population, so that $0 < \Lambda(0) \ll 1$. As we shall see it suffices to assume that $\Lambda(0) > 0$ rather than specifying the age-distribution of the initial infectious cases.

The description of the seasonal epidemic model can be simplified substantially. Since immune hosts play no role in the transmission dynamics, the r_k-equations are redundant and may be omitted. Furthermore by introducing the force of infection

as a dynamic variable, we can eliminate direct reference to the infectious classes i_k. Summing over all age-classes, one finds that Λ follows the equation

$$\frac{d\Lambda}{dt} = \left(c \sum \sigma_k \tau_k s_k - \nu\right) \Lambda. \tag{2.4}$$

The transmission dynamics during the epidemic season is therefore captured by (2.1) and (2.4).

These equations can be solved explicitly by Gart's modification to the analysis of the single epidemic [30]. We first introduce a reference age-class s_f, namely an age-class for which $\sigma_f = 1$. Eliminating time, we find that $ds_k/ds_f = \sigma_k s_k/s_f$, so that

$$s_k(t) = s_k(0) \left(\frac{s_f(t)}{s_f(0)}\right)^{\sigma_k}. \tag{2.5}$$

By substituting these values for s_k into (2.4), we obtain

$$\frac{d\Lambda}{ds_f} = \frac{c\Lambda \sum \sigma_k \tau_k s_k - \nu\Lambda}{-\Lambda s_f} = -c \sum \sigma_k \tau_k \left(\frac{s_f}{s_f(0)}\right)^{\sigma_k - 1} \frac{s_k(0)}{s_f(0)} + \frac{\nu}{s_f},$$

which may be solved to yield

$$\Lambda(t) - \Lambda(0) = -c \sum \tau_k \left(\frac{s_f(t)}{s_f(0)}\right)^{\sigma_k} s_k(0) + \nu \log(s_f(t)) + c \sum \tau_k s_k(0) - \nu \log(s_f(0)).$$

We are now ready to apply our separation of time scales, in that we express the final size of the epidemic in terms of $\theta = s_f(\infty)/s_f(0)$.

Provided that the threshold condition

$$\frac{d\Lambda}{dt}_{|t=0} = c \left(\sum \tau_k \sigma_k s_k(0) - \nu\right) \Lambda > 0 \tag{2.6}$$

is satisfied, an epidemic will occur. Since the epidemic will eventually die out, we have that $\Lambda(t) \to 0$ for $t \to \infty$. By assumption, $\Lambda(0) \ll 1$, so we find that if the threshold condition (2.6) holds, then θ satisfies the equation

$$G(\theta) = \Lambda(\infty) - \Lambda(0) = \log\theta + \gamma \sum \tau_k s_k(0)(1 - \theta^{\sigma_k}) = 0,$$

where $\gamma = c/\nu$.

It is straightforward to show that if $G'(1) < 0$, this equation has exactly one solution in the interval $(0, 1)$, and none if $G'(1) > 0$; for details see [6]. Clearly the condition on $G'(1)$ is identical to the threshold condition (2.6). If the threshold condition does not hold, no epidemic can occur, and consequently we set $\theta = 1$.

These observations lead to the following statement.

DEFINITION OF θ. *If the threshold condition*

$$\gamma \sum \tau_k \sigma_k s_k > 1 \tag{2.7}$$

holds, then $0 < \theta < 1$ *is the unique solution to*

$$\log\theta + \gamma \sum \tau_k s_k (1 - \theta^{\sigma_k}) = 0; \tag{2.8}$$

if the threshold condition is not satisfied, then

$$\theta = 1.$$

Notice that, by (2.5), θ now determines the changes in all susceptible classes, since we have

$$s_k(\infty) = \theta^{\sigma_k} s_k(0). \tag{2.9}$$

Thus the entire outcome of the epidemic season is captured in a single (implicit) function θ of the initial conditions.

Since we no longer will be concerned with the dynamics within a season, we use the notation $s_k = s_k(0)$ here and for the remainder of this paper. The definition of θ generalizes the concept of the final size of an epidemic in a closed population (cf. [12, p. 10]) to a closed structured population.

The quantity $\gamma \sum \tau_k \sigma_k s_k$ gives the number of secondary infections per primary infection at the onset of the epidemic, and it may therefore be thought of as the *effective replacement number*, R_e, at the onset of the epidemic.

We are now ready to describe the complete dynamics over a full year, from the beginning of one epidemic season to the beginning of the next season. Since $s_k + r_k = n_k = 1/m$, it suffices to specify the age-distribution of susceptible hosts $(s_1, s_2, \ldots, s_m)$, and the process taking the susceptible population from the onset of one school-year to the onset of the next year is described by the following two steps:

$$\begin{aligned}
\text{Start of year:} &\quad (s_1, s_2, \ldots, s_m) \\
\text{Epidemic season:} &\quad \downarrow \\
\text{End of year:} &\quad (\theta^{\sigma_1} s_1, \theta^{\sigma_2} s_2, \ldots, \theta^{\sigma_m} s_m) \\
\text{Change of school class:} &\quad \downarrow \\
\text{Start of next year:} &\quad (1/m, \theta^{\sigma_1} s_1, \theta^{\sigma_2} s_2, \ldots, \theta^{\sigma_{m-1}} s_{m-1}),
\end{aligned}$$

where the first step is given by (2.9) and the second step by the cohortwise age-progression. The long-term behavior of the epidemics is now determined by this map connecting the age-distribution of susceptible hosts from one year to the next. To facilitate the analysis of the system we introduce the following notation. Since at the beginning of each season $s_1 = 1/m$, the population structure is characterized by the $m - 1$ values of $S = (s_2, \ldots, s_m)$. The size of a susceptible cohort can only decrease, so we have $0 \le s_k \le 1/m$ for all k. Thus the state space for the system is $B = [0, 1/m]^{m-1}$. With this notation the season-to-season map $F : B \to B$ can be written in the form

$$F : \begin{pmatrix} s_2 \\ \vdots \\ s_m \end{pmatrix} \mapsto \begin{pmatrix} \theta^{\sigma_1}/m \\ \vdots \\ \theta^{\sigma_{m-1}} s_{m-1} \end{pmatrix}, \tag{2.10}$$

where the value of θ is defined above with the understanding that $s_1 = 1/m$.

The definition makes θ a continuous function $\theta : B \to [0, 1]$, which is nondifferentiable on the hyperplane where the threshold quantity equals unity. To get an impression of the function θ, let us for a moment set $\sigma_k = 1$ and define w as the weighted sum of all susceptibles $w = \sum \tau_k s_k$. We can now consider θ as a composite

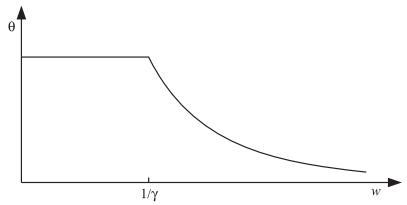


FIG. 2.1. *The function $\theta(w)$. The function is continuous but not differentiable at $w = 1/\gamma$ and has no inverse.*

function $S \mapsto w \mapsto \theta$, where θ is determined as function of w by

$$\theta(w) = \begin{cases} \text{the solution to} \\ 0 = \log\theta + \gamma w(1-\theta) & \text{if } \gamma w > 1; \\ 1 & \text{otherwise.} \end{cases}$$

Figure 2.1 shows the graph of $\theta(w)$ and its two unusual features. The function $\theta(w)$ is continuous but not differentiable at $w = \gamma$—in fact, we have $\theta'(1/\gamma^-) = 0$ and $\theta'(1/\gamma^+) = -2\gamma$. Furthermore, for $w \le 1/\gamma$ the function is a constant, so no inverse map exists. The constant section of $\theta(w)$ introduces a "folding" where a whole line-segment is mapped into the same point; this folding turns out to play a major role in the long-term dynamics of the system for some parameter values. The nondifferentiability of $\theta(w)$ leads to some technical (but less significant) complications in the mathematical analysis.

3. Repeated epidemics. The map (2.10) describes how the age-structure of susceptible hosts changes from the beginning of one school-year to the next, and we now turn our attention to the dynamics on the slow season-to-season time scale.

We first determine equilibria $\hat{S} \in B$ of the system in terms of $\vartheta = \theta(\hat{S})$, noting that an equilibrium $F(\hat{S}) = \hat{S}$ with $\vartheta < 1$ corresponds to a situation where an epidemic of the same size occurs in all seasons.

If $F(\hat{S}) = \hat{S}$ and $\vartheta = \theta(\hat{S})$, it follows that

$$\hat{s}_k = \vartheta^{\sigma_{k-1}}\hat{s}_{k-1} = \frac{\vartheta^{u_{k-1}}}{m}, \qquad k = 2, \dots, m,$$

where $u_k = \sum_{j=1}^{k} \sigma_j$ and $u_0 = 0$. Equilibria of the model are therefore characterized by those $\vartheta \in (0, 1]$ that solve the equation

$$H(\vartheta) = \log\vartheta + \frac{\gamma}{m}\sum \tau_k \vartheta^{u_{k-1}}(1 - \vartheta^{\sigma_k}) = 0. \tag{3.1}$$

Using the method in [6, Proposition 1], we have that the following properties hold for H and $\sigma^* = \min\{\, \sigma_k \,|\, \sigma_k > 0 \,\}$:

(1) $H(0^+) = -\infty$.
(2) $H(1) = 0$.
(3) $H'(1) = 1 - \frac{\gamma}{m}\sum \tau_k\sigma_k$.
(4) The function $H(\vartheta)/(1 - \vartheta^{\sigma^*})$ is increasing on the interval $(0, 1)$.

Property (2) ensures that H is positive to the left of $\vartheta = 1$ if $H'(1) < 0$, and in combination with property (1), this shows that $H(\vartheta)$ has a zero on the interval $(0, 1)$, while property (4) shows that this zero is unique. If $H'(1) > 0$, the function must

have an even number to zeroes on $(0,1)$, and from property (4) we conclude that no zeroes exist in $(0,1)$.

Since $H(1)=0$, $\vartheta=1$ will always satisfy (3.1). However, because of our assumption that an epidemic occurs if the threshold condition (2.7) is satisfied, θ evaluated at the disease-free state $S_0=(1,\dots,1)/m$ equals unity only if condition (2.7) does not hold at S_0. Therefore the disease-free state is an equilibrium only when the effective replacement number at S_0 is below 1.

We summarize our observations in the following claim.

PROPOSITION 3.1. *The model* (2.10) *has exactly one equilibrium. If the threshold condition*

$$R_0=\frac{\gamma}{m}\sum_{k=1}^{m}\tau_k\sigma_k>1 \tag{3.2}$$

holds, then the equilibrium is endemic, corresponding to an annual epidemic. If the threshold condition does not hold, then the equilibrium is disease-free, i.e., $s_k=1/m$, $k=1,\dots,m$.

The quantity R_0 gives the *basic replacement number*, i.e., the number of secondary infections per primary infection in a totally susceptible population, and Proposition 3.1 therefore shows that the model has an endemic equilibrium exactly when R_0 exceeds unity, as one expects in epidemic models [12]. However, in contrast to most epidemic models, there is no (unstable) disease-free equilibrium when condition (3.2) holds. This is due to our basic assumption that an infection from an external source will cause an epidemic in every season where it is possible.

The present model can exhibit quite complicated dynamics, including multiple steady states, chaos, and cycles where epidemics occur only in some seasons. We cannot provide a complete analysis of the model, but in the following section we analyze the stability of the endemic equilibrium for a particularly simple mixing structure, and in the subsequent section we provide some numerical examples of the complicated dynamics that can arise.

4. Local stability. From the previous section we know that for $R_0>1$ there exists a unique endemic equilibrium. In this section we will determine the local stability of this equilibrium for the case where disease susceptibility is independent of age, $\sigma_k=1, k=1,\dots,m$. This structure is chosen for its mathematical convenience rather than for its epidemiological interest, but it suffices when demonstrating the possibility of destabilization of the endemic equilibrium through variation in infection rates. Assuming constant susceptibility $\sigma_k=1$ (but age-dependent infectivity τ_k) simplifies the analysis considerably, especially because $\theta(S)$, the size of the epidemic, can be treated as a composite map consisting of a linear map followed by a (complicated) one-dimensional (1-D) map, as discussed in section 2.

The endemic equilibrium now takes the form

$$\hat{S}=(\hat{s}_2,\hat{s}_3,\dots,\hat{s}_m)=\frac{1}{m}(\vartheta,\vartheta^2,\dots,\vartheta^m),$$

where ϑ is the fraction of hosts that get infected during a season at equilibrium. It turns out that it is more convenient to use ϑ, rather than $\gamma=c/\nu$ as a (bifurcation) parameter. To see that this is possible we start by proving that the equilibrium condition expressed in terms of ϑ can be used to define a one-to-one map between γ and ϑ.

LEMMA 4.1. *For fixed age-specific infectivity* $(\tau_1, \dots, \tau_m)$ *the implicit equation*

$$\log \vartheta + \frac{\gamma}{m} \sum_{k=1}^{m} \tau_k \vartheta^{k-1}(1-\vartheta) = 0$$

defines a one-to-one map $\gamma \mapsto \vartheta$ *from* $(m/\sum \tau_k, +\infty)$ *onto* $(0,1)$.

Proof. According to Proposition 3.1, the map is well defined, and the inverse map $\vartheta \mapsto \gamma$ is given explicitly as

$$\gamma = \frac{-m \log \hat{\vartheta}}{\sum_{k=1}^{m} \tau_k \vartheta^{k-1}(1-\vartheta)}, \tag{4.1}$$

showing that the map is one-to-one. Clearly $\gamma(\vartheta)$ is positive on $(0,1)$ and $\gamma(0^+) = +\infty$, while $\gamma(1^-) = m/\sum \tau_k$.

The stability of the equilibrium $\hat{S}$ is determined by the Jacobian of F, given by (2.10) evaluated at $\hat{S}$. A straightforward computation yields

$$DF = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ \vartheta & 0 & \dots & 0 & 0 \\ 0 & \vartheta & & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \vartheta & 0 \end{pmatrix} + \begin{pmatrix} \hat{s}_1 \\ \hat{s}_2 \\ \hat{s}_3 \\ \vdots \\ \hat{s}_{m-1} \end{pmatrix} \begin{pmatrix} \theta_2' & \theta_3' & \dots & \theta_m' \end{pmatrix},$$

where

$$\theta_k' = \frac{\partial \theta}{\partial s_k \,|\hat{s}} = \frac{\gamma \tau_k (1-\vartheta)}{\gamma \sum_1^m \tau_j \hat{s}_j - \vartheta^{-1}} = -\frac{\gamma \tau_k (1-\vartheta)}{\frac{\log \vartheta}{1-\vartheta} + \frac{1}{\vartheta}} = \frac{\tau_k m}{\sum_1^m \tau_j \vartheta^{j-1}} \psi(\vartheta) < 0$$

denotes the partial derivative of θ after s_k evaluated at the equilibrium. The third equality is obtained by using the equilibrium condition

$$\log \vartheta + \gamma \sum \tau_k \hat{s}_k (1-\vartheta) = 0,$$

while the subsequent equality follows from (4.1).

The elementary function

$$\psi(\vartheta) = \frac{\log \vartheta}{\frac{\log \vartheta}{1-\vartheta} + \frac{1}{\vartheta}}$$

will play a central role in the analysis, and we note the following claim.

LEMMA 4.2. *The function* ψ *satisfies the inequality*

$$0 > \vartheta + \psi(\vartheta) > -1$$

on the interval $(0,1)$, *and* $\psi(0^+) = 0$, *while* $\psi(1^-) = -2$.

Proof. We find

$$\begin{aligned} 1 + \vartheta + \psi(\vartheta) &= 1 + \vartheta + \frac{\log \vartheta}{\frac{1}{\vartheta} + \frac{\log \vartheta}{1-\vartheta}} \\ &= 1 - \frac{\vartheta - 1 - \log \vartheta}{\vartheta^{-1} - 1 - \log \vartheta^{-1}} \qquad (4.2) \\ &= \frac{\vartheta^{-1} - \vartheta + 2 \log \vartheta}{\vartheta^{-1} - 1 - \log \vartheta^{-1}}. \qquad (4.3) \end{aligned}$$

By applying the elementary inequality $y - 1 - \log y > 0$ to the numerator and the denominator in expression (4.2), we observe that $\vartheta + \psi(\vartheta) < 0$. To see that $\vartheta + \psi(\vartheta) + 1 > 0$, apply the same inequality to the denominator of (4.3) and observe that $t(\vartheta) = \vartheta^{-1} - \vartheta + 2\log\vartheta > 0$ since $t(1) = 0$ and $t'(\vartheta) = -(1 - \vartheta^{-1})^2$.

In principle, the characteristic polynomial can be determined by a straightforward computation of the determinant of $DF - zE$. However, the following two lemmas considerably simplify the computation.

LEMMA 4.3. *For an $n \times n$-matrix M and (column) vectors c_1, c_2, and d we have*

$$\det(M + d(c_1 + c_2)^T) = \det(M + dc_1^T) + \det(M + dc_2^T) - \det M.$$

Proof. The proof follows from Theorem 8.4.3 in [22].

LEMMA 4.4. *Let $P_{nj}(a)$ denote the determinant of the $n \times n$ matrix*

$$L = \begin{pmatrix} -z & 0 & \dots & 0 & 0 \\ u & -z & \dots & 0 & 0 \\ 0 & u & & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & u & -z \end{pmatrix} + a \begin{pmatrix} 1 \\ u \\ u^2 \\ \vdots \\ u^{n-1} \end{pmatrix} e_j^T,$$

where e_j denotes the jth unit vector; then

$$P_{nj}(a) = \det L = (-1)^n \left(z^n - au^{j-1}(z^{n-1} + \ldots + z^{n-j})\right).$$

Proof. Expansion of the determinant after the first row of L gives the recursion formula

$$P_{nj}(a) = -zP_{n-1\,j-1}(au) + (-1)^{j+1}u^{j-1}(-z)^{n-j},$$

and the result follows by induction.

Combining the two lemmas, we find that the characteristic polynomial of $DF(\hat{S})$ is

$$\begin{aligned} p(z) &= (-1)^{m-1}\left(z^{m-1} - \frac{1}{m}\sum_{k=2}^{m}\theta_k'\vartheta^{k-2}(z^{m-2} + \cdots + z^{m-k})\right) \\ &= (-1)^{m-1}\left(z^{m-1} - \frac{1}{m}\sum_{k=2}^{m}\theta_k'\vartheta^{k-2}z^{m-k}\frac{z^{k-1} - 1}{z - 1}\right). \end{aligned} \tag{4.4}$$

Here the reader should bear in mind that since the first age-class is omitted, DF is in fact an $(m-1) \times (m-1)$-matrix, where θ_j' appears in the $(j-1)$th column. Since $\theta_k' < 0$, the penultimate expression is a polynomial where all coefficients are nonnegative, showing that the polynomial cannot have positive real zeroes. Therefore the last expression may be used to determine the eigenvalues of DF.

The polynomial (4.4) is too complex to allow for a general analysis, and we will study only a—somewhat artificial—situation where infected hosts in age-class $j, j \geq 2$, can infect with full strength $\tau_j = 1$, while hosts in other age-classes infect with reduced strength $\tau_k = \tau \leq 1$, for $k \neq j$. In particular, for $\tau = 1$ we obtain as a special case the homogeneous model where all hosts infect equally well, and we shall analyze this case in some detail below.

Since $\theta'_k = \tau\theta'_j$, $k \neq j$, we set $\theta'_j = \theta'$ and find that

$$\theta' = \frac{\psi(\vartheta)}{\tau\frac{\vartheta^m-1}{\vartheta-1} + (1-\tau)\vartheta^{j-1}}.$$

The characteristic polynomial takes the form

$$\begin{aligned} p(z) &= (-1)^{m-1}\left(z^{m-1} - \frac{1-\tau}{m}\theta'\vartheta^{j-2}z^{m-j}\frac{z^{j-1}-1}{z-1} - \frac{\tau}{m}\sum_{k=2}^{m}\theta'\vartheta^{k-2}z^{m-k}\frac{z^{k-1}-1}{z-1}\right) \\ &= \frac{(-z)^{m-1}}{z-1}\Bigg((z-1) - \frac{(1-\tau)\theta'}{m}\vartheta^{j-2}(1-z^{-j+1}) \\ &\qquad\qquad - \frac{\tau\theta'}{m}\left(\frac{\vartheta^{m-1}-1}{\vartheta-1} - z\frac{\left(\frac{\vartheta}{z}\right)^{m-1}-1}{\frac{\vartheta}{z}-1}\right)\Bigg). \end{aligned}$$

Thus eigenvalues of DF will satisfy the equation

$$0 = p(z) = \frac{z^{m+1} + az^m + bz^{m-1} - (c + dz^{m-j} + fz^{m-j+1})}{(z-1)(\vartheta - z)}, \tag{4.5}$$

where

$$\begin{aligned} a &= -\left(1 + \vartheta + \tau\frac{\theta'}{m}\frac{\vartheta^{m-1}-1}{\vartheta-1} + (1-\tau)\frac{\theta'}{m}\vartheta^{j-2}\right), \\ c &= \tau\frac{\theta'}{m}\vartheta^{m-1}, \\ d &= (1-\tau)\frac{\theta'}{m}\vartheta^{j-1}, \\ f &= -(1-\tau)\frac{\theta'}{m}\vartheta^{j-2}, \\ b &= -1 - a + c + d + f = \vartheta + \psi(\vartheta). \end{aligned}$$

From Lemma 4.2 we have that $0 > b > -1$.

We now describe the eigenvalues of DF by analyzing the roots of

$$h(z) = p(z)(z-1)(\vartheta - z) = z^{m+1} + az^m + bz^{m-1} - (c + dz^{m-j} + fz^{m-j+1}),$$

noticing that we have introduced artificial roots at $z = \vartheta, 1$, which do not correspond to eigenvalues of DF. Since DF cannot have positive real eigenvalues, double roots at these two values of z cannot occur.

The equation $h(z) = 0$ is not amenable to analytic solution, and we first study the two extreme cases $\tau = 1$, corresponding to homogeneous disease spread, and $\tau = 0$, corresponding to the situation where only age-class j can spread the disease.

4.1. Homogeneous disease transmission. For homogeneous disease transmission, $\sigma_k = \tau_k = 1$, the reproduction ratio simplifies to $R_0 = \gamma$, and from Proposition 3.1, we know that an internal (endemic) equilibrium exists exactly when $R_0 > 1$, while a disease-free equilibrium exists only when $R_0 < 1$. The picture is clear in that we can prove the following result.

PROPOSITION 4.5. *When disease spread is age-independent and $\gamma > 1$, the endemic equilibrium is always locally stable.*

The proof naturally breaks into two lemmas.

LEMMA 4.6. *If disease transmission is independent of age $\sigma_k = \tau_k = 1$, then the eigenvalues of the Jacobian DF cannot lie on the unit circle.*

Proof. When $\tau = 1$, the coefficients d and f vanish, and the equation $h(z) = 0$ simplifies to

$$z + a + bz^{-1} = cz^{-m}, \tag{4.6}$$

where the coefficients a and c are

$$a = -\left(1 + \vartheta + \psi(\vartheta)\frac{\vartheta^{m-1} - 1}{\vartheta^m - 1}\right),$$
$$c = \psi(\vartheta)\frac{\vartheta^m - \vartheta^{m-1}}{\vartheta^m - 1}.$$

Taking the absolute value of both sides of the equation, we find that roots of unit length $z = e^{i\omega}$ must satisfy the equation

$$((1 + b)\cos\omega + a)^2 + (1 - b)^2 \sin^2\omega = c^2.$$

Using the fact that $a + b + 1 = c$, the equation simplifies to

$$(\cos\omega - 1)[2a(1 + b) + 4b(1 + \cos\omega)] = 0.$$

Clearly $z = 1$ solves (4.6). However, (4.6) was obtained by multiplying the characteristic equation by $(z - 1)$ to simplify the expression. We have already observed that DF has no positive real eigenvalues, excluding the possibility of an eigenvalue at $z = 1$. Characteristic roots of unit length $z = e^{i\omega}$ are therefore possible only if

$$\cos\omega = -1 - \frac{a(b + 1)}{2b},$$

and thus it suffices to show that

$$\frac{a(b + 1)}{2b} > 0.$$

Since we know that $0 > b > -1$, we need only show that $a < 0$.

To see that $a < 0$, observe that

$$-a = 1 + \vartheta + \psi(\vartheta)\frac{1 - \vartheta^{m-1}}{1 - \vartheta^m}$$
$$> 1 + \vartheta + \psi(\vartheta) > 0.$$

We conclude that roots of the characteristic polynomial cannot cross the unit circle in the case of homogeneous mixing, completing the proof of Lemma 4.6.

LEMMA 4.7. *If disease transmission is independent of age $\tau_k = \sigma_k = 1$, then for large epidemics ($\vartheta \ll 1$) all eigenvalues of DF lie within the unit circle.*

Proof. At $\vartheta = 0$, the characteristic polynomial (and, in fact, the entire model!) is singular in the sense that 0 is a root of multiplicity m, and we apply a singular perturbation analysis to the characteristic equation. For ϑ small, $\psi(\vartheta) \approx \vartheta \log\vartheta$, so that (4.6) may be written in the form

$$z^{m+1} - (1 - \vartheta\log\vartheta)z^m + \vartheta\log\vartheta z^{m-1} - \vartheta^{m-1}\log\vartheta = 0, \tag{4.7}$$

where we have retained only the terms of leading order. By inspection, it is clear that the equation has the following $m+1$ approximate solutions: $z=1$, $z=\vartheta\log\vartheta$, and $z=w_k\vartheta, k=1,\dots,m-1$, where w_k denotes the $m-1$ roots of unity solving the equation $w^{m-1}=1$. Since (4.7) must have exactly $m+1$ complex solutions, our list is exhaustive. During our simplification of the characteristic polynomial we introduced additional roots at $z=1$ and $z=\vartheta$, and we conclude that the remaining $m-1$ solutions are in fact the eigenvalues of DF, all of which lie well within the unit circle.

Since, by the implicit function theorem, solutions of the characteristic equation are continuous in ϑ, Lemmas 4.6 and 4.7 immediately give us Proposition 4.5.

4.2. Age-dependent transmissibility. We now return to the situation where age-class j, $j\geq 2$, can infect at full strength $\tau_j=1$, while all other age-classes infect with reduced strength $\tau_k=\tau$, $k\neq j$. We first study the extreme case of $\tau=0$ and show the following result.

PROPOSITION 4.8. *If only age-class j, $j\geq 2$, can infect, then the endemic equilibrium is always unstable.*

From an intuitive viewpoint the proposition is quite simple. Since all hosts are equally susceptible, an epidemic will reduce the susceptible population in all age-classes by the same amount, and since the infective age-class j can no longer support an epidemic, neither can younger age-classes. Thus a new epidemic cannot arise before a new cohort has been born and has lived for j seasons, bringing it into the age-class where it is capable of infecting others. We now offer a formal proof.

Proof of Proposition 4.8. The proof follows the same method as that of Proposition 4.5, and we first show that roots of the characteristic polynomial cannot cross the unitcircle. For $\tau=0$, the equation $h(z)=0$ simplifies to

$$z+a+bz^{-1}=z^{-j}(d+zf),$$

where $a=-\left(1+\vartheta+\psi(\vartheta)/\vartheta\right)$, $b=\vartheta+\psi(\vartheta)$, $d=\psi(\vartheta)$, and $f=-\psi(\vartheta)/\vartheta$. Since age-classes $j+1,\dots,m$ do not contribute to disease transmission, the corresponding variables span an $(m-j)$-dimensional generalized nullspace, so that we need to identify only $j-1$ eigenvalues (plus the two roots at $z=1,\vartheta$). As in Lemma 4.6, we can derive a necessary condition for the existence of roots of unit length $z=e^{i\omega}$ and find that ω must satisfy the equation

$$\begin{aligned}\cos\omega &= -1-\frac{a(b+1)-df}{2b}\\ &= -1+\frac{\left(1+\vartheta+\frac{\psi(\vartheta)}{\vartheta}\right)\left(1+\vartheta+\psi(\vartheta)\right)-\frac{\psi(\vartheta)^2}{\vartheta}}{2(\vartheta+\psi(\vartheta))}\\ &= -1+\frac{\frac{(1+\vartheta)^2}{\vartheta}}{2}>1,\end{aligned}$$

excluding the existence of such roots.

It remains to see that there exists one value of ϑ for which the equation has solutions outside the unit circle. This may be done by studying small ϑ, as in Lemma 4.7. One finds that the leading terms of the $j+1$ nonzero roots of $h(z)$ are ϑ, $\log\vartheta$, and the $j-1$ roots of unity solving $w^{j-1}=1$. Clearly the root near $\log\vartheta$ lies outside the unit circle.

For fixed ϑ, the endemic equilibrium is stable for $\tau=1$ and unstable for $\tau=0$, so by the continuity of the characteristic roots, destabilization occurs for an intermediate value of $\tau=\tau_B$. Our analysis cannot exclude the possibility of multiple stability

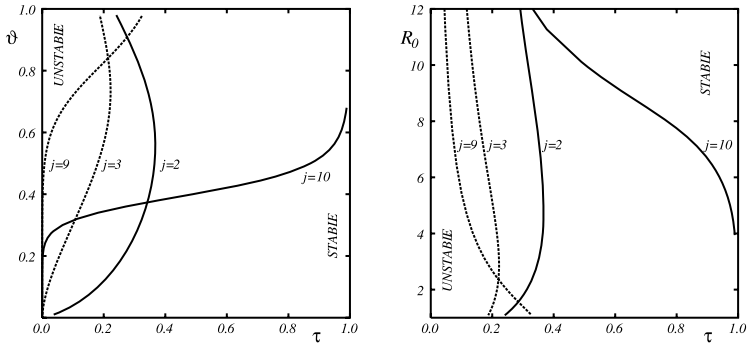


FIG. 4.1. *The location of the critical value $\tau = \tau_B$ for which the endemic equilibrium loses its stability. The parameter τ gives the strength of disease transmission from all age-classes except the highly infective age-class j. The value of τ_B is determined by numerically solving the modified characteristic equation $h(z) = 0$ using Laguere's method* [36]*, excluding the artificial solutions at $z = 1$ and applying a bisection method (in τ-space) to locate the value for which the largest solution is of unit length. The number of age-classes is fixed at $m = 10$. The left-hand panel shows the diagram in (τ, ϑ)-space, which is used in the analysis. The right-hand panel shows the diagram in (τ, R_0)-space, which is more natural for biological interpretation.*

switches, but numerical investigations of the equation $h(z) = 0$ indicate that such switches do not occur. Figure 4.1 shows τ_B for various j and $m = 10$. Clearly odd and even j lead to qualitatively different stability regions; furthermore, for odd j the bifurcation seems to be always a Hopf-type bifurcation through a pair of complex eigenvalues, while the bifurcation for even j is a flip bifurcation through an eigenvalue of -1. Since the age-structure of disease transmission is chosen for its mathematical convenience rather than for biological reasons, the phenomenon probably has no biological significance.

5. Nonequilibrium dynamics. In addition to the endemic equilibrium, more complicated dynamics may occur, and we finish our discussion of the school model with some examples of these complications.

For $m = 3$, i.e., three age-classes, the underlying dynamics is 2-D. Using the methods of [6], we have studied in some detail the case of age-independent susceptibility for $\tau_3 = 1$ and $\gamma = 5$. The analysis is quite similar to that of [6], so we only sketch the results; details of the analysis and the exact location of the bifurcations may be found in [19].

Figure 5.1 shows the main features of the bifurcations for $\tau_1 = 0.1$. The structure is most easily explained by starting at high values of τ_2, i.e., with the lower right-hand panel of Figure 5.1. For $\tau_2 = 0.60$ the endemic equilibrium is stable. When τ_2 is decreased, the equilibrium undergoes a Hopf bifurcation, giving rise to a stable limit-cycle, e.g., for $\tau_2 = 0.53$. As τ_2 decreases further, the limit-cycle degenerates, in that a segment of the limit-cycle meets the region of state space where no epidemic occurs, and three "arms" are created, $\tau_2 = 0.32$. Numerical simulations suggest that the dynamics on this attractor is chaotic, in that the trajectories of neighboring points diverge. In addition, the third iterate of F undergoes a saddle-node bifurcation, creating a stable and an unstable 3-cycle. As τ_2 decreases even more, the unstable 3-cycle meets the three "arms" on the chaotic attractor; a segment of the "arms" now lies in the basin of attraction for the stable triennial-annual cycle, and the chaotic attractor loses its stability.

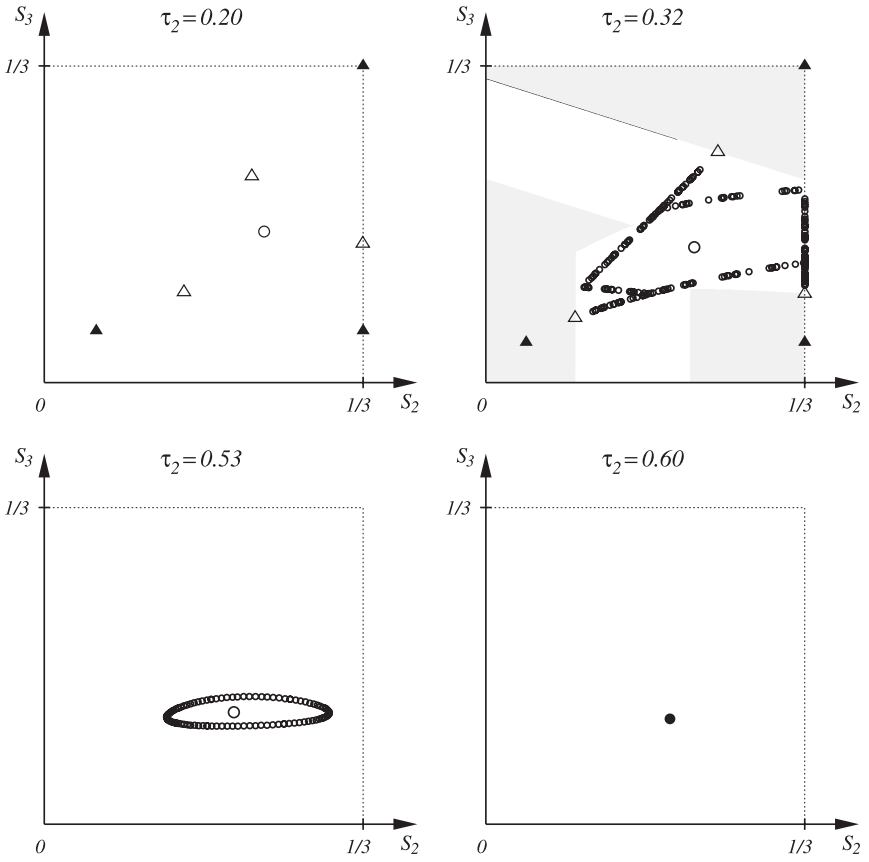


FIG. 5.1. *Trajectories and equilibria for the map F, showing the bifurcations that give rise to the complicated dynamics for $m = 3$. In all panels $\gamma = 5$, $\tau_1 = 0.1$, and $\tau_3 = 1$. For $\tau_2 = 0.60$ the endemic equilibrium is stable. For $\tau = 0.53$ the endemic equilibrium is unstable, and a stable periodic (or quasi-periodic) cycle is created through a Hopf bifurcation. For $\tau_2 = 0.32$ the limit cycle has met the boundary, and "arms" appear on the cycle. In addition, a stable and an unstable 3-cycle have appeared through a saddle-node bifurcation of the third iterate F^3. Notice that the "arms" of the attractor almost touch the unstable 3-cycle. The stable attractor will disappear in a global bifurcation for slightly smaller τ_2, and for $\tau_2 = 0.20$ the stable 3-cycle is globally stable. Other symbols: ○ endemic equilibrium; △ three-cycle. Filled marks indicate that the point is stable; open marks indicate that the point is unstable. For $\tau_2 = 0.32$ the shaded area indicates the basin of attraction for the stable 3-cycle.*

The dynamics of the model appears to be similar—but not identical—to those of the "realistic age-structured models" based on Schenzle's approach where the transfer of infection from season to season is accounted for explicitly [42, 10]. Figure 5.2 shows the bifurcation diagrams for model (2.10) and for a modified version of Schenzle's model. Schenzle's original model does not assume proportionate mixing in disease transmission, so it is not directly comparable to the present model. In addition, Schenzle assumes that births and deaths are distributed evenly over the year, and he accounts in detail for changes in disease transmission in response to school closure during weekends and breaks. In the modified model we neglect the seasonal variation in transmission and set the transmission coefficient from cohort j to cohort k to

$$\beta_{jk} = R_0 b_j b_k N \nu,$$

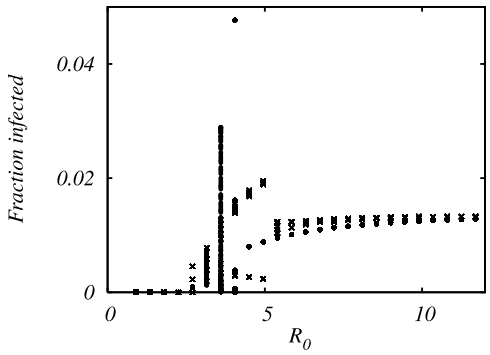


FIG. 5.2. *Bifurcation diagram showing the fraction of the host population that is infected in a season as a function of the basic reproduction number R_0. Crosses: model* (2.10)*; dots: a modified version of Schenzle's "realistic age-structured model," where the size of the infection is carried over between seasons; for details see the text.*

where

$$b_k = \begin{cases} \sqrt{0.5} & \text{for } 1 \le k \le 6, \\ 1 & \text{for } 7 \le k \le 10, \\ \sqrt{3.5/9.0} & \text{for } 11 \le k \le 20, \\ \sqrt{1/3} & \text{for } k = 21. \end{cases}$$

Here $k = 21$ is Schenzle's adult age-class. The strange values of b_k are chosen to ensure that the values of $\beta_{kk} = b_k^2$ are identical to those used by Schenzle. Following Schenzle, we assume that new individuals are born susceptible and at a constant rate during the year, while deaths occur only in the adult class and at a constant rate corresponding to a mean residence time of 55 years. To compare this "Schenzle" model to model (2.10), we have set $\tau_k = \sigma_k = b_k$ and included a total of $m = 75$ cohorts. Clearly the carry-over of infection from one season to the next has significant impact on the dynamics for small values of R_0, suggesting that stochastic effects during the summer break may play a role in this situation.

6. Discussion. The school model gives a somewhat different picture of disease transmission dynamics than that of the well-known SIR-model [24, 3]. In the present model the disease appears after the introduction of a new susceptible cohort, it produces an epidemic, and disappears, quite similarly to our everyday experience of epidemic diseases such as measles and influenza. The role of the school-year, class-progression, and admission of new students is quite explicit. The emphasis is on the dynamics of the age-specific herd immunity rather than on the virus population as such. Consequently the model does not address the question of disease persistence during periods of low disease transmission. In fact, we are assuming that the virus population will survive between epidemics in some unspecified reservoir. The model, for example, could describe transmission dynamics in a medium-sized town or island that is fairly well connected to a large city. This suggests a way to separate the non-linear transmission dynamics during periods of high transmission from the persistence problem where the stochastic effects due to the finite population size may play a significant role [21]. In the case of isolated island populations it will allow us to separate the (stochastic) arrival of an index case from the subsequent transmission dynamics during the epidemic; we are currently exploring this issue.

Since the endemic equilibrium appears to be stable unless the age-dependent variation in transmissibility is unrealistically pronounced, our findings do not corroborate Schenzle's [42] suggestion that the remarkably stable two-year cycle of measles epidemics in many urban areas is caused by the uptake of new susceptible hosts immediately after the summer break. Still we have not explored the entire parameter space, and in particular we have not studied the effect of age-dependent susceptibility in any detail, so a final conclusion awaits further investigations.

In the model, focus is on the slow year-to-year time scale, neglecting the details within the epidemic season, and the model does not give information about the precise onset of the epidemic. In some way this may be an advantage, since the exact timing of the epidemics may depend on specific geographic details of the meta-population in question [17]. However, measles epidemics tend to last several months, and it has been suggested that the exact timing of school breaks may in fact influence the dynamics of childhood diseases [41]. The present model, of course, is not designed to capture such subtleties.

Similarly to other epidemic models, we identified a basic reproduction number R_0 giving the number of secondary infections per primary infection in a susceptible population and showed that when $R_0 > 1$ there exists an endemic equilibrium. When disease transmission is independent of age, this equilibrium is always stable, but when infectivity is concentrated in just one age-class, the equilibrium is unstable. These results parallel those for the SIR-model with continuous aging of the host population, where the endemic equilibrium is stable for homogeneous disease transmission and short duration of infection [4] but may lose stability when disease transmission is age-dependent [5, 43, 27]. It is remarkable that the stability results are considerably easier to obtain for the present model than they were for the PDE model with continuous age-progression.

In the unstable region of parameter-space, the dynamics can be quite complicated, with multiple stable states, complicated attractors, and an unusual global bifurcation. These features are due to the model formulation in which the dynamics are governed by a differential equation for a period followed by a discrete map. Such hybrid models have become increasingly common in biology and epidemiology [35, 39, 40], and also general results on "time scale calculus" are available [1]. Usually the models do not exhibit as complicated behavior as the present model. However, in time scale calculus—as in most specific applications—one assumes that the dynamics follows the differential equation model for a *finite time*, ensuring that the composite map is in fact smooth. In contrast, for our epidemic model the fast time scale in effect runs over the entire time interval $(-\infty, +\infty)$. Smoothness of the map is no longer guaranteed, and in fact our map is continuous but nondifferentiable along the hyperplane defining the epidemic threshold. In addition, our map is constant below the threshold, allowing for a "folding" that seems to be responsible for the observed chaotic dynamics. It is unclear how one might extend time scale calculus to include such dynamics.

REFERENCES

[1] R. Agarwal, M. Bohner, D. O'Regan, and A. Peterson, *Dynamic equations on time scales: A survey*, J. Comput. Appl. Math., 141 (2002), pp. 1–26.

[2] R. M. Anderson and R. M. May, *Vaccination against rubella and measles: Quantitative investigations of different policies*, J. Hyg. Camb., 90 (1983), pp. 259–325.

[3] R. M. Anderson and R. M. May, *Infectious Diseases of Humans*, Oxford University Press, Oxford, U.K., 1991.

[4] V. Andreasen, *Age-dependent host mortality in the dynamics of endemic infectious diseases*, Math. Biosci., 114 (1993), pp. 29–58.
[5] V. Andreasen, *Instability in an SIR-model with age-dependent susceptibility*, in Mathematical Population Dynamics, O. Arino, D. Axelrod, M. Kimmel, and M. Langlais, eds., Wuerz Publishing, Winnipeg, ON, 1995, Vol. 1, pp. 3–14.
[6] V. Andreasen, *Dynamics of annual influenza A epidemics with immuno-selection*, J. Math. Biol., 46 (2003), pp. 504–536.
[7] J. L. Aron and I. B. Schwartz, *Seasonality and period-doubling bifurcations in an epidemic model*, J. Theoret. Biol., 110 (1984), pp. 665–679.
[8] C. T. Bauch and D. J. D. Earn, *Transients and attractors in epidemics*, Proc. Roy. Soc. London Ser. B Biol. Sci., 270 (2003), pp. 1573–1578.
[9] L. Billings and I. B. Schwartz, *Exciting chaos with noise: Unexpected dynamics in epidemic outbreaks*, J. Math. Biol., 44 (2002), pp. 31–48.
[10] B. Bolker, *Chaos and complexity in measles models: A comparative numerical study*, IMA J. Math. Appl. Med. Biol., 10 (1993), pp. 83–95.
[11] M. F. Boni, J. R. Gog, V. Andreasen, and F. B. Christiansen, *Influenza drift and epidemic size: The race between generating and escaping immunity*, Theoret. Pop. Biol., 65 (2004), pp. 179–191.
[12] O. Diekmann and J. A. P. Heesterbeek, *Mathematical Epidemiology of Infectious Diseases*, Wiley, Chichester, UK, 2000.
[13] K. Dietz, *Transmission and control of arbovirus diseases*, in Epidemiology, D. Ludwig and K. L. Cooke, eds., SIAM, Philadelphia, 1975, pp. 104–121.
[14] K. Dietz and D. Schenzle, *Proportionate mixing models for age-dependent infection transmission*, J. Math. Biol., 22 (1985), pp. 117–120.
[15] G. Dwyer, J. Dushoff, J. S. Elkinton, and S. A. Levin, *Pathogen driven outbreaks in forest defoliators revisited: Building models from experimental data*, Amer. Natur., 156 (2000), pp. 105–120.
[16] D. J. D. Earn, P. Rohani, B. M. Bolker, and B. T. Grenfell, *A simple model for complex dynamical transitions in epidemics*, Science, 287 (2000), pp. 667–670.
[17] B. Finkenstädt and B. Grenfell, *Empirical determinants of measles metapopulation dynamics in England and Wales*, Proc. Roy. Soc. London Ser. B Biol. Sci., 265 (1998), pp. 211–220.
[18] B. F. Finkenstädt and B. T. Grenfell, *Time series modelling of childhood diseases: A dynamical systems approach*, Appl. Statist., 49 (2000), pp. 187–205.
[19] T. Frommelt, *Periodisk forekomst af børnesygdomme (Periodic recurrence of childhood diseases)*, Masters Thesis, Department of Mathematics and Physics, Roskilde University, Roskilde, Denmark, 2002.
[20] J. H. Gillespie, *Natural selection for resistance to epidemics*, Ecology, 56 (1975), pp. 493–495.
[21] J. R. Gog, G. F. Rimmelzwaan, A. D. M. E. Osterhaus, and B. T. Grenfell, *Population dynamics of rapid fixation in cytotoxic T lymphocyte escape mutants of influenza A*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 11143–11147.
[22] F. A. Graybill, *Introduction to Matrices with Applications in Statistics*, Wadsworth, Belmont, CA, 1969.
[23] B. T. Grenfell, A. Kleczkowski, S. P. Ellner, and B. M. Bolker, *Measles as a case study in nonlinear forecasting and chaos*, Phil. Trans. Roy. Soc. London Ser. A Phys. Sci., 348 (1994), pp. 515–530.
[24] H. W. Hethcote, *Asymptotic behavior and stability of epidemic models*, in Mathematical Problems in Biology, P. van der Driessche, ed., Springer-Verlag, Berlin, Heidelberg, New York, 1974, pp. 83–92.
[25] H. W. Hethcote, *The mathematics of infectious diseases*, SIAM Rev., 42 (2000), pp. 599–653.
[26] F. Hoppensteadt, *Mathematical Theories of Populations: Demographics, Genetics, and Epidemics*, CBMS-NSF Conf. Ser. Appl. Math. 20, SIAM, Philadelphia, 1975.
[27] H. Inaba, *Threshold and stability results for an age-structured epidemic model*, J. Math. Biol., 28 (1990), pp. 411–434.
[28] M. J. Keeling and B. T. Grenfell, *Disease extinction and community size: Modeling the persistence of measles*, Science, 275 (1997), pp. 65–67.
[29] M. J. Keeling and B. T. Grenfell, *Understanding the persistence of measles: Reconciling theory, simulation and observation*, Proc. Roy. Soc. London Ser. B Biol. Sci., 269 (2002), pp. 335–343.
[30] W. O. Kermack and A. G. McKendrick, *Contributions to the mathematical theory of epidemics* 1, Proc. Roy. Stat. Soc. A, 115 (1927), pp. 700–721; reprinted in Bull. Math. Biol., 80 (1997), pp. 243–248.

[31] W. P. London and J. A. Yorke, *Recurrent outbreaks of measles, chickenpox, and mumps: I. Seasonal variation in contact rates*, Amer. J. Epidemiol., 98 (1973), pp. 453–468.
[32] R. M. May, *Regulation of populations with nonoverlapping generations by microparasites: A purely chaotic system*, Amer. Natur., 125 (1985), pp. 573–584.
[33] L. F. Olsen and W. M. Schaffer, *Chaos versus noisy periodicity: Alternative hypotheses for childhood epidemics*, Science, 249 (1990), pp. 499–504.
[34] R. Pearl, *The Rate of Living*, Knopf, New York, 1928.
[35] E. T. Poulsen, *A model for population regulation with density- and frequency-dependent selection*, J. Math. Biol., 8 (1979), pp. 325–343.
[36] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes*, 2nd ed., Cambridge University Press, New York, 1992.
[37] D. A. Rand and H. B. Wilson, *Chaotic stochasticity: A ubiquitous source of unpredictability in epidemics*, Proc. Roy. Soc. London Ser. B Biol. Sci., 246 (1991), pp. 179–184.
[38] T. A. Reichert, N. Sugaya, D. S. Fedson, W. P. Glezen, L. Simonsen, and M. Tashiro, *The Japanese experience with vaccinating school-children against influenza*, New England J. Med., 344 (2001), pp. 889–896.
[39] M. G. Roberts and J. A. P. Heesterbeek, *A simple parasite model with complicated dynamics*, J. Math. Biol., 37 (1998), pp. 272–290.
[40] M. G. Roberts and R. R. Kao, *The dynamics of an infectious disease in a population with birth pulses*, Math. Biosci., 149 (1998), pp. 23–36.
[41] P. Rohani, M. J. Keeling, and B. T. Grenfell, *The interplay between determinism and stochasticity in childhood diseases*, Amer. Natur., 159 (2002), pp. 469–481.
[42] D. Schenzle, *An age-structured model of pre- and post-vaccination measles transmission*, IMA J. Math. Appl. Med. Biol., 1 (1984), pp. 169–191.
[43] H. R. Thieme, *Stability change of the endemic equilibrium in age-structured models for the spread of SIR type infectious diseases*, in Differential Equations Models in Biology, Epidemiology and Ecology, S. Busenberg and M. Martelli, eds., Springer-Verlag, Berlin, Heidelberg, New York, 1991, pp. 139–158.
[44] D. W. Tudor, *An age-dependent epidemic model with application to measles*, Math. Biosci., 73 (1985), pp. 131–147.

SIAM J. APPL. MATH.
Vol. 65, No. 6, pp. 1888–1912

WAVELET MIE REPRESENTATIONS FOR SOLENOIDAL VECTOR FIELDS WITH APPLICATIONS TO IONOSPHERIC GEOMAGNETIC DATA*

THORSTEN MAIER†

Abstract. A wavelet technique, the wavelet Mie representation, is introduced for the analysis and modeling of the earth's magnetic field and corresponding electric current distributions from geomagnetic data obtained within the ionosphere. The considerations are essentially based on two well-known geomathematical keystones, (i) the Helmholtz decomposition of spherical vector fields and (ii) the Mie representation of solenoidal vector fields in terms of poloidal and toroidal parts. The wavelet Mie representation is shown to provide an adequate tool for geomagnetic modeling in the case of ionospheric magnetic contributions and currents which exhibit spatially localized features. An important example is ionospheric currents flowing radially onto or away from the earth. To demonstrate the functionality of the approach, such radial currents are calculated from vectorial data of the MAGSAT and CHAMP satellite missions.

Key words. Mie representation, Helmholtz decomposition, vectorial wavelets, geomagnetic field modeling

AMS subject classifications. 42C40, 65Z05, 86A25

DOI. 10.1137/040603796

1. Introduction.

Macroscopic electrodynamics is the theoretical basis for dealing with the subject of satellite magnetometry in geomagnetism. The fundamental equations governing that branch are Maxwell's equations for polarizable media. Since typical time-scales in satellite magnetometry are on the order of days and typical length-scales are on the order of the earth's radius, the typical system velocities are much smaller than the speed of light, and therefore the quasi-static (or stationary) approximations of Maxwell's equations (i.e., the pre-Maxwell equations) can be used (cf., e.g., [2]). As far as the magnetic field is concerned, these equations read

$$\nabla \cdot b = 0,$$
$$\nabla \wedge b = \mu_0 j,$$

where b (in classical geophysical notation usually denoted by $\vec{B}$) is the magnetic induction, i.e., the magnetic field; j is the electric current density; and μ_0 is the vacuum permeability, $\mu_0 = 4\pi \cdot 10^{-7}\mathrm{VsA}^{-1}\mathrm{m}^{-1}$. Note that, in this approximation, the electric current density j is also of zero divergence, i.e.,

$$\nabla \cdot j = 0.$$

Many concepts in geomagnetic modeling assume that the geomagnetic data are solely collected within a spherical shell $\Omega_{(R_1,R_2)}$ around the origin—with inner radius R_1 and outer radius R_2—between the earth's surface and the ionosphere, so that the current density j can be neglected. This results in $\nabla \wedge b = 0$, $\nabla \cdot b = 0$, which implies

*Received by the editors February 4, 2004; accepted for publication (in revised form) January 27, 2005; published electronically August 3, 2005. This work was supported by the priority program "Geomagnetic Variations" of the German Research Foundation (DFG FR 761/10-1).
http://www.siam.org/journals/siap/65-6/60379.html
†Department of Mathematics, Geomathematics Group, University of Kaiserslautern, P.O. Box 3049, 67663 Kaiserslautern, Germany (tmaier@mathematik.uni-kl.de).

that there exists a scalar potential U in $\Omega_{(R_1,R_2)}$ such that $b = -\nabla U$ and $\Delta U = 0$ in $\Omega_{(R_1,R_2)}$. In order to model the magnetic field b the potential is expanded into a Fourier series of (scalar) spherical harmonics and the expansion coefficients are chosen such that the gradient of the potential fits—in the sense of a least-square metric—the given vectorial data as well as possible. This method, which is known as Gauss representation, has been used and constantly improved for more than 150 years now, so that profound numerical and theoretical techniques exist (see, e.g., [23]).

Satellite missions (like MAGSAT, Oersted, and CHAMP) collect their data *within* the ionosphere. Due to the intense solar radiation on the earth's dayside (i.e., the hemisphere directed to the sun), the electric conductivity of the ionosphere is increased, and tidal forces, due to solar heating as well as solar and lunar attraction, can drive large electric current systems. Among the most important ionospheric current systems are the so-called equatorial electro jet (EEJ) and the polar electro jets (PEJ), as well as the so-called field aligned currents that are flowing radially towards and away from the geomagnetic poles. In connection with polarization effects in the ionospheric plasma, the geomagnetic field produces an enhanced Hall conductivity (Cowling effect) in the vicinity of the geomagnetic equator. This increased conductivity results in an amplified current system—the EEJ—flowing roughly along the magnetic equator. With regard to our later considerations, it is worth mentioning that the EEJ, though mainly tangential, also provides a notable radial current density, which is known as the radial contribution of the meridional current system of the EEJ. The PEJ is mainly due to an increased conductivity and large horizontal electric field contributions in the polar ionosphere. Currents flowing along the geomagnetic field lines—the field aligned currents—are caused by magnetospheric and ionospheric coupling or imbalances of Sq-current systems (see, e.g., [29] and the references therein). In the polar regions field aligned currents flow onto or away from the earth's body, thus contributing large radial current densities confined to these areas. The radial currents and the resulting magnetic effects, as well as the corresponding modeling approaches, are increasingly the focus of research (see, for example, in chronological order, [30], [34], [9], [23], [29], [11], [4], [27], [31], and [35]). The numerical examples presented in this article also deal with the determination of such radial ionospheric currents from geomagnetic vectorial satellite data.

Due to the electric currents, the magnetic field measured by satellites in the ionosphere is no longer a gradient-field. In fact, it now contains magnetic contributions from current densities on the satellite's track. But this means that new vectorial methods, not based on the existence of a scalar potential, must be derived in close orientation on a (quasi-static) formulation of Maxwell's equations. The authors of [1], [2], [18], [34] suggest the resolution of the magnetic field by means of the Mie representation as an adequate replacement of the Gauss approach. The Mie representation, i.e., splitting the magnetic field into poloidal and toroidal parts, has the advantage that it can equally be applied in regions of vanishing as well as nonvanishing electric current densities. The poloidal fields are due to toroidal current densities below and above the satellite's track, whereas the toroidal fields are created by the radial currents which are crossing the satellite's orbit. It is this characteristic that makes the Mie approach a powerful tool for dealing with geomagnetic source problems, i.e., the problems of calculating magnetic effects due to given electric currents (direct source problem) and, conversely, determining those current distributions that produce a predefined magnetic field (inverse source problem).

There remains the question of how to computationally obtain, in terms of suitable trial functions, the Mie representation from a given set of vectorial data. Most

of the considerations in [1], [2] and all the results in [11], [25], [29] are based on a spherical harmonic parametrization; i.e., the starting points of the considerations are expansions of the poloidal and toroidal scalars in terms of spherical harmonics. On the one hand, this approach is advantageous since it admits the possibility of incorporating radial dependencies of magnetic fields and electric currents in a natural way. On the other hand, the global support of the spherical harmonics limits the practicability of this technique since it cannot cope with electric currents (and corresponding magnetic effects) that vary rapidly with latitude or longitude or that are confined to certain regions. In fact, Backus [1] states that it might be advantageous to find a field parametrization in terms of functions that take efficient account of the specific concentration of the current densities in space. The uncertainty principle (see the scalar theory by Freeden and Windheuser [17] and their generalization to the vector case by Beth [5]) provides an adequate tool for the classification of (spherical restrictions of) poloidal and toroidal vector fields by determining a trade-off between two "spreads," one for the position (space) and the other for the momentum (frequency). The main statement is that sharp localizations in space and in frequency are mutually exclusive. The varieties of space/frequency localization can be illustrated by considering different poloidal and toroidal trial fields on the sphere as suitable for constructive approximation. Vector (spherical) harmonics show an ideal frequency localization but no space localization. The spectrum of (band-limited and non–band-limited) kernel functions known from harmonic and vectorial spline theory (cf. [12], [33], [14], [15]) shows all intermediate cases of space/frequency localization. But in view of the amount of space/frequency localization, it is also worth distinguishing band-limited from non–band-limited kernels. As a matter of fact, it turns out that non–band-limited kernels show a much stronger space localization than their comparable band-limited counterparts. Roughly speaking, this is due to the fact that band-limited kernels can be represented as finite sums of polynomials and therefore—though strongly smoothed compared to polynomial functions—tend to oscillate. In contrast, non–band-limited kernels cannot be displayed as finite sums of polynomials and hence yield a stronger space localization. Finally, the Dirac kernels show ideal space localization but no frequency localization. Thus they provide the final stage in the spatial resolution of the magnetic field by trial functions. In conclusion, vector harmonics and Dirac kernels are "extreme trial functions" for purposes of geomathematical modeling. These facts help us to find a suitable characterization and categorization of the trial functions for modeling and approximation: Fourier methods (in terms of scalar/vector spherical harmonics, for example) are the canonical starting point for obtaining an approximation of low frequency contributions (global modeling), while band-limited kernel functions can be used for the intermediate cases between long and short wavelengths (global to regional modeling). Due to their extreme space localization, non–band-limited kernels can be utilized to deal with short wavelength phenomena (local modeling). Most data show correlation in space as well as in frequency, and the kernel functions with their simultaneous space and frequency localization allow for the efficient detection and approximation of essential features in the data by using only a fraction of the original information (decorrelation). Using kernels at different scales (*multiscale modeling*), the corresponding approximation techniques can be constructed so as to be suitable for the particular data situation.

In this article we are concerned with *wavelet techniques* for the parametrization of the Mie representation, i.e., methods based on certain classes of kernel functions, the *scaling functions and wavelets*. Suitably constructed wavelets admit a basis property in certain function spaces, the elements of which—the data functions—admit a series

representation in terms of a structured sequence of kernels at different positions and at different scales (*multiscale approximation*). It is thus possible to break up complicated functions like the geomagnetic field, electric current densities, or geopotentials into different pieces and to study these pieces separately. Consequently, the efficiency of wavelets lies in the fact that only a few wavelet coefficients are needed in areas where the data are smooth, and in regions where the data exhibit more complicated features, higher resolution approximations can be derived by "zooming-in" with more and more wavelets of higher scales and consequential stronger space localization.

The outline of the paper is as follows. In section 2 the fundamentals, such as necessary notation and representation theorems for vector fields (the Helmholtz decomposition theorem for spherical and the Mie representation theorem for solenoidal vector fields) are presented. In section 3 we recapitulate how the Mie representation can be applied in satellite magnetometry in order to interpret different source terms and their geomagnetic effects. In section 4 scalar and vectorial scaling functions and wavelets for the analysis of square-integrable scalar and vectorial spherical functions are introduced. In section 5 the Helmholtz decomposition theorem is utilized to combine the wavelet techniques and the Mie representation of the geomagnetic field into what is called the *wavelet Mie representation*. The resulting method of data analysis is illustrated in section 6, where the wavelet Mie representation is used to calculate radial ionospheric current distributions from the toroidal geomagnetic contributions extracted from MAGSAT and CHAMP vectorial data sets.

2. Fundamentals.

2.1. Notation and preliminaries. In order to avoid notational complications we will, unless stated otherwise, use the following scheme: Scalar fields will be denoted by capital roman letters (F, G, etc.), while vector fields are symbolized by lower-case roman letters (f, g, etc.).

A sphere of radius R centered in the origin, i.e., the set $\{x \in \mathbb{R}^3 : |x| = R\}$, will be denoted by Ω_R. In particular, $\Omega(= \Omega_1)$ is the unit sphere in $\mathbb{R}^3$. A spherical shell with inner radius R_1 and outer radius R_2, $R_2 > R_1 > 0$, is given by $\Omega_{(R_1,R_2)} = \{x \in \mathbb{R}^3 : R_1 \le |x| \le R_2\}$. Any element $x \in \mathbb{R}^3$ with $|x| \neq 0$ may be written in the form $x = r\xi$, where $r = |x|$ and $\xi = \frac{x}{r} \in \Omega$, $\xi = (\xi_1, \xi_2, \xi_3)^T$, is the uniquely determined directional unit vector of x. Using this separation, the gradient ∇ in $\mathbb{R}^3$ reads

$$\nabla_x = \xi \frac{\partial}{\partial r} + \frac{1}{r} \nabla^*_\xi,$$

where the horizontal part ∇^* is the *surface gradient* on the unit sphere Ω. Moreover, the Laplace operator $\Delta = \nabla \cdot \nabla$ in $\mathbb{R}^3$ has the representation

$$\Delta_x = \left(\frac{\partial}{\partial r} \right)^2 + \frac{2}{r} \frac{\partial}{\partial r} + \frac{1}{r^2} \Delta^*_\xi,$$

where Δ^* is the *Beltrami operator* on the unit sphere Ω. The *surface curl gradient* L^* on Ω can be calculated from ∇^* by the relation $L^*_\xi = \xi \wedge \nabla^*_\xi$, $\xi \in \Omega$ (where "$\wedge$" denotes the usual cross product).

A function is said to be of class $\mathcal{C}^{(k)}(\Omega_R)$, $0 \le k < \infty$, if it possesses k continuous derivatives on Ω_R. The set $c^{(k)}(\Omega_R)$, $0 \le k < \infty$, denotes the space of k-times continuously differentiable vector fields on Ω_R. The Hilbert spaces of measurable square-integrable scalar and vector fields on the sphere Ω_R are denoted by $\mathcal{L}^2(\Omega_R)$

and $l^2(\Omega_R)$, respectively. Let $H_n : \mathbb{R}^3 \to \mathbb{R}$ be a homogeneous harmonic polynomial of degree n; then the restriction $Y_n = H_n|_\Omega$ is called a (scalar) spherical harmonic of degree n. The space of all spherical harmonics of degree n is of dimension $2n+1$. Spherical harmonics of different degrees are orthogonal in the sense of the $\mathcal{L}^2(\Omega)$-inner product

$$(Y_n, Y_m)_{\mathcal{L}^2(\Omega)} = \int_\Omega Y_n(\xi) Y_m(\xi) d\omega(\xi) = 0, \quad n \neq m.$$

Throughout the remainder of this work, we denote by $\{Y_{n,k}\}$, $n = 0, 1, \ldots$, $k = 1, \ldots, 2n+1$, a complete orthonormal system in the Hilbert space $\mathcal{L}^2(\Omega)$. It is obvious that $\{Y^{R_1}_{n,k}\}$, $n = 0, 1, \ldots$, $k = 1, \ldots, 2n+1$, with $Y^{R_1}_{n,k} = \frac{1}{R_1} Y_{n,k}$ denotes an $\mathcal{L}^2(\Omega_{R_1})$-orthonormal system. Let $F \in \mathcal{C}^{(0_i)}(\Omega)$; then the operators $o^{(i)}_\xi : \mathcal{C}^{(0_i)}(\Omega) \to c(\Omega)$ are given by

$$\begin{aligned} o^{(1)}_\xi F(\xi) &= \xi F(\xi), \qquad \xi \in \Omega, \\ o^{(2)}_\xi F(\xi) &= \nabla^*_\xi F(\xi), \quad \xi \in \Omega, \\ o^{(3)}_\xi F(\xi) &= L^*_\xi F(\xi), \quad \xi \in \Omega, \end{aligned}$$

where 0_i is an abbreviation given by $0_1 = 0$ and $0_i = 1$ for $i \in \{2, 3\}$. Clearly, $o^{(1)}_\xi F(\xi)$ is a radial field. From the definitions of the operators ∇^* and L^* it is easy to see that $o^{(2)}_\xi F(\xi)$ and $o^{(3)}_\xi F(\xi)$ are purely tangential. Furthermore, $o^{(2)}_\xi F(\xi)$ is curl-free, whereas $o^{(3)}_\xi F(\xi)$ is divergence-free, which is clear from $\nabla^*_\xi F(\xi)$ being a gradient- and $L^*_\xi F(\xi)$ being a curl-field. Additionally, it is not difficult to see that

$$o^{(i)}_\xi F(\xi) \cdot o^{(j)}_\xi F(\xi) = 0 \quad \text{for all } i \neq j, \quad i, j \in \{1, 2, 3\}.$$

Using a complete system of scalar spherical harmonics, we are able to introduce a complete orthonormal set $\{y^{(i)}_{n,k}\}$ of vector spherical harmonics in $l^2(\Omega)$ (e.g., see [15]):

$$y^{(i)}_{n,k} = (\mu^{(i)}_n)^{-1/2} o^{(i)} Y_{n,k}, \tag{2.1}$$

$i = 1, 2, 3$, $n \in \mathbb{N}_{0_i}$, $k = 1, \ldots, 2n+1$. The normalization factor is chosen to be

$$\mu^{(i)}_n = \begin{cases} 1 & \text{if} \quad i = 1, \\ n(n+1) & \text{if} \quad i = 2, 3. \end{cases}$$

2.2. Helmholtz decomposition and Mie representation. The wavelet Mie representations are based on two main results for the decomposition of vector fields: the Helmholtz decomposition of spherical vector fields and the Mie representation of solenoidal vector fields. We start with the *Helmholtz decomposition theorem* (cf. [15]), as follows.

THEOREM 2.1. *Let $f \in c^{(1)}(\Omega)$. Then there exist uniquely determined scalar functions $F_1 \in \mathcal{C}^{(1)}(\Omega)$ and $F_2, F_3 \in \mathcal{C}^{(2)}(\Omega)$ satisfying*

$$\int_\Omega F_i(\xi) d\omega(\xi) = 0, \quad i = 2, 3,$$

such that

$$f = \sum_{i=1}^{3} o^{(i)} F_i.$$

It should be mentioned that F_1 is just the radial projection of f, while representations for the Helmholtz scalars F_2 and F_3 are available in terms of the Green function with respect to the Beltrami operator (cf. [15]). Note that the above theorem is also valid for vector fields on Ω_R, since they are isomorphic to those on Ω.

In addition to the Helmholtz representation presented above, we will make use of the so-called Mie representation for solenoidal vector fields. A vector field f on an open subset $\mathcal{U} \subset \mathbb{R}^3$ is called solenoidal if and only if the integral $\int_S f(x) \cdot \nu(x) d\omega(x)$ vanishes for every closed surface S lying entirely in U (ν denotes the outward normal of S). Every such solenoidal vector field admits a representation in terms of two uniquely defined scalar functions by means of the *Mie representation theorem* (e.g., [1, 2, 18, 32]), which follows.

THEOREM 2.2. *Let $0 < R_1 < R_2$, and let $f : \Omega_{(R_1,R_2)} \to \mathbb{R}^3$ be a solenoidal vector field in the spherical shell $\Omega_{(R_1,R_2)}$. Then there exist unique scalar functions $P_f, Q_f : \Omega_{(R_1,R_2)} \to \mathbb{R}$ such that*

(1) $\int_{\Omega_r} P_f(x) d\omega_r(x) = \int_{\Omega_r} Q_f(x) d\omega_r(x) = 0$,

(2) $f = \nabla \wedge LP_f + LQ_f$

for all $r \in (R_1, R_2)$ with the operator L given by $L_x = x \wedge \nabla_x$.

Vector fields of the form $\nabla \wedge LP_f$ are called poloidal, while vector fields of the form LQ_f are denoted toroidal. For the sake of completeness we present the following theorem (cf. [2]).

THEOREM 2.3. *Let $0 < R_1 < R_2$, and let $f : \Omega_{(R_1,R_2)} \to \mathbb{R}^3$ be a solenoidal vector field in the spherical shell $\Omega_{(R_1,R_2)}$. Then there exist a unique poloidal field p as well as a unique toroidal field t such that*

$$f = p + t$$

in $\Omega_{(R_1,R_2)}$.

For our further considerations it is important that, for each $x = r\xi$ with $R_1 < r < R_2$ and $\xi \in \Omega$, the Mie representation $f = \nabla \wedge LP_f + LQ_f$ can be rewritten as

$$f(r\xi) = \xi \frac{\Delta_\xi^* P_f(r\xi)}{r} - \nabla_\xi^* \frac{\partial_r r P_f(r\xi)}{r} + L_\xi^* Q_f(r\xi) \tag{2.2}$$

(cf., e.g., [1, 2, 25, 29]), where we have used the abbreviation $\partial_r = \partial/\partial r$. (Actually, with regard to the second term, it is mathematically correct to write

$$\left(\frac{\partial}{\partial \tilde{r}} \tilde{r} P_f(\tilde{r}\xi) \right) \Big|_{\tilde{r}=r}.$$

We avoid this awkward notation, however, and stick to the easy nomenclature.) Note that (2.2) is the Helmholtz decomposition of the Mie representation of f and links the previously defined vector spherical harmonics to the Mie representation of vector fields.

Finally, we mention the following last result, which is concerned with the curl of a Mie representation.

COROLLARY 2.4. *Let $f, g : \Omega_{(R_1,R_2)} \to \mathbb{R}^3$ be two solenoidal vector fields with representations*

$$f = \nabla \wedge LP_f + LQ_f,$$
$$g = \nabla \wedge LP_g + LQ_g,$$

and which are connected via $\nabla \wedge f = \lambda g$, $\lambda \in \mathbb{R} \setminus \{0\}$. *Then the Mie scalars are related via*

$$P_g = \frac{1}{\lambda} Q_f,$$
$$Q_g = -\frac{1}{\lambda} \Delta P_f.$$

This shows us that the curl of a poloidal field is a toroidal field, and vice versa.

3. The (geo)magnetic field in Mie representation. If we assume the typical length- and time-scales of the magnetic field b and the electric current densities j to be such that retarding effects (and displacement currents) are negligible, then we can consider the quasi-static approximation of Maxwell's equations, the pre-Maxwell equations, to be valid:

$$\nabla \cdot b = 0,$$
$$\nabla \wedge b = \mu_0 j,$$

where μ_0 is the vacuum permeability. Since the magnetic field is divergence-free everywhere, it can be split up into a poloidal and a toroidal part (see Theorem 2.2):

$$b = b_{pol} + b_{tor} = \nabla \wedge L P_b + L Q_b. \tag{3.1}$$

The quasi-static approximation being true is equivalent to the current densities being divergence-free everywhere. Consequently the electric currents also admit a Mie representation:

$$j = j_{pol} + j_{tor} = \nabla \wedge L P_j + L Q_j. \tag{3.2}$$

According to Corollary 2.4 the Mie scalars of the magnetic field and the electric currents are related via

$$P_j = \frac{1}{\mu_0} Q_b, \tag{3.3}$$

$$Q_j = -\frac{1}{\mu_0} \Delta P_b. \tag{3.4}$$

Using (2.2), we can, for each $x = r\xi$ with $r \neq 0$ and $\xi \in \Omega$, rewrite (3.1) and (3.2) as

$$b = \xi \frac{\Delta_\xi^* P_b}{r} - \nabla_\xi^* \frac{\partial_r r P_b}{r} + L_\xi^* Q_b \tag{3.5}$$

and

$$j = \xi \frac{\Delta_\xi^* P_j}{r} - \nabla_\xi^* \frac{\partial_r r P_j}{r} + L_\xi^* Q_j. \tag{3.6}$$

The first two terms in (3.5) and (3.6) can be interpreted as the restriction of the poloidal magnetic field b_{pol} and the poloidal currents j_{pol}, respectively, to the sphere Ω_r. The last terms represent the toroidal field b_{tor} and currents j_{tor} on Ω_r. These equations will serve as a starting point for the wavelet Mie representations in section 5.

Following Backus [1], Engels and Olsen [11], and Maus [25], we assume either the geomagnetic field b or the electric current distributions j to be sampled within

a spherical shell $\Omega_{(R_1,R_2)}$, $0 < R_1 < R_2 < \infty$. This assumption takes into account elliptical satellite orbits as well as the decrease in altitude with the lifetime of the satellite. The geomagnetic field within the shell $\Omega_{(R_1,R_2)}$ consists of four different parts (cf. [29]), i.e.,

$$b = b_{pol}^{int} + b_{pol}^{ext} + b_{pol}^{sh} + b_{tor}.$$

b_{pol}^{int} denotes the poloidal magnetic field due to internal toroidal currents in the region with $r < R_1$. b_{pol}^{ext} is the poloidal part caused by external toroidal current densities in the region with $r > R_2$, and b_{pol}^{sh} is the poloidal magnetic field due to the toroidal electric currents within $\Omega_{(R_1,R_2)}$. Finally, b_{tor} is the toroidal part of b generated by poloidal currents in $\Omega_{(R_1,R_2)}$. If there are no currents in the shell $\Omega_{(R_1,R_2)}$, then $b_{pol}^{sh} = b_{tor} = 0$, and b can be represented as the gradient-field of a scalar harmonic potential or by means of the Mie representation equivalently. If only the toroidal currents vanish within the shell, then $b_{pol}^{sh} = 0$, and the magnetic field within the shell can be represented by

$$b = b_{pol}^{int} + b_{pol}^{ext} + b_{tor}. \tag{3.7}$$

The situation changes if the toroidal currents within the shell $\Omega_{(R_1,R_2)}$ do not vanish. Let us suppose that the radii of the shell satisfy

$$R_2 - R_1 \ll \frac{R_2 + R_1}{2},$$

i.e., the thickness of the shell is small compared to the mean radius. Such a shell is called a *thin shell*. As pointed out by Backus [1] and Olsen [29], even for non-vanishing (toroidal) current densities in the shell, the magnetic field within a thin shell can (approximately) be represented by (3.7), i.e., the poloidal field b_{pol}^{sh} tends to zero in thin shells, while the toroidal part b_{tor} remains finite. Actually, for thin shells, it holds that $b_{pol}^{sh} \to 0$ as $(R_2 - R_1)/H \to 0$, where H is a reference length characterizing the vertical scale of the current density (e.g., [1], [29]). In more detail, if in a thin shell,

$$R_2 - R_1 \ll H \simeq \frac{R_2 + R_1}{2},$$

i.e., the current density changes significantly on vertical scales that can be compared to the mean radius and that are much larger than the thickness of the shell, then the thin shell approximation (3.7) is surely valid. If, in a thin shell,

$$R_2 - R_1 \simeq H \ll \frac{R_2 + R_1}{2},$$

i.e., the currents change significantly on vertical length scales that are small compared to the mean radius but that can be compared to the thickness of the shell, then the thin shell approximation can fail as well. For more details the interested reader is directed to [1]. In what follows we assume the thin shell approximation to be valid (which is a reasonable assumption for the examples presented in section 6; see, e.g., [29] and [25]).

At this point, there remains the question of how to numerically obtain—in terms of suitable trial functions—the Mie representation of a given set of vectorial data.

As we have already mentioned, the global support of the spherical harmonics limits the practicability of spherical harmonic parametrizations since most of the relevant ionospheric currents vary rapidly with latitude and longitude and/or are confined to certain regions. Consequently, it seems reasonable to find a field parametrization in terms of functions that take efficient account of the specific concentration of the current densities in space. In [4] we have already presented first methods to deal with the Mie representation in terms of space-localizing trial functions, so-called spherical vectorial wavelets, which are able to reflect various levels of space localization (see also [3]). The techniques developed in section 5 are generalizations and enhancements of this approach. For a complete and comprehensive description, the interested reader is directed to the thesis [24].

4. Scaling functions and wavelets in $\mathcal{L}^2(\Omega)$ and $l^2(\Omega)$. As far as this article is concerned, it suffices to introduce scaling functions and wavelets for the spaces of square-integrable scalar and vector fields on the unit sphere, i.e., $\mathcal{L}^2(\Omega)$ and $l^2(\Omega)$. This theory is well known since, starting from classical wavelet theory (see, e.g., [7] and [6] for an overview), the concept of multiresolution has been adapted to spherical geometries for scalar fields by, e.g., Freeden and Windheuser [16], [17] and, for vector fields, by Bayer, Beth, and Freeden [3] and Freeden, Gervens, and Schreiner [15], for example. We therefore just repeat some results which are useful for our further considerations as follows.

DEFINITION 4.1. *A real sequence $\{(\Phi_J)^\wedge(n)\}$, $J \in \mathbb{Z}$, $n \in \mathbb{N}_0$, is called a generator (or symbol) of an $\mathcal{L}^2(\Omega)$-scaling function if it satisfies*

(i) $\sum_{n=0}^{\infty} \left((\Phi_J)^\wedge(n)\right)^2 < \infty$,
(ii) $\sum_{n=0}^{\infty} \left((\Phi_J)^\wedge(n) Y_{n,k}(\xi)\right)^2 < \infty$ *for all* $\xi \in \Omega$,
(iii) $\lim_{J\to\infty} \left((\Phi_J)^\wedge(n)\right)^2 = 1$, $n \in \mathbb{N}$,
(iv) $\left((\Phi_J)^\wedge(n)\right)^2 \geq \left((\Phi_{J-1})^\wedge(n)\right)^2$,
(v) $\lim_{J\to-\infty} \left((\Phi_J)^\wedge(n)\right)^2 = 0$,
(vi) $\left((\Phi_J)^\wedge(0)\right)^2 = 1$, $J \in \mathbb{Z}$.

The corresponding family $\{\Phi_J\}$ of kernels given by

$$\Phi_J(\xi,\eta) = \sum_{n=0}^{\infty} \sum_{k=1}^{2n+1} (\Phi_J)^\wedge(n) Y_{n,k}(\xi) Y_{n,k}(\eta), \qquad \xi,\eta \in \Omega,$$

is called $\mathcal{L}^2(\Omega)$-scaling function.

Wavelets come into play via the refinement equation, as in the following.

DEFINITION 4.2. *The real sequence $\{(\Psi_J)^\wedge(n)\}$, $J \in \mathbb{Z}$, $n \in \mathbb{N}_0$, defined via the refinement equation*

$$(\Psi_J)^\wedge(n) = \left(\left((\Phi_{J+1})^\wedge(n)\right)^2 - \left((\Phi_J)^\wedge(n)\right)^2\right)^{\frac{1}{2}},$$

is called the generator (or symbol) of the $\mathcal{L}^2(\Omega)$-wavelet $\{\Psi_J\}$ given as

$$\Psi_J(\xi,\eta) = \sum_{n=0}^{\infty} \sum_{k=1}^{2n+1} (\Psi_J)^\wedge(n) Y_{n,k}(\xi) Y_{n,k}(\eta), \qquad \xi,\eta \in \Omega.$$

The concept of scaling functions and wavelets can also be carried over to the vectorial case as in the following.

DEFINITION 4.3. *Let, for $i \in \{1,2,3\}$, the real sequence $\{(\varphi_J^{(i)})^\wedge(n)\}$, $J \in \mathbb{Z}$, $n \in \mathbb{N}_{0_i}$, be generators of $\mathcal{L}^2(\Omega)$-scaling functions; then the vectorial kernels*

$$\varphi_J^{(i)}(\xi,\eta) = \sum_{n=0_i}^{\infty} \sum_{k=1}^{2n+1} (\varphi_J^{(i)})^\wedge(n) Y_{n,k}(\xi) y_{n,k}^{(i)}(\eta) \tag{4.1}$$

are called $l^2(\Omega)$-scaling functions of type i. The $l^2(\Omega)$-wavelets of type i are given by

$$\psi_J^{(i)}(\xi,\eta) = \sum_{n=0_i}^{\infty} \sum_{k=1}^{2n+1} (\psi_J^{(i)})^\wedge(n) Y_{n,k}(\xi) y_{n,k}^{(i)}(\eta), \tag{4.2}$$

with generators $\{(\psi_J^{(i)})^\wedge(n)\}$ satisfying the refinement equation

$$(\psi_J^{(i)})^\wedge(n) = \left(\left((\varphi_{J+1}^{(i)})^\wedge(n) \right)^2 - \left((\varphi_J^{(i)})^\wedge(n) \right)^2 \right)^{\frac{1}{2}}.$$

With these definitions at hand, we can find approximations of $\mathcal{L}^2(\Omega)$ and $l^2(\Omega)$ functions in terms of the respective scaling functions and wavelets (e.g., [16]).

THEOREM 4.4. *Let the families $\{\Phi_J\}$, $\{\Psi_J\}$ be $\mathcal{L}^2$-scaling functions and wavelets. For any $F \in \mathcal{L}^2(\Omega)$ it holds that*

$$F = \Phi_{J'} * \Phi_{J'} * F + \sum_{J=J'}^{\infty} \Psi_J * \Psi_J * F \tag{4.3}$$

$$= \Phi_0 * \Phi_0 * F + \sum_{J=0}^{\infty} \Psi_J * \Psi_J * F, \tag{4.4}$$

where the convolution operator "$*$" for scalar kernels and functions is defined by

$$K * F = \int_\Omega K(\cdot,\eta) F(\eta)\, d\omega(\eta).$$

In the case of vector fields $f \in l^2(\Omega)$ it is possible to show the following (cf. [3]).

THEOREM 4.5. *Let the families $\{\varphi_J^{(i)}\}$, $\{\psi_J^{(i)}\}$ be vectorial scaling functions and wavelets. Then, for $f \in l^2(\Omega)$,*

$$f = \sum_{i=1}^{3} \varphi_{J'}^{(i)} \star \varphi_{J'}^{(i)} * f + \sum_{J=J'}^{\infty} \sum_{i=1}^{3} \psi_J^{(i)} \star \psi_J^{(i)} * f \tag{4.5}$$

$$= \sum_{i=1}^{3} \varphi_0^{(i)} \star \varphi_0^{(i)} * f + \sum_{J=0}^{\infty} \sum_{i=1}^{3} \psi_J^{(i)} \star \psi_J^{(i)} * f, \tag{4.6}$$

where the convolution "$*$" of a vectorial kernel against a vector field is given as

$$k * f = \int_\Omega k(\cdot,\eta) \cdot f(\eta)\, d\omega(\eta),$$

and the convolution "$\star$" of a vectorial kernel against a scalar field is

$$k \star F = \int_\Omega k(\eta,\cdot) F(\eta)\, d\omega(\eta).$$

The representations of square-integrable scalar and vectorial functions in terms of scaling functions and wavelets build one of the fundamentals for our considerations in the next section. It is noteworthy that the vectorial wavelets are defined in correspondence to the vector spherical harmonics in section 2.1 and can therefore be linked to the Mie representation via (2.2), (3.5), and (3.6), which is the task of the next section.

It should be remarked that, in the case of F being one of the Mie-scalars, the first term in (4.4) vanishes, since the Mie-scalars have vanishing zeroth order moment. A similar argument holds true for the case of f being the magnetic field or the electric current density; i.e., since both are of zero divergence, the first term in (4.6) vanishes.

Before we go on, we mention some properties of the above kernel functions (scaling functions and wavelets) which are important from a numerical point of view. Let K and $k^{(i)}$ be either scalar scaling functions or wavelets, or vectorial scaling functions or wavelets, respectively. Using the addition theorem for spherical harmonics, each scalar kernel K admits the following representation:

$$\begin{aligned} K(\xi,\eta) &= \sum_{n=0}^{\infty}\sum_{k=1}^{2n+1} (K)^\wedge(n) Y_{n,k}(\xi) Y_{n,k}(\eta) \\ &= \sum_{n=0}^{\infty} (K)^\wedge(n) \frac{2n+1}{4\pi} P_n(\xi\cdot\eta), \end{aligned}$$

where P_n is the Legendre polynomial of degree n. For the evaluation of such series of Legendre polynomials there exist fast and stable recursive algorithms (e.g., [8]).

In the case of vectorial kernels, the situation is only slightly more complicated. From the definition of the vector spherical harmonics and the vectorial kernel functions (see (2.1), (4.1), and (4.2)) we see that

$$k^{(i)}(\xi,\eta) = \sum_{n=0_i}^{\infty}\sum_{k=1}^{2n+1} (k^{(i)})^\wedge(n)\, Y_{n,k}(\eta)\, (\mu_n^{(i)})^{-\frac{1}{2}}\, o_\xi^{(i)} Y_{n,k}(\xi) \tag{4.7}$$

$$= o_\xi^{(i)} \sum_{n=0_i}^{\infty}\sum_{k=1}^{2n+1} (k^{(i)})^\wedge(n)\, Y_{n,k}(\eta)\, (\mu_n^{(i)})^{-\frac{1}{2}}\, Y_{n,k}(\xi) \tag{4.8}$$

$$= o_\xi^{(i)} \sum_{n=0_i}^{\infty} \frac{2n+1}{4\pi}\, (k^{(i)})^\wedge(n)\, (\mu_n^{(i)})^{-\frac{1}{2}}\, P_n(\xi\cdot\eta). \tag{4.9}$$

For $\eta\in\Omega$ fixed, the Legendre polynomials are isotropic functions on the unit sphere and the $o^{(i)}$ can be applied. This results in

$$o_\xi^{(1)} P_n(\xi\cdot\eta) = \xi P_n(\xi\cdot\eta), \tag{4.10}$$

$$o_\xi^{(2)} P_n(\xi\cdot\eta) = (\eta - (\xi\cdot\eta)\,\xi)\; P_n'(\xi\cdot\eta), \tag{4.11}$$

$$o_\xi^{(3)} P_n(\xi\cdot\eta) = (\xi\wedge\eta)\, P_n'(\xi\cdot\eta)\,. \tag{4.12}$$

Using this, the kernels in (4.9) admit the representation

$$k^{(1)}(\xi,\eta) = \xi \sum_{n=0}^{\infty} \frac{2n+1}{4\pi}\, (k^{(i)})^\wedge(n)\, (\mu_n^{(i)})^{-\frac{1}{2}}\, P_n(\xi\cdot\eta), \tag{4.13}$$

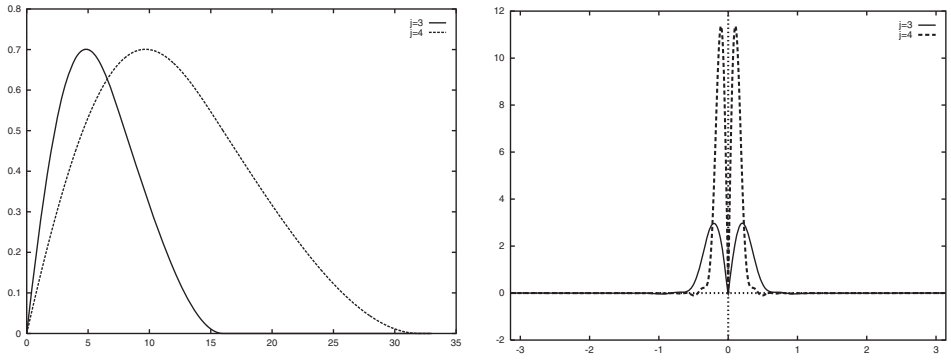


FIG. 1. *Left: Generators of CuP wavelets of scales* 3 *and* 4. *Right: Cross section through the corresponding CuP wavelets of scales* 3 *and* 4. *Note that an increasing scale leads to a decreasing localization in the Fourier domain (generator) but to an increasing localization in the space domain.*

$$k^{(2)}(\xi,\eta) = (\eta - (\xi\cdot\eta)\,\xi)\sum_{n=1}^{\infty}\frac{2n+1}{4\pi}\,(k^{(i)})^{\wedge}(n)\,(\mu_n^{(i)})^{-\frac{1}{2}}\,P_n'(\xi\cdot\eta), \tag{4.14}$$

$$k^{(3)}(\xi,\eta) = (\xi\wedge\eta)\sum_{n=1}^{\infty}\frac{2n+1}{4\pi}\,(k^{(i)})^{\wedge}(n)\,(\mu_n^{(i)})^{-\frac{1}{2}}\,P_n'(\xi\cdot\eta) \tag{4.15}$$

such that the fast and stable one-dimensional recursive algorithms can also be used for calculating the vectorial kernels. It should be noted that, if the kernel functions are non–band-limited (nondegenerate), the sums in the above equations need to be truncated if no analytic representations for the kernels are known.

For later use we present, as a certain choice of possible kernels, the vectorial cubic polynomial (CuP) wavelets, which can be derived by using a generator of the form

$$\varphi_0^{(i)}(x) = \begin{cases} (1-x)^2(1+2x), & x\in[0,1),\\ 0, & x\in[1,\infty). \end{cases}$$

The left-hand side of Figure 1 shows generators of CuP wavelets of scales 3 and 4, while the right-hand side presents cross sections of the corresponding CuP wavelets. It is obvious that—with increasing scale—the localization in the Fourier domain (generators) decreases, while the localization in the space domain (wavelets) increases. Figure 2 provides illustrations of tangential CuP vectorial wavelets in the longitude-latitude plane. Note again that the significant support of the wavelets decreases with increasing scale, a feature typical for wavelets. It is this property that, via the wavelet Mie representation, allows for the analysis and modeling of spatially confined structures in the geomagnetic field and the corresponding current distributions.

5. Wavelet Mie parametrizations. In what follows we restrict ourselves to the wavelet parametrization of toroidal magnetic fields and the corresponding poloidal electric current densities in the spherical shell $\Omega_{(R_1,R_2)}$. This case is sufficient for the applications presented in section 6. For details on the wavelet parametrization of poloidal magnetic fields, the reader may consult our treatise in [24]. This approach, however, requires the introduction of inner and outer harmonic wavelets (e.g., [13]), which is beyond the scope of this article.

The starting point for our considerations is a separation of variables for the

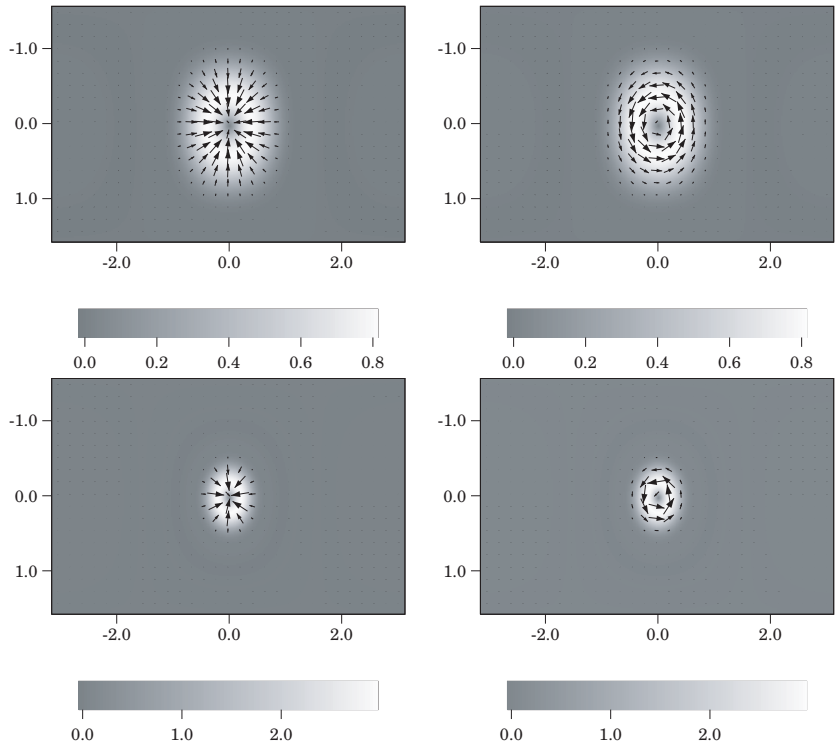


FIG. 2. *CuP wavelets in the longitude-latitude plane. Arrows indicate direction and shading indicates magnitude. Top: Curl-free ($i = 2$, left) and divergence-free ($i = 3$, right) CuP wavelets at scale $j = 2$. Bottom: Curl-free ($i = 2$, left) and divergence-free ($i = 3$, right) CuP wavelets at scale $j = 3$. Note how the increasing scale leads to a sharper localization in the space domain.*

toroidal field scalar Q_b; i.e., we assume that

$$Q_b(r\xi) = Q_{b,1}(r)Q_{b,2}(\xi) \quad \text{in } \Omega_{(R_1,R_2)}. \tag{5.1}$$

Relation (3.3) suggests proceeding likewise in the case of the scalar P_j for the poloidal currents, and hence we suppose that

$$\begin{aligned} P_j(r\xi) &= P_{j,1}(r)P_{j,2}(\xi) \\ &= \frac{1}{\mu_0} Q_{b,1}(r)Q_{b,2}(\xi) \quad \text{in } \Omega_{(R_1,R_2)}. \end{aligned}$$

The results of section 4 yield that the angular parts $Q_{b,2}$ and $P_{j,2}$ can be expanded in terms of scalar spherical $\mathcal{L}^2(\Omega)$-wavelets $\{\Psi_J\}$; i.e.,

$$Q_{b,2} = \sum_{J=0}^{\infty} \Psi_J * \Psi_J * Q_{b,2}, \tag{5.2}$$

$$P_{j,2} = \sum_{J=0}^{\infty} \Psi_J * \Psi_J * P_{j,2}. \tag{5.3}$$

Combining this with (3.1)–(3.2) and (3.5)–(3.6), we can come up with the following representations for the toroidal magnetic field and the corresponding poloidal current density.

THEOREM 5.1. *Let, for $J \in \mathbb{Z}$, $\{\Psi_J\}$ be an $\mathcal{L}^2(\Omega)$-wavelet. Under the assumptions above, the toroidal magnetic field in $\Omega_{(R_1,R_2)}$ can be represented via*

$$b_{tor}(r\cdot) = Q_{b,1}(r)\left(\bar{\varphi}^{(3)}_{J'} \star \Phi_{J'} * Q_{b,2} + \sum_{J=J'}^{\infty} \bar{\psi}^{(3)}_J \star \Psi_J * Q_{b,2}\right) \tag{5.4}$$

$$= Q_{b,1}(r)\left(\bar{\varphi}^{(3)}_0 \star \Phi_0 * Q_{b,2} + \sum_{J=0}^{\infty} \bar{\psi}^{(3)}_J \star \Psi_J * Q_{b,2}\right) \tag{5.5}$$

$$= Q_{b,1}(r) \sum_{J=0}^{\infty} \bar{\psi}^{(3)}_J \star \Psi_J * Q_{b,2}, \tag{5.6}$$

where the vectorial kernels $\bar{\varphi}^{(3)}_J$ and $\bar{\psi}^{(3)}_J$ are given via $\bar{\varphi}^{(3)}_J(\xi,\eta) = L^*_\xi \Phi_J(\xi,\eta)$ and $\bar{\psi}^{(3)}_J(\xi,\eta) = L^*_\xi \Psi_J(\xi,\eta)$. The corresponding poloidal current density in $\Omega_{(R_1,R_2)}$ is given by

$$\mu_0 j_{pol}(r\cdot) = \frac{1}{r} Q_{b,1}(r)\left(\tilde{\varphi}^{(1)}_{J'} \star \Phi_{J'} * Q_{b,2} + \sum_{J=J'}^{\infty} \tilde{\psi}^{(1)}_J \star \Psi_J * Q_{b,2}\right) \tag{5.7}$$

$$+ \left(\partial_r + \frac{1}{r}\right) Q_{b,1}(r)\left(\hat{\varphi}^{(2)}_{J'} \star \Phi_{J'} * Q_{b,2} + \sum_{J=J'}^{\infty} \hat{\psi}^{(2)}_J \star \Psi_J * Q_{b,2}\right) \tag{5.8}$$

$$= \frac{1}{r} Q_{b,1}(r)\left(\tilde{\varphi}^{(1)}_0 \star \Phi_0 * Q_{b,2} + \sum_{J=0}^{\infty} \tilde{\psi}^{(1)}_J \star \Psi_J * Q_{b,2}\right) \tag{5.9}$$

$$+ \left(\partial_r + \frac{1}{r}\right) Q_{b,1}(r)\left(\hat{\varphi}^{(2)}_0 \star \Phi_0 * Q_{b,2} + \sum_{J=0}^{\infty} \hat{\psi}^{(2)}_J \star \Psi_J * Q_{b,2}\right) \tag{5.10}$$

$$= \frac{1}{r} Q_{b,1}(r) \sum_{J=0}^{\infty} \tilde{\psi}^{(1)}_J \star \Psi_J * Q_{b,2} \tag{5.11}$$

$$+ \left(\partial_r Q_{b,1}(r) + \frac{1}{r} Q_{b,1}(r)\right) \sum_{J=0}^{\infty} \hat{\psi}^{(2)}_J \star \Psi_J * Q_{b,2}, \tag{5.12}$$

where the kernel functions $\tilde{\varphi}^{(1)}_J$ and $\hat{\varphi}^{(2)}_J$ as well as $\tilde{\psi}^{(1)}_J$ and $\hat{\psi}^{(2)}_J$ are defined to be $\tilde{\varphi}^{(1)}_J(\xi,\eta) = \xi \Delta^*_\xi \Phi_J(\xi,\eta)$ and $\hat{\varphi}^{(2)}_J(\xi,\eta) = -\nabla^*_\xi \Phi_J(\xi,\eta)$, as well as $\tilde{\psi}^{(1)}_J(\xi,\eta) = \xi \Delta^*_\xi \Psi_J(\xi,\eta)$ and $\hat{\psi}^{(2)}_J(\xi,\eta) = -\nabla^*_\xi \Psi_J(\xi,\eta)$.

Proof. Equation (5.4) follows from (3.1), (3.5), and (5.2). Theorems 4.4 and 4.5 lead to (5.5). The fact that the magnetic field is of zero divergence everywhere implies—via the Gauss theorem—that the magnetic field has vanishing zeroth order moment (i.e., the magnetic field is solenoidal), which means that

$$Q_{b,1}(r)\left(\bar{\varphi}^{(3)}_0 \star \Phi_0 * Q_{b,2}\right) = 0.$$

Equations (5.7) and (5.8) follow from (3.2), (3.6), and (5.3) in combination with (3.4). Theorems 4.4 and 4.5 then imply (5.9) and (5.10). Since in the pre-Maxwell approximation the current density is solenoidal, too, it follows that

$$\frac{1}{r} Q_{b,1}(r)\left(\tilde{\varphi}^{(1)}_0 \star \Phi_0 * Q_{b,2}\right) = 0$$

and

$$\left(\partial_r + \frac{1}{r}\right) Q_{b,1}(r) \left(\hat{\varphi}_0^{(2)} \star \Phi_0 * Q_{b,2}\right) = 0. \qquad \square$$

Note that all of the kernel functions that appear in Theorem 5.1 can be calculated using the rules and results of (4.10)–(4.15), as well as the fact that scalar spherical harmonics of degree n are eigenfunctions of the Beltrami operator with respect to eigenvalues $-n(n+1)$.

Theorem 5.1 presents the wavelet Mie representation of the toroidal magnetic field and the corresponding poloidal electric currents in the spherical shell $\Omega_{(R_1,R_2)}$. Due to the space localization of the ansatz functions, this representation yields the possibility of using or deriving different models of Q_b in different regions, depending on the underlying physical effects and, of course, the data situation.

The ansatz (5.1) is quite simple and might fail if the radial dependency is very complex (see also the considerations in [25]). Nevertheless, assumption (5.1) is reasonable as long as the data situation is such that the radial behavior of the field is difficult to extract. This is arguably the case when using data from single satellite missions. (See also the comments in [1], [29], and [25] concerning time-variations and single satellite missions.) Nevertheless, if the data situation allows for determination of higher order radial dependencies (e.g., if data from multisatellite missions are used, or if measurements from satellites are combined with terrestrial observations), we might expand our ansatz by adding further toroidal scalars with different radial behavior (cf. [24]).

The product ansatz for the toroidal field scalar Q_b is reflected in the corresponding toroidal magnetic field as well as in the representation of the corresponding poloidal current density. As regards the poloidal current, both its radial and its tangential parts admit a product representation, too. In more detail, let j_{rad} and j_{∇^*} be the radial and the tangential parts, respectively, of j_{pol}. Then (5.11) and (5.12) of Theorem 5.1 show that j_{rad} and j_{∇^*} can be represented as

$$j_{rad}(r\xi) = J_{rad,1}(r) j_{rad,2}(\xi)$$

and

$$j_{\nabla^*}(r\xi) = J_{\nabla^*,1}(r) j_{\nabla^*,2}(\xi),$$

where the scalar functions $J_{rad,1}(r)$ and $J_{\nabla^*,1}(r)$ are given via

$$\begin{aligned} \mu_0 J_{rad,1}(r) &= \frac{1}{r} Q_{b,1}(r), \\ \mu_0 J_{\nabla^*,1}(r) &= \left(\partial_r Q_{b,1}(r) + \frac{1}{r} Q_{b,1}(r)\right) \end{aligned}$$

and the vectorial parts are

$$\begin{aligned} \mu_0 j_{rad,2} &= \sum_{J=0}^{\infty} \tilde{\psi}_J^{(1)} \star \Psi_J * Q_{b,2}, \\ \mu_0 j_{\nabla^*,2} &= \sum_{J=0}^{\infty} \hat{\psi}_J^{(2)} \star \Psi_J * Q_{b,2}. \end{aligned}$$

Using the ansatz (5.1) together with (3.6) immediately leads us to the same results for $J_{rad,1}$ and $J_{\nabla^*,1}$ but, with regard to $j_{rad,2}$ and $j_{\nabla^*,2}$, we end up with

$$\mu_0 j_{rad,2}(\xi) = \xi \Delta_\xi^* Q_{b,2}(\xi),$$

$$\mu_0 j_{\nabla^*,2}(\xi) = -\nabla_\xi^* Q_{b,2}(\xi),$$

which is independent from any parametrization of Q_b. Nevertheless, we know from section 4 that we can expand the radial vector field $\mu_0 j_{rad,2}$ and the tangential vector field $\mu_0 j_{\nabla^*,2}$ using vectorial $l^2(\Omega)$-wavelets $\{\psi_J^{(i)}\}$ of type $i = 1$ and $i = 2$, respectively. Consequently we are led to the following alternative representation in terms of $l^2(\Omega)$-wavelets.

COROLLARY 5.2. *Let the families $\{\varphi_J^{(i)}\}$, $\{\psi_J^{(i)}\}$, $i = 1, 2$, be vectorial scaling functions and wavelets. The radial part j_{rad} and tangential part j_{∇^*} of the poloidal current density can be represented via*

$$j_{rad}(r\cdot) = \varphi_{J'}^{(1)} \star \left(\varphi_{J'}^{(1)} * j_{rad}\right)(r) + \sum_{J=J'}^{\infty} \psi_J^{(1)} \star \left(\psi_J^{(1)} * j_{rad}\right)(r) \tag{5.13}$$

$$= \varphi_0^{(1)} \star \left(\varphi_0^{(1)} * j_{rad}\right)(r) + \sum_{J=0}^{\infty} \psi_J^{(1)} \star \left(\psi_J^{(1)} * j_{rad}\right)(r) \tag{5.14}$$

$$= \sum_{J=0}^{\infty} \psi_J^{(1)} \star \left(\psi_J^{(1)} * j_{rad}\right)(r) \tag{5.15}$$

$$= \frac{1}{r} Q_{b,1}(r) \sum_{J=0}^{\infty} \psi_J^{(1)} \star \psi_J^{(1)} * j_{rad,2} \tag{5.16}$$

$$= \sum_{J=0}^{\infty} \psi_J^{(1)} \star \left(\psi_J^{(1)} * j\right)(r) \tag{5.17}$$

and

$$j_{\nabla^*}(r\cdot) = \varphi_{J'}^{(2)} \star \left(\varphi_{J'}^{(2)} * j_{rad}\right)(r) + \sum_{J=J'}^{\infty} \psi_J^{(2)} \star \left(\psi_J^{(2)} * j_{rad}\right)(r) \tag{5.18}$$

$$= \varphi_0^{(2)} \star \left(\varphi_0^{(2)} * j_{rad}\right)(r) + \sum_{J=0}^{\infty} \psi_J^{(2)} \star \left(\psi_J^{(2)} * j_{rad}\right)(r) \tag{5.19}$$

$$= \sum_{J=0}^{\infty} \psi_J^{(2)} \star \left(\psi_J^{(2)} * j_{\nabla^*}\right)(r) \tag{5.20}$$

$$= \left(\partial_r + \frac{1}{r}\right) Q_{b,1}(r) \sum_{J=0}^{\infty} \psi_J^{(2)} \star \psi_J^{(2)} * j_{\nabla^*,2} \tag{5.21}$$

$$= \sum_{J=0}^{\infty} \psi_J^{(2)} \star \left(\psi_J^{(2)} * j\right)(r). \tag{5.22}$$

Note that (5.17) and (5.22) are true since only the poloidal current density contains a radial or ∇^*-contribution (see (3.6)). In other words, on each Ω_r with $R_1 < r < R_2$, the radial current density can be derived from expanding the total current density in terms of spherical vectorial wavelets of type $i = 1$, while the tangential

part of the poloidal current density can be calculated via spherical vectorial wavelets of type $i = 2$. Equations (5.15)–(5.22) can therefore be used to determine the toroidal field scalar or, of course, the corresponding toroidal magnetic field.

A similar approach can be applied in order to determine the poloidal current density j_{pol} in $\Omega_{(R_1,R_2)}$ from the corresponding toroidal field b_{tor}. Assuming the product ansatz for Q_b and applying (3.5), we see that the toroidal magnetic field admits a product representation as well, i.e.,

$$b_{tor}(r\xi) = B_{tor,1}(r)b_{tor,2}(\xi),$$

where $b_{tor,2} = L^*Q_{b,2}$ can be expressed in terms of spherical vectorial $l^2(\Omega)$-wavelets $\{\psi_J^{(3)}\}$ of type $i = 3$ as follows:

$$b_{tor,2} = \sum_{J=0}^{\infty} \psi_J^{(3)} \star \psi_J^{(3)} * b_{tor,2}.$$

From our previous results we know that the scalar $B_{tor,1}$ is just given by

$$B_{tor,1}(r) = Q_{b,1}(r).$$

Since the toroidal magnetic field b_{tor} is the only part of b that contributes an L^*-portion, it is clear that

$$\begin{aligned} b_{tor}(r\cdot) &= \sum_{J=0}^{\infty} \psi_J^{(3)} \star \left(\psi_J^{(3)} * b_{tor}\right)(r) \\ &= \sum_{J=0}^{\infty} \psi_J^{(3)} \star \left(\psi_J^{(3)} * b\right)(r) \end{aligned}$$

on any sphere Ω_r with $R_1 < r < R_2$. Summarizing the above considerations, we are led to the following claim.

COROLLARY 5.3. *Let the families $\{\varphi_J^{(3)}\}$ and $\{\psi_J^{(3)}\}$ be vectorial scaling functions and wavelets of type* 3. *The toroidal magnetic field b_{tor} can be represented via*

$$b_{tor}(r\cdot) = Q_{b,1}(r)\left(\varphi_{J'}^{(3)} \star \varphi_{J'}^{(3)} * b_{tor,2} + \sum_{J=J'}^{\infty} \psi_J^{(3)} \star \psi_J^{(3)} * b_{tor,2}\right) \tag{5.23}$$

$$= Q_{b,1}(r)\left(\varphi_0^{(3)} \star \varphi_0^{(3)} * b_{tor,2} + \sum_{J=0}^{\infty} \psi_J^{(3)} \star \psi_J^{(3)} * b_{tor,2}\right) \tag{5.24}$$

$$= Q_{b,1}(r) \sum_{J=0}^{\infty} \psi_J^{(3)} \star \psi_J^{(3)} * b_{tor,2} \tag{5.25}$$

$$= \sum_{J=0}^{\infty} \psi_J^{(3)} \star \left(\psi_J^{(3)} * b\right)(r) \tag{5.26}$$

on any sphere Ω_r with $R_1 < r < R_2$.

This yields one possible way of determining the poloidal field scalar (and consequently the corresponding poloidal electric current density) from magnetic measurements in $\Omega_{(R_1,R_2)}$.

Assuming that the data are given only at constant altitude (or with negligible radial dependencies), the previous approach can be easily applied to calculate radial

current densities on a sphere Ω_r, with $R_1 < r < R_2$, from measurements of the magnetic field on that very sphere. We assume that the magnetic field b is sampled on a dense grid on the sphere Ω_r. We make use of the fact that, with a suitably chosen maximum scale J_{max}, we can approximate the toroidal part b_{tor} on Ω_r via a series expansion in terms of $l^2(\Omega)$-wavelets (see Corollary 5.3):

$$b_{tor}(r\xi) \simeq \left(\sum_{J=0}^{J_{max}} \psi_J^{(3)} \star \left(\psi_J^{(3)} * b \right)(r) \right)(\xi).$$

Using the fact that $b_{tor}(r,\cdot) = L^* Q_b$, we immediately get an approximation for the toroidal scalar, i.e.,

$$Q_b(r\xi) \simeq \left(\sum_{J=0}^{J_{max}} \tilde{\Psi}_J * \left(\psi_J^{(3)} * b \right)(r) \right)(\xi), \tag{5.27}$$

where the kernel $\tilde{\Psi}_J$ is given such that the relation $\psi_J^{(3)}(\eta,\xi) = L_\xi^* \tilde{\Psi}_J(\eta,\xi)$ holds true. Using (5.27) together with (3.6), we arrive at an approximation of the radial current density on Ω_r corresponding to the toroidal magnetic field there:

$$\begin{aligned} \mu_0 j_{rad}(r\xi) &= \frac{1}{r} \xi \Delta_\xi^* Q_b(r\xi) \\ &\simeq \frac{1}{r} \left(\sum_{J=0}^{J_{max}} \tilde{\psi}_J^{(1)} \star \left(\psi_J^{(3)} * b \right)(r) \right)(\xi), \end{aligned} \tag{5.28}$$

with $\tilde{\psi}_J^{(1)}(\eta,\xi) = \xi \Delta_\xi^* \tilde{\Psi}_J(\eta,\xi)$. Note that this equation is just a different expression of a well known fact; i.e., the toroidal magnetic field at a certain altitude is solely due to the radial current distributions at that very height. Equation (5.28) is the starting point of the examples in the next section. It is noteworthy that, from a practical point of view, (5.28) can successfully be applied if the geomagnetic measurements available are sampled within a comparatively short period of time; i.e., the time-scale under consideration is such that the variations in the satellite's altitude can be neglected to some extent (cf. [27] and [24] for first applications). If the radial variations start to play a role, one can still neglect these variations if the data are appropriately preprocessed, i.e., if suitable geomagnetic field models are subtracted prior to the numerical applications (see, e.g., [28]).

6. Applications to geomagnetic satellite data. As examples of the wavelet Mie representation of the magnetic field, electric current distributions at satellite altitudes are determined from data sets of vectorial MAGSAT and CHAMP data. The method is based on our considerations in section 5, especially (5.28). The current distributions under consideration are due to ionospheric F region currents, which are extensively treated in the literature (see [29] and the references therein).

The data sets used in the first example are similar to those used by Olsen [29] for a spherical harmonic approach to the Mie representation and have kindly been made available by him. MAGSAT was orbiting the earth in a sun-synchronous orbit, thus acquiring data only at dawn and dusk local times. Neglecting the variations in altitude of the MAGSAT satellite, one month of MAGSAT data (centered at March 21, 1980) is transformed to geomagnetic components and is averaged onto the equiangular longitude-latitude grid (90×90 grid points) proposed in [10], which is then used to

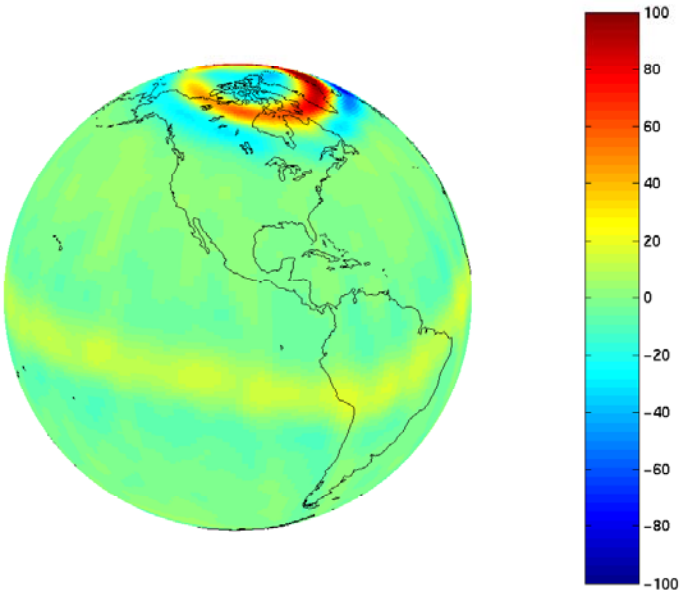


FIG. 3. *Radial current density during evening local time obtained from a wavelet Mie representation of MAGSAT data with vectorial CuP wavelets up to scale* 5. *[nA/m^2]*

discretize the convolution integrals. This averaging process is performed using a robust Tuckeys biweight method (cf. [19]). The dusk and dawn data are treated separately such that two separate data sets are obtained. Prior to the averaging process, a geomagnetic field model (GSFC(12/83) up to degree and order 12) due to [21] is subtracted from the measurements in order to avoid spurious effects due to the neglected altitude variations (cf. [29]).

According to (5.28), the radial current distribution at a fixed height can be calculated from the wavelet coefficients of the toroidal field at that altitude, i.e., $(\psi_J^{(3)} * b)(r)$. With regard to the present example, we calculate these coefficients by means of spherical vectorial CuP wavelets up to scale 5 from the evening data set. Then, in a second step, these coefficients are utilized to calculate the corresponding radial current distribution. Figure 3 shows the reconstruction of the radial current density $J_{rad} = (\xi \cdot j_{rad}(\xi))$. (Note that, for enhancing the visible features, the color scale has been driven in saturation; i.e., although there are currents with absolute values larger than 100 nano-Amperes per square meter (nA/m^2), we use a color bar ranging from -100 to 100 nA/m^2.)

The largest radial current densities ($|J_{rad}| \lesssim 150$ nA/m^2) are present in the polar regions. In agreement with the results in [29], the main current flow in the polar cap is directed into the ionosphere ($J_{rad} > 0$) during evening. At the poleward boundary of the polar oval the currents flow out of the ionosphere, while the main current direction is into the ionosphere at the equatorward boundary. At the magnetic dip equator one realizes comparatively weak upward currents ($|J_{rad}| \lesssim 25$ nA/m^2) accompanied by even weaker downward currents at low latitudes. These current distributions are the radial components of the so-called meridional current system of the EEJ. Figure 4 presents the same results as Figure 3 but in a different projection, thus enabling a better view of the meridional currents. As can be expected from theoretical considerations, the corresponding signatures follow the geomagnetic dip equator.

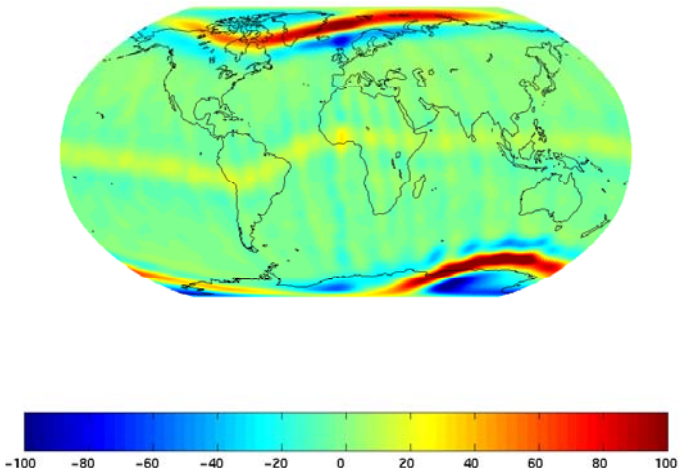


FIG. 4. *Radial current density during evening local time obtained from a wavelet Mie representation of MAGSAT data with vectorial CuP wavelets up to scale* 5. *[nA/m²]*

With regard to the convergence of our approach, there remains the question of how to find a suitable maximum scale J_{max} where the calculations should be stopped. Thus far, we have not investigated this matter in mathematical detail; nevertheless, our numerical results hint at a reasonably good convergence rate (see [16] for a first theoretical discussion). In this study, we use a heuristic way of determining a suitable maximum scale: Wavelet-approximations of the poloidal and toroidal parts of the magnetic field are calculated, added up, and then compared to the input magnetic field. As long as increasing the wavelet-scale reduces the residual, we have not yet reached the suitable maximum scale and, consequently, the corresponding toroidal coefficients should be used to calculate the associated details of the current distribution. In our numerical example this leads to a maximum wavelet-scale $J_{max} = 5$. If a wavelet-scale higher than 5 is chosen, one immediately realizes numerical artifacts in the detail information; e.g., the magnetic field and the electric currents reach unreasonably high peak values all along the satellite's tracks, and the detail information is dominated by spiky features. Though this method of determining the maximum scale is quite heuristic, we can also estimate a reasonable maximum scale using the Shannon sampling theorem. The Shannon sampling theorem tells us that the resolution of the features detectable in the data is limited, depending on the sampling rate of the data set. In the example above, we have used a 90×90 point equiangular grid. This corresponds to a resolvable horizontal wavelength of approximately 898 kilometers. In terms of spherical harmonics, this yields a maximum of approximately 43. Since our wavelets are constructed as linear combinations of spherical harmonics, the maximum scale of the wavelets is determined by this maximum degree. In the present example this estimated maximum wavelet-scale turns out to be 5, which is in accordance with our numerical results. Figures 5 and 6 show a multiresolution of the radial currents and illustrate the convergence of the approximation process up to wavelet-scale 5. The upper rows show the scale information of consecutive scales ($J = 0, 1, 2, 3$ in Figure 5, $J = 4, 5, 6$ in Figure 6), while the lower rows present the respective detail information. Note that adding the scale information and the details of corresponding scales leads to the scale information of the next higher scale, and so on. Though the typical structures of the polar field aligned currents and the EEJ

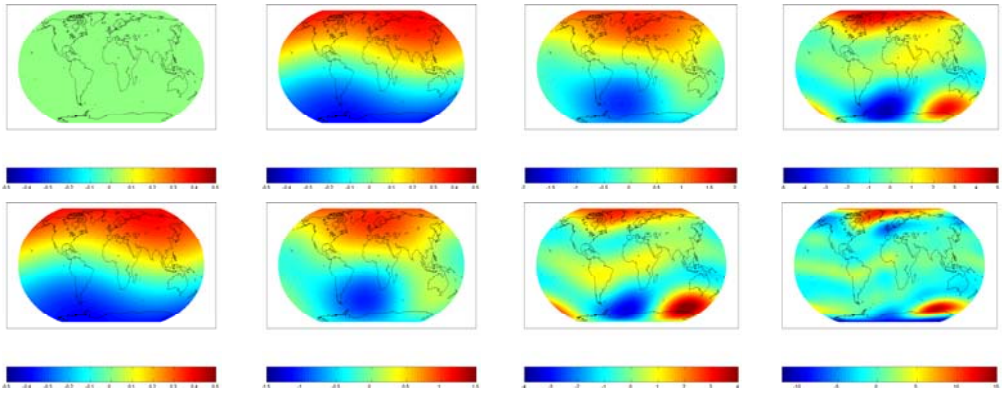


FIG. 5. *Multiscale reconstruction of the radial current density from MAGSAT data using CuP kernels. Upper row: approximations* $\tilde{\varphi}_J \star \varphi_J * b$ *from scaling functions of scales* $J = 0, 1, 2, 3$*; lower row: corresponding detail information* $\tilde{\psi}_J \star \psi_J * b$ *with* $J = 0, 1, 2, 3$*. Note that adding the information of an upper image to the detail information contained in the corresponding subjacent image results in the next upper image.* $[nA/m^2]$

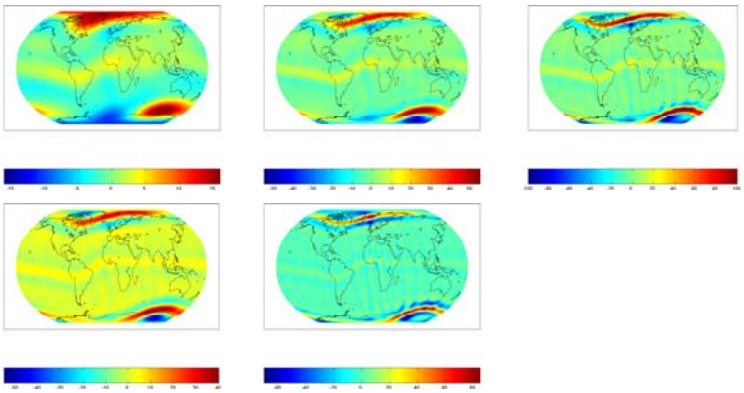


FIG. 6. *Multiscale reconstruction of the radial current density from MAGSAT data using CuP kernels. Upper row: approximations* $\tilde{\varphi}_J \star \varphi_J * b$ *from scaling functions of scales* $J = 4, 5, 6$*; lower row: corresponding detail information* $\tilde{\psi}_J \star \psi_J * b$*, with* $J = 4, 5$*. Note that adding the information of an upper image to the detail information contained in the corresponding subjacent image results in the next upper image.* $[nA/m^2]$

gradually start to become visible in the detail information of scale 3, it is obvious from the magnitude of the respective currents that the contributions of scales 4 and 5 are predominant.

In order to demonstrate the possibility of regional calculations, Figure 7 presents a reconstruction of the radial current systems during dusk local times over the polar region. These results are obtained using vectorial CuP wavelets of scales 4 and 5 (recall the predominance of these scales in Figures 5 and 6) and a data window centered at the geographic north pole with a half angle of 60° as well as an integration window with the same center but a half angle of 55° (the white border approximately illustrates the extent of the calculation region). The visualization window is a little smaller than the calculation window in order to suppress Gibbs phenomena. Comparing Figure 7 with Figure 3 shows that the structures of the radial currents are clearly visible

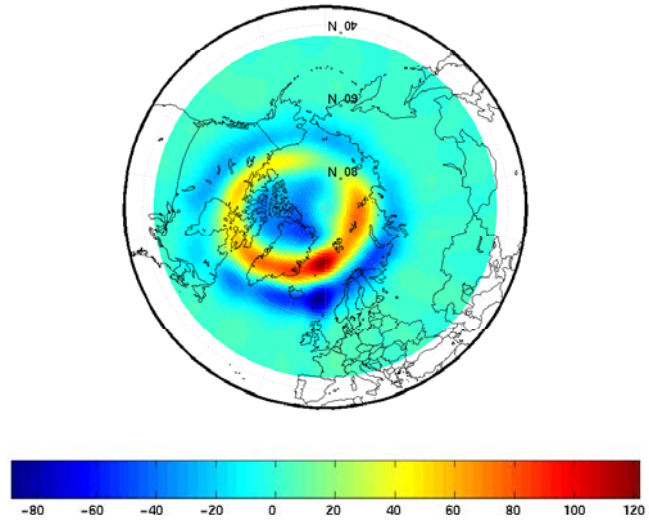


FIG. 7. *Local reconstruction of radial current density during evening local time obtained from a wavelet Mie representation of MAGSAT data with vectorial CuP wavelets at scales* 4 *and* 5. *The white area corresponds to the calculation region. [nA/m^2]*

though slightly weaker in magnitude. This slight difference is due to the fact that we have omitted the contributions of wavelet-scales up to 3, i.e., features of coarse spatial resolution. The signatures seen in our results are the effects of higher wavelet-scales (4 and 5) and consequently are of more or less confined spatial extent. As might be expected from the physical point of view, these are clearly the main radial current contributions in the polar region. The effects of lower scales can be neglected. This, however, demonstrates the regional character of the radial current distributions and suggests the use of space-adaptive methods like the one presented here.

The results of the previous example illustrate the geometry of the ionospheric currents at a fixed (magnetic) local time, i.e., the Earth-satellite-Sun geometry was fixed during the process of data accumulation. The reader should be aware of the fact that the current distributions presented in Figures 3–7 do not illustrate the global distribution of the radial currents but show a small strip of the currents moving over the earth (along longitude) during the course of the day. This is because ionospheric current systems are not properly described in earth-fixed coordinate systems like geographic longitude and latitude. Since the conductance of the ionosphere is varying with the influence of the sun, the magnetic field induced by ionospheric currents is linked to the position of the sun and the distance of the observing satellite to the geomagnetic equator. Consequently, a sun-fixed reference frame should rather be used to parameterize ionospheric currents. Very advantageous, in that sense, is the coordinate system of magnetic local time $MLT \in [0, 24]$ (instead of longitude) and quasi-dipole latitude $QDlat \in [-90, 90]$. The magnetic local time thereby denotes the relative position of the satellite with respect to the magnetic field and the sun, while the quasi-dipole latitude gives the relative position of the satellite with respect to the geomagnetic equator. For more information on these coordinate systems, the reader might consult [22] and the references therein. In order to use MLT and QDlat as a parametrization, one needs to utilize geomagnetic data from satellites with polar but not sun-synchronous orbits thus covering the whole span of magnetic local times

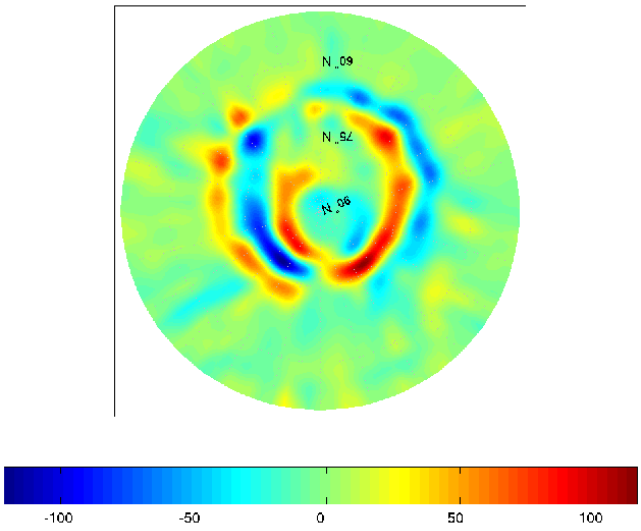


FIG. 8. *Local reconstruction of radial current density from a wavelet Mie representation of CHAMP data in MLT and QDlat, calculated with a CuP scaling function of scale* 6. *[nA/m^2] Figure courtesy of Carsten Mayer* [26]*; used by permission.*

(i.e., from 0 h to 24 h). The German geoscientific satellite CHAMP, operated by the GeoForschungsZentrum in Potsdam, Germany, is such a satellite. Among other instruments, the CHAMP satellite is equipped with high precision vector and scalar magnetometers and, in contrast to MAGSAT, covers all magnetic local times within four months. In what follows we present a result calculated by Mayer [26] analyzing a CHAMP data set via a wavelet Mie representation parameterized in QDlat and MLT. Three days of CHAMP vector data (September 10, 16, and 17, 2001) are used. In polar regions these data suffice to cover the whole span of magnetic local times (see [26]). The polar data are transformed to the QDlat-MLT coordinate system and then averaged to an equiangular integration grid using a robust method. A wavelet Mie representation is performed over the geomagnetic north pole, and the radial current distributions are calculated via (5.28) from the toroidal magnetic field contribution. Figure 8 shows the resulting radial currents in the northern polar region, which are in accordance with the physical models presented in [20] and [11]. This result can now be interpreted as the evolution of the currents' morphology in magnetic local time; for example, it is clearly visible how the currents' polarity changes at the noon-midnight plane.

7. Summary and outlook. The Mie representation for the geomagnetic field has the advantage that it can equally be applied in regions of vanishing as well as nonvanishing electric current densities. The standard method of deriving the Mie representation is given by a spherical harmonic parametrization, i.e., by expanding the corresponding Mie scalars in terms of spherical harmonics. Considering the measurements (magnetic field or currents) to be given in a spherical shell, we have presented a wavelet parametrization of the magnetic field and the corresponding electric current densities in Mie representation, i.e., a wavelet Mie representation. The use of wavelets as trial functions for field parametrization enables us to cope with electric currents

(and corresponding magnetic effects) that vary rapidly with latitude or longitude, or that are confined to certain regions. Consequently, we are able to reflect the various levels of space localization in the form of a vectorial multiresolution analysis and can thus take efficient account of the specific concentration of the current densities in space. Using our approach, the direct as well as the inverse geomagnetic source problem now admit a treatment within a vectorial multiscale framework.

Neglecting variations in altitude, we have provided numerical examples that illustrate the multiscale approximation of radial current distributions from sets of vectorial geomagnetic field data from the MAGSAT as well as the CHAMP satellite. Global as well as regional reconstructions of the radial current densities are calculated and demonstrate the functionality of the approach. With regard to future studies, the next reasonable step is to incorporate the variations in altitude of the satellite—at least to some extent—since this would allow for the determination of horizontal current distributions, too. Additionally—in studies using synthetic data, based on satellite data sampled over large time intervals, or using data from multisatellite missions—a simultaneous wavelet parametrization of the poloidal and toroidal magnetic fields from the corresponding electric currents (or vice versa) should be derived in future works.

Acknowledgments. The author thanks Prof. Dr. W. Freeden (TU Kaiserslautern), Dr. C. Mayer (TU Kaiserslautern), and Dr. N. Olsen (Danish Space Research Institute, Copenhagen) for their support, their comments, and many fruitful discussions. The detailed and helpful comments of the referees are also gratefully acknowledged.

REFERENCES

[1] G. E. Backus, *Poloidal and toroidal fields in geomagnetic field modeling*, Rev. Geophys., 24 (1986), pp. 75–109.
[2] G. E. Backus, R. Parker, and C. Constable, *Foundations of Geomagnetism*, Cambridge University Press, Cambridge, UK, 1996.
[3] M. Bayer, S. Beth, and W. Freeden, *Geophysical field modeling by multiresolution analysis*, Acta Geod. Geophys. Hung., 33 (1998), pp. 289–319.
[4] M. Bayer, W. Freeden, and T. Maier, *A vector wavelet approach to iono- and magnetospheric geomagnetic satellite data*, J. Atmospher. Solar-Terrestrial Phys., 63 (2001), pp. 581–597.
[5] S. Beth, *Multiscale Approximation by Vector Radial Basis Functions on the Sphere*, Shaker Verlag, Aachen, Germany, 2000.
[6] C. K. Chui, *An Introduction to Wavelets*, Academic Press, New York, 1992.
[7] I. Daubechies, *The wavelet transform, time-frequency localization and signal analysis*, IEEE Trans. Inform. Theory, 36 (1990), pp. 961–1005.
[8] P. Deuflhard, *On algorithms for the summation of certain special functions*, Computing, 17 (1976), pp. 37–48.
[9] E. F. Donovan, *Modeling the magnetic effects of field-aligned currents*, J. Geophys. Res., 98 (1993), pp. 529–543.
[10] J. R. Driscoll and D. M. Healy, *Computing Fourier transforms and convolutions on the 2-sphere*, Adv. Appl. Math., 15 (1994), pp. 202–250.
[11] U. Engels and N. Olsen, *Computation of magnetic fields within source regions of ionospheric and magnetospheric currents*, J. Atmospher. Solar-Terresterial Phys., 60 (1998), pp. 1585–1592.
[12] W. Freeden, *On Approximation by Harmonic Splines*, Manuscr. Geod., 6 (1981), pp. 193–244.
[13] W. Freeden, *Multiscale Modeling of Spaceborne Geodata*, B. G. Teubner, Stuttgart, Leipzig, 1999.
[14] W. Freeden and T. Gervens, *Vector spherical spline interpolation—Basic theory and computational aspects*, Math. Methods Appl. Sci., 16 (1993), pp. 151–183.

[15] W. Freeden, T. Gervens, and M. Schreiner, *Constructive Approximation on the Sphere (With Applications to Geomathematics)*, Oxford Science Publications, Clarendon, Oxford, UK, 1998.

[16] W. Freeden and U. Windheuser, *Spherical wavelet transform and its discretization*, Adv. Comput. Math., 5 (1996), pp. 51–94.

[17] W. Freeden and U. Windheuser, *Combined spherical harmonic and wavelet expansion—A future concept in earth's gravitational potential determination*, Appl. Comput. Harm. Anal., 4 (1997), pp. 1–37.

[18] G. Gerlich, *Magnetfeldbeschreibung mit verallgemeinerten poloidalen und toroidalen Skalaren*, Z. Naturforsch., 8 (1972), pp. 1167–1172.

[19] R. V. Hogg, *An introduction to robust estimation*, in Robustness in Statistics, G. N. Launer and R. L. Wilkinson, eds., Academic, San Diego, CA, 1979, pp. 1–17.

[20] T. Iijima and A. Potemra, *Large-scale characteristics of field-aligned currents associated with substorms*, J. Geophys. Res., 90 (1978), pp. 2593–2598.

[21] R. A. Langel and R. H. Estes, *The near-Earth magnetic field at* 1980 *determined from MAGSAT data*, J. Geophys. Res., 90 (1985), pp. 2495–2510.

[22] R. A. Langel, N. Olsen, and T. J. Sabaka, *A Comprehensive Model of the Near-Earth Magnetic Field: Phase* 3, Technical Report TM-2000-209894, NASA, Washington, DC, 2000.

[23] R. A. Langel, T. J. Sabaka, R. T. Baldwin, and J. A. Conrad, *The near-earth magnetic field from magnetospheric and quiet-day ionospheric sources and how it is modeled*, Phys. Earth Planet. Interiors, 98 (1996), pp. 235–267.

[24] T. Maier, *Multiscale Geomagnetic Field Modeling from Satellite Data—Theoretical Aspects and Numerical Applications*, Ph.D. thesis, Department of Mathematics, University of Kaiserslautern, Kaiserslautern, Germany, 2003; available online from http://kluedo.ub.uni-kl.de/volltexte/2003/1553/.

[25] S. Maus, *New Statistical Methods in Gravity and Magnetics*, Habilitation, Gemeinsame Naturwissenschaftliche Fakultät der Technischen Universität Carolo-Wilhelmina zu Braunschweig, Braunschweig, Germany, 2001.

[26] C. Mayer, *Wavelet Modelling of Ionospheric Currents and Induced Magnetic Fields From Satellite Data*, Ph.D. thesis, Department of Mathematics, University of Kaiserslautern, Kaiserslautern, Germany, 2003; available online from http://kluedo.ub.uni-kl.de/volltexte/2003/1623/.

[27] C. Mayer and T. Maier, *Multiscale determination of radial current distribution from CHAMP FGM-data*, in First CHAMP Mission Results for Gravity, Magnetic and Atmospheric Studies, C. Reigber, H. Luehr, and P. Schwintzer, eds., Springer, New York, 2003, pp. 339–345.

[28] N. Olsen, *A new tool for determining ionospheric currents from magnetic satellite data*, Geophys. Res. Lett., 24 (1996), pp. 3635–3638.

[29] N. Olsen, *Ionospheric F region currents at middle and low latitudes estimated from MAGSAT data*, J. Geophys. Res. A, 3 (1997), pp. 4563–4576.

[30] A. D. Richmond, *The computation of magnetic effects of field-aligned magnetospheric currents*, J. Atmospher. Terrest. Phys., 36 (1974), pp. 245–252.

[31] P. Ritter, A. Viljanen, H. Luehr, O. Amm, and N. Olsen, *Ionospheric currents from CHAMP magnetic field data—Comparison with ground based measurements*, in First CHAMP Mission Results for Gravity, Magnetic and Atmospheric Studies, C. Reigber, H. Luehr, and P. Schwintzer, eds., Springer, New York, 2003, pp. 346–352.

[32] B.J. Schmitt, *The poloidal-toroidal representation of solenoidal fields in spherical domains*, Analysis, 15 (1995), pp. 257–277.

[33] L. Shure, R. L. Parker, and G. E. Backus, *Harmonic splines for geomagnetic modeling*, Phys. Earth Planet. Interiors, 28 (1982), pp. 215–229.

[34] D. P. Stern, *Representation of magnetic fields in space*, Rev. Geophys., 14 (1976), pp. 199–214.

[35] J. Waterman, F. Christiansen, V. Popov, P. Stauning, and O. Rasmussen, *Field-aligned currents inferred from low-altitude earth-orbiting satellites and ionospheric currents inferred from ground-based magnetometers—Do they render consistent results?*, in First CHAMP Mission Results for Gravity, Magnetic and Atmospheric Studies, C. Reigber, H. Luehr, and P. Schwintzer, eds., Springer, New York, 2003, pp. 361–368.

SIAM J. APPL. MATH.
Vol. 65, No. 6, pp. 1913–1932

RETRIEVING TOPOLOGICAL INFORMATION FOR PHASE FIELD MODELS*

QIANG DU†, CHUN LIU†, AND XIAOQIANG WANG†

Abstract. The phase field approach has become a popular tool in modeling interface motion, microstructure evolution, and more recently the shape transformation of vesicle membranes under elastic bending energy. While it is advantageous to employ phase field models in numerical simulations to automatically handle topological changes to the microstructures or the configurations of vesicle membranes, detecting topological events may also become important for many applications such as those in the simulation of blood cells. Motivated by such considerations, a new quantity is formulated to retrieve some topological information based on the phase field formulation and to capture the occurrence of topological events. It can also be used as a control method to avoid unphysical changes of topology due to the numerical methods, should it become necessary for particular practical applications. Through numerical experiments, we demonstrate the effectiveness and the robustness of the new quantity in detecting the topology of fluid bubbles and vesicle membranes.

Key words. phase field, elastic bending energy, Gauss–Bonnet formula

AMS subject classifications. 57M50, 74A50, 74S99, 92C05

DOI. 10.1137/040606417

1. Introduction.

Phase field modeling of mezoscopic morphology and microstructure evolution has become popular in recent years (see [6, 7, 8, 9, 12, 13, 21, 22, 24, 34, 35, 38, 39] and the references therein). These phase field approaches are usually combined with energetic variational formulations that lead to a diffuse interface modeled by a mixing energy. This allows topological changes of the interface to take place naturally [28, 31, 37]. Such a feature gives such approaches many advantages in numerical simulation of the interface variation and the interfacial motion (cf. [11]). More recently, we have given another successful application of the phase field model for computing the equilibrium configurations of vesicle membranes that minimize the bending elastic energy [19, 18]. The methods developed in [19] can be very useful in morphological studies of vesicle membranes under elastic bending energy, which has many interesting applications.

In this paper, we continue the study of the phase field model but from a different, and perhaps also very novel, point of view. We are motivated by the fact that in many engineering and biological applications, such as the modeling of blood cells in vascular systems, topological information about the vesicle membranes is of critical value. Thus, how to detect and control the topological change of the interface in the phase field modeling and numerical simulation becomes an important issue. Partly due to the nature of the phase field method (and all other level set methods) in their standard formulation, there is no mechanism preventing the topological change of the membranes or other interfaces. In fact, some of the topological changes are due purely to the formulations, instead of the underlying physics. To our knowledge, there has not

*Received by the editors April 7, 2004; accepted for publication (in revised form) January 13, 2005; published electronically August 3, 2005. The research of the first and third authors was supported in part by NSF-DMR 0205232, NSF-DMS 0196522, and NSF-DMS 0400297. The research of the second author was supported by NSF-DMS 040850 and NSF-DMS 0509094.

http://www.siam.org/journals/siap/65-6/60641.html

†Department of Mathematics, Pennsylvania State University, University Park, PA 16802 (qdu@math.psu.edu, liu@math.psu.edu, wang@math.psu.edu).

been any discussion in the literature on how to recover relevant topological information from the phase field simulations, nor to address further control mechanisms.

The general phase field model is based on the introduction of a phase function (or order parameter) $\phi = \phi(x)$, defined on the physical (computational) domain Ω that encloses Γ. The function ϕ is used to label the inside and the outside of the vesicle Γ: the level set $\{x : \phi(x) = 0\}$ gives the membrane, while $\{x : \phi(x) > 0\}$ represents the interior of the membrane and $\{x : \phi(x) < 0\}$ the exterior. A *transition thickness* parameter is also chosen to characterize the typical length scale of the transition layer (or the thickness of the *regularized* diffuse interface). For time dependent problems, it is natural to allow $\phi = \phi(x, t)$. The objective of our study here is to propose some robust and efficient methods for retrieving or recovering interesting topological information within the phase field framework. In particular, we will develop some robust formulae for computing the Euler number for the modeled interface based on order parameter ϕ.

We expect that our study here has the potential of opening up a host of exciting new applications of phase field modeling, including the use of the topological quantities in a control setting. Our numerical simulation indicates that the generalized Euler number ($\frac{\chi}{2}$ in three dimensions and χ in two dimensions) not only is a better indicator of topological changes than the energy functional we had been using in our previous paper [19], but in fact, gives a quantized jump upon a completion of the topological change (a direct consequence of the Gauss–Bonnet formula) for regular surfaces. Moreover, when the computed surface passes singularity, the new formula for χ based on the phase field formulation gives a fractional interpolation of the usual Euler number.

The study of the Euler number in terms of χ, or the Euler–Poincaré index number η, may provide guidance for the study of other topological quantities within the phase field framework. The ideas proposed in this paper may be equally applicable to other simulation methods for free boundary and interface problems such as the level set methods. Our work here is also likely to shed light on the study of many geometrical modeling problems, where providing topological information or controlling topological changes may be highly desirable.

The rest of the paper is organized as follows. In section 2, we briefly recall the phase field model, and we discuss a formulation of the Euler number (in terms of χ) within the phase field framework that can be used to recover some topological information. We present a set of formulae in various three-dimensional (3-D) and two-dimensional (2-D) cases. We also discuss the Euler–Poincaré index number η when the surfaces involve singularities. In section 4, we illustrate the method of computing the quantity χ in two applications and demonstrate the effectiveness and the robustness of the our formulation. Finally, some concluding remarks are given in section 5.

2. The Euler number. Given an oriented (regular) compact (i.e., without boundary) surface Γ, the well-known Gauss–Bonnet formula states that

$$\int_{\Gamma} K \, ds = 2\pi\chi, \tag{2.1}$$

where $K = k_1 k_2$ is the Gaussian curvature of the surface in R^3, ds is the area element, and $\frac{\chi}{2}$ in three dimensions (χ in two dimensions) is the Euler number [16]. The number χ is a commonly used topological quantity. For some frequently encountered surfaces, we have $\chi = 2$ for a sphere, $\chi = 0$ for a torus, and $\chi = -2$ for a torus with two holes.

For 2-D curves, K is the curvature and $\chi = 1$ for a circle. For convenience, in this paper we call $\frac{\chi}{2}$ the Euler number in 3-D cases and χ in 2-D cases.

In this section, it is our goal to find a suitable expression of the Euler number when the surface is implicitly defined by a phase field formulation. We first give the general formula of the Euler number under a general phase field definition for both 2-D and 3-D spaces. Then we give some simplified formulae under some specific ansatz assumptions corresponding to our applications. For simplicity, we focus on only the static case when deriving the formulae. In the time dependent case, since we are mostly interested in the topological information of a spatially distributed interface at a fixed time stance, the generalizations to time dependent problems are obvious.

We note that many of our derivations are through formal asymptotics, though more detailed and more rigorous justifications will be provided in our subsequent work [17].

2.1. The 3-D case. Let Γ be a smooth oriented compact surface in a domain Ω in $\mathbf{R}^3$; we note that Γ is allowed to have multiple disconnected pieces. Let p be a monotone increasing function defined from $\mathbf{R}$ to $\mathbf{R}$ with $p(0) = 0$. For each function p, we take a phase field function $\phi = \phi(x)$ of Ω as $\phi(x) = p(d(x, \Gamma))$ where the *signed* distance function $d = d(x, \Gamma)$ is defined to be positive inside Ω and negative outside Ω. The level sets of ϕ are denoted by $\Gamma_\mu = \{x \in \Omega | \phi(x) = \mu\}$. In particular, we have $\Gamma = \Gamma_0$. We also define $\Omega(a, b) = \{x \in \Omega \,|\, b < \phi(x) < a\}$, which forms a banded (layered) neighborhood around the surface for $b < 0 < a$.

Define $\Lambda(M) = \mathrm{Tr}(\mathrm{Adj}(M))$ for a matrix M. Clearly, $\Lambda(M)$ is the coefficient of the linear term of the characteristic polynomial of M. In particular, for the singular matrix $M = \nabla^2 d$ we have $\Lambda(M) = \lambda_1(M)\lambda_2(M)$, with λ_1 and λ_2 being the two nonzero eigenvalues of M.

THEOREM 2.1. *Using the notation above, for any monotone increasing function p there exists $b < 0 < a$ such that the matrix M, defined by*

$$M(x)_{ij} = \frac{1}{2\sqrt{\pi(a-b)|\nabla\phi|}} \left(\nabla_i \nabla_j \phi - \frac{\nabla |\nabla\phi|^2 \cdot \nabla\phi}{2|\nabla\phi|^4} \nabla_i \phi \nabla_j \phi \right), \tag{2.2}$$

is a singular matrix for any $x \in \Omega(a, b)$ and the Euler number of Γ is given by

$$\frac{\chi}{2} = \int_{\Omega(a,b)} \Lambda(M(x))\, dx. \tag{2.3}$$

Proof. Since ϕ depends only on the distance d, and p is monotone increasing, there exist real numbers a and b, with $b < 0 < a$, sufficiently close to 0 such that Γ_μ are close to the parallel translations of Γ in the normal direction for all $b \le \mu \le a$, and all Γ_μ have the same topology as Γ.

Direct computation shows

$$\nabla_i \nabla_j d = \frac{\nabla^2 \phi}{p'} - \frac{p''}{p'} \nabla_i d \nabla_j d = \frac{1}{p'} [\nabla_i \nabla_j \phi - p'' \nabla_i d \nabla_j d].$$

The matrix $\nabla^2 d$, with $d = d(x, \Gamma)$ being the signed distance from x to the surface Γ, always has a zero eigenvalue with ∇d as the eigenvector. This is due to the fact that $\nabla^2 d$ is symmetric and $|\nabla d| = 1$. On the surface Γ, the two other eigenvalues are actually the two principle curvatures of Γ (k_1 and k_2 in this case). The Gaussian curvature K is of course the product of these two eigenvalues, while the mean curvature H is

given by the mean of the two eigenvalues. Both quantities can in fact be defined and computed on all the level sets in $\Omega(a,b)$. For instance, the Gaussian curvature K can be conveniently computed from $\mathrm{Tr}(\mathrm{Adj}(\nabla^2 d))$ (the sum of the three principal 2×2 minors of $\nabla^2 d$; see section 2.6 for an example in the cylindrically symmetric case).

Assuming that $k_1$and k_2 remain constant along the normal directions in Ω', we have

$$\begin{aligned}
\frac{\chi}{2} &= \frac{1}{4\pi}\int_\Gamma k_1(x)k_2(x)\,ds \\
&= \frac{1}{4\pi(a-b)}\int_{p^{-1}(b)}^{p^{-1}(a)} p'(\tau)d\tau \int_\Gamma k_1(x)k_2(x)\,ds \\
&= \frac{1}{4\pi(a-b)}\int_{\Omega(a,b)} p'(d(x,\Gamma))k_1(x)k_2(x)\,dx \qquad (2.4)\\
&= \frac{1}{4\pi(a-b)}\int_{\Omega(a,b)} p'(d(x,\Gamma))\Lambda(\nabla^2 d(x,\Gamma))\,dx \\
&= \frac{1}{4\pi(a-b)}\int_{\Omega(a,b)} \frac{1}{p'(d(x,\Gamma))}\Lambda(\nabla^2\phi - p''\nabla_i d\nabla_j d)\,dx. \qquad (2.5)
\end{aligned}$$

Now, since p is monotone increasing, we have $p'(d(x,\Gamma)) = |\nabla\phi(x)|$ and

$$p''(d(x,\Gamma)) = \nabla|\nabla\phi|\cdot\frac{\nabla\phi}{|\nabla\phi|}(x) = \frac{\nabla|\nabla\phi|^2\cdot\nabla\phi}{2|\nabla\phi|^2}(x).$$

Hence we get the general formula for the Euler number in three dimensions:

$$\begin{aligned}
\frac{\chi}{2} &= \frac{1}{4\pi(a-b)}\int_{\Omega(a,b)}\frac{1}{|\nabla\phi|}\Lambda\left(\nabla_i\nabla_j\phi - \frac{\nabla|\nabla\phi|^2\cdot\nabla\phi}{2|\nabla\phi|^4}\nabla_i\phi\nabla_j\phi\right)dx \qquad (2.6)\\
&= \frac{1}{4\pi(a-b)}\int_{\Omega(a,b)}\frac{1}{|\nabla\phi|}\Lambda(M(x))\,dx.
\end{aligned}$$

From the definitions of M and $\Lambda(M)$, we know that $M(x)$ is singular for any $x\in\Omega(a,b)$ in the sense that it always has a zero eigenvalue, and (2.3) holds. ∎

The formula (2.3) forms the basis for our efforts to recover topological information, in particular the Euler number.

2.2. The 2-D case.

THEOREM 2.2. *If $\Omega\in\boldsymbol{R}^2$, with the same notation as in Theorem* 2.1*, for any monotone increasing function p there exists $b<0<a$ such that*

$$\chi = \frac{1}{2\pi(a-b)}\int_{\Omega(a,b)}\left(-\triangle\phi + \frac{\nabla|\nabla\phi|^2\cdot\nabla\phi}{2|\nabla\phi|^2}\right)dx. \qquad (2.7)$$

Proof. Using the same argument as above, this can be derived as follows:

$$\begin{aligned}
\chi &= \frac{1}{2\pi}\int_\Gamma K(x)\,ds \\
&= \frac{1}{2\pi(a-b)}\int_{\Omega(a,b)} p'(d(x,\Gamma))K(x)\,dx \\
&= \frac{1}{2\pi(a-b)}\int_{\Omega(a,b)} (-\triangle\phi + p''(d(x,\Gamma)))\,dx \\
&= \frac{1}{2\pi(a-b)}\int_{\Omega(a,b)}\left(-\triangle\phi + \frac{\nabla|\nabla\phi|^2\cdot\nabla\phi}{2|\nabla\phi|^2}\right)dx. \qquad (2.8)
\end{aligned}$$

∎

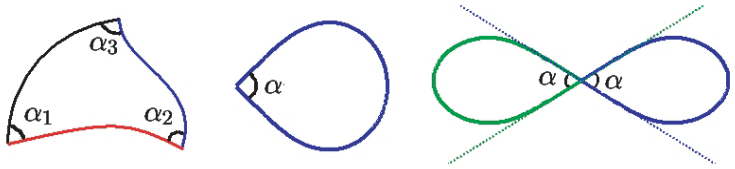


FIG. 2.1. *Singular 2-D cases.*

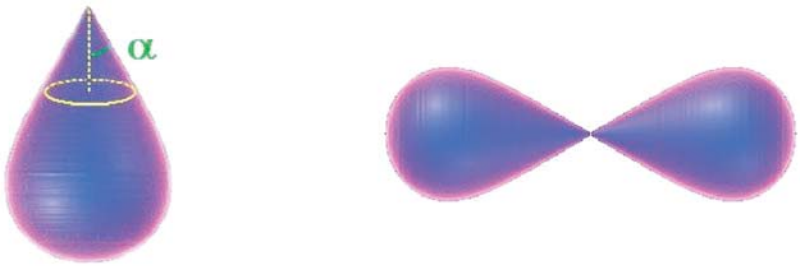


FIG. 2.2. *Singular 3-D cases.*

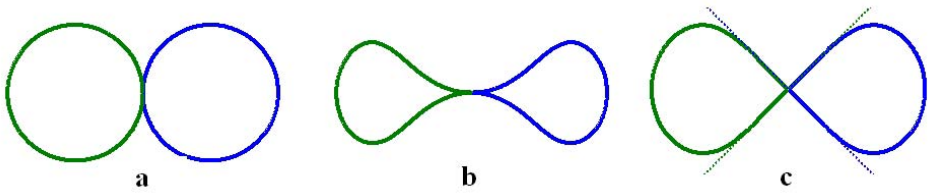


FIG. 2.3. *Singular 2-D cases. The inner intersect angles are* π, 0, $\pi/2$ *for cases a, b, and c, respectively.*

2.3. Cases involving singularities on the interfaces. Both Theorems 2.1 and 2.2 require that Γ be a smooth oriented compact surface. However, in realistic physical applications, we always encounter the singular cases where either the curves or the surfaces intersect or have some sharp angles or cones. Figure 2.1 illustrates several simple singular cases in two dimensions, while Figure 2.2 shows two singular shapes in three dimensions.

With the possible occurrence of the singularities, we will employ the general Gauss–Bonnet formula. In the 2-D case, suppose that the curves are piecewise smooth with n vertices (sharp corners) and that the inner angles for each vertices are $\{\alpha_i, i = 1, \ldots, n\}$. The Gauss–Bonnet formula reads as

$$2\pi\eta = \int_\Gamma K\, ds + \sum_{i=1}^{n}(\pi - \alpha_i) \; = 2\pi\chi + \sum_{i=1}^{n}(\pi - \alpha_i),$$

where η, the Euler–Poincaré index number, is the topological integer, the genus of the surface [16].

We give illustrations, in Figures 2.1 and 2.3, of the values of χ defined by $2\pi\chi = \int_\Gamma K\, ds$ in the singular cases. For configurations such as the third image in Figure

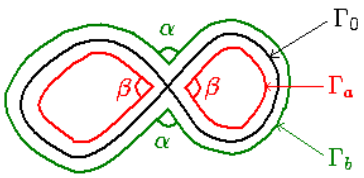


FIG. 2.4. Γ_0, Γ_a, *and* Γ_b *have different topologies but the same Euler number.*

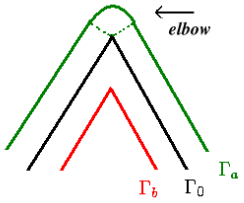


FIG. 2.5. *By getting rid of the singular complement,* Γ_0, Γ_a, *and* Γ_b *have the same Euler number.*

2.1, $2\pi\eta = \int_\Gamma K\,ds + 2(\pi - \alpha)$. In particular, for the cases in Figure 2.3, $\alpha = \pi, 0, \frac{\pi}{2}$ for the cases a, b, and c, respectively. In all of these cases, the Euler–Poincaré index number η is always 2, representing the number of bubbles.

On the other hand, the value of χ can be used to detect the change of topology even with the presence of the singular cases. Again, using the example in Figure 2.3, the Euler numbers χ are 2, 1, and 1.5, respectively, with the singularity being signaled by the fractional value. We want to point out that in the first case, the Euler number χ is still equal to the Euler–Poincaré index number η, even with the singularity.

When we consider the definition of χ for the 3-D singular cases, we may also get a fractional Euler number when singularity appears. For example, for the configuration in the first picture of Figure 2.2, suppose that the angle of the tangent cone is α; the Euler number can be derived from the following explicit analytic formula by calculating the ratio between the spherical cap cut by the α-cone and the total area of the sphere:

$$\frac{\chi}{2} = \frac{1}{4\pi}\int_\Gamma K\,ds = 1 - \frac{1}{4\pi}(2\pi(1-\sin\alpha)) = \frac{1}{2} + \frac{1}{2}\sin\alpha. \tag{2.9}$$

Finally, we discuss the validity of formulae (2.3) and (2.7) in singular cases. Figure 2.4 shows that Γ_μ seem to have (visual) topology different from that of Γ for some $b < \mu < a$.

However, in reference to Figure 2.4, because each Γ_μ has a singular vertex with angle either α or $\beta = \pi - \alpha$, the Euler number ξ for every Γ_μ is the same as that of Γ_0. Thus Theorems 2.1 and 2.2 still hold for this singular case where the Euler number is a fractional number.

One will encounter the issue of a *singular complement* in the numerical simulations of the interfaces involving singularities. This is crucial in calculating the fractional Euler number correctly. To clarify this phenomenon, let us examine Figure 2.5.

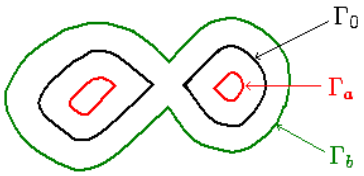


FIG. 2.6. *An illustration of the choice of a and b. Γ_0, Γ_a, and Γ_b have the same Euler number.*

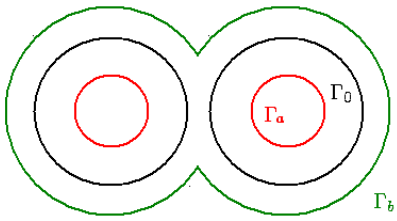


FIG. 2.7. *Γ_0, Γ_b have different Euler numbers.*

Because of the *elbow* shape associated with Γ_a, Γ_a has a different Euler number from Γ_0. This phenomenon, hereafter referred to as the *singularity compensation*, occurs in all numerical experiments involving interfacial singularities due to the numerical smoothing of the surfaces. This appears to introduce difficulties into the application of Theorems 2.1 and 2.2. However, noticing that the elbow has very large curvature, if a method can be designed to filter out such an elbow, the Euler number calculated for the remaining Γ_a would remain the same as that of Γ_0. Such a technique will be introduced and developed later in section 3.

2.4. Stability and autoselection of parameters. Although the proofs of Theorems 2.1 and 2.2 are given under the condition that for every $b < \mu < a$, Γ_μ has the same topology as Γ_0, to correctly identify the Euler number, one can relax such a condition to requiring that the Euler number of Γ_μ be the same as that of Γ_0. By the analysis of section 2.3, it is desirable that the parameters a and b are chosen to be quite different from 0. Figure 2.6 gives an illustration of such a case.

The advantage of choosing a and b some distance away from 0 is to ensure that $\Omega(a, b)$ contains plenty of grid points, which in turn makes the integration for the Euler number more accurate and stable.

However, in general, it is not true that all the neighboring curves or surfaces always share the same Euler number. In Figure 2.7, Γ_0 and Γ_b clearly have different Euler numbers. The largest b can be selected only where the two circles of Γ_b are tangent to each other. How to choose the value of b in such situations becomes a very relevant issue in our computation of the Euler numbers.

On the other hand, in Figure 2.7, we may notice that Γ_a always has the same Euler number as Γ_0, and a can be very different from 0. Thus the selection of a and b may be problem dependent. In some cases, we can select the integration area to be either $\Omega(0, a)$ or $\Omega(b, 0)$ in order to get the correct Euler number.

Denote the Euler number as $E(b,a)$, corresponding to the integration region $\Omega(a,b)$; in order to see the impact of a and b, we employ the following algorithm in our numerical simulations.

ALGORITHM: PARAMETER AUTOSELECTION FOR a. Given a phase field function ϕ defined on region Ω, select a relatively far away from 0 (for instance, $a=1$). Select a tolerance number ν (for instance, $\nu=0.1$) and a small step h (we can use, as an example, $h=a/50$).

- Step 1. If $|E(0,a)-E(0,a/2)|<\nu$, exit; else set $a=a-h$.
- Step 2. If $a<h$, exit with $a=0$; else loop back to Step 1.

The above algorithm is used to select the best a. A similar algorithm can be used to select the best b. A bisection method can be adopted to make the autoselection more efficient. The resulting a and b are used to compute the correct Euler numbers. As verified in the earlier theorems, the above algorithm is assured to terminate with suitable a and b if the step size h and the tolerance ν are chosen to be reasonably small.

2.5. Formula simplification. The formulae given in the above discussion of the Euler number are computable numerically, but for many practical situations they can be further simplified if some appropriate ansatz can be taken. We now discuss some of these simplifications.

First let us take the ansatz

$$\phi=\tanh\left(\frac{d}{\sqrt{2}\epsilon}\right)=p\left(\frac{d}{\sqrt{2}\epsilon}\right), \tag{2.10}$$

which is actually an accurate description of the phase field function in many models such as the basic Allen–Cahn (Ginzburg–Landau) equations, which are popular prototype phase field models for interface and microstructure evolution. Such an ansatz also captures well the phase field function in our energetic phase field model of the vesicle membranes under bending elastic energy.

With such an ansatz, we have the following theorem.

THEOREM 2.3. *As in Theorem* 2.1, *if* $\phi=\tanh(\frac{d}{\sqrt{2}\epsilon})=p(\frac{d}{\sqrt{2}\epsilon})$, *we have the following:*

3-*d case: Let*

$$M_{ij}=\sqrt{\frac{3\epsilon}{8\sqrt{2}\pi}}\left(\nabla^2_{ij}\phi+\frac{2\phi}{1-\phi^2}\nabla_i\phi\nabla_j\phi\right) \tag{2.11}$$

and $\Lambda(M)$ *be defined as before. The Euler number of* Γ *is given by*

$$\frac{\chi}{2}=\lim_{\epsilon\to 0}\int_\Omega \Lambda(M(x))\,dx. \tag{2.12}$$

2-*d case: The Euler number of* Γ *is given by*

$$\chi=\lim_{\epsilon\to 0}\frac{1}{4\pi}\int_\Omega\left(-\triangle\phi+\frac{1}{\varepsilon^2}(\phi^2-1)\phi\right)\,dx. \tag{2.13}$$

Remark 2.4. Based on the ansatz (2.10), we have that

$$p'=1-p,\quad p''=(p-1)p. \tag{2.14}$$

Hence each term in the matrix M is in fact nonsingular.

Proof. A direct calculation shows that

$$\nabla\phi = p'\frac{1}{\sqrt{2}\epsilon}\nabla d, \quad \nabla^2_{ij}\phi = p''\frac{1}{2\epsilon^2}\nabla_i d\nabla_j d + p'\frac{1}{\sqrt{2}\epsilon}\nabla^2 d,$$

and

$$\nabla^2 d = \frac{\sqrt{2}\epsilon\nabla^2\phi}{p'} - \frac{p''}{\sqrt{2}\epsilon p'}\nabla_i d\nabla_j d.$$

Several elementary facts are in order:

$$p'(x) = \tanh'(x) = 1 - \tanh^2(x) = 1 - p^2(x), \qquad p'' = -2p(1-p^2), \quad p'' = -2pp'.$$

Since $\nabla d = \sqrt{2}\epsilon\nabla\phi/p'$, we have

$$\begin{aligned}\nabla^2 d &= \frac{\sqrt{2}\epsilon\nabla^2\phi}{p'} - \frac{p''}{\sqrt{2}\epsilon p'}\nabla_i d\nabla_j d = \sqrt{2}\epsilon\frac{\nabla^2\phi}{p'} + \sqrt{2}\epsilon\frac{2p}{(1-p^2)^2}\nabla_i\phi\nabla_j\phi \\ &= \frac{\sqrt{2}\epsilon}{1-p^2}\left(\nabla^2\phi + \frac{2p}{1-p^2}\nabla_i\phi\nabla_j\phi\right) = \frac{\sqrt{2}\epsilon}{1-\phi^2}\left(\nabla^2\phi + \frac{2\phi}{1-\phi^2}\nabla_i\phi\nabla_j\phi\right).\end{aligned}$$

Simple calculation yields that

$$\int_{-\infty}^{+\infty}(1-\phi^2)^2 dx = \sqrt{2}\epsilon\int_{-\infty}^{+\infty}(1-\tanh^2(x))^2 dx = \frac{4\sqrt{2}\epsilon}{3}.$$

With sufficiently small ϵ, the function ϕ goes to 1 or -1 very quickly (exponentially) away from Γ. Thus, we can just take $a = -b = 1$ with $\epsilon \to 0$, and the matrix in the formula (2.2) becomes

$$M = \sqrt{\frac{3\epsilon}{8\sqrt{2}\pi}}\left(\nabla^2\phi + \frac{2\phi}{1-\phi^2}\nabla_i\phi\nabla_j\phi\right),$$

and (2.12) holds.

For the 2-D cases, by taking a similar ansatz and putting

$$|\nabla\phi|^2 = \frac{1}{2\varepsilon^2}(1-\phi^2)^2$$

into formula (2.7), we finally have (2.13). ∎

More simplifications can be made for problems where periodic boundary conditions are used, since $\int_\Omega \triangle\phi = 0$; the above formula can be further simplified to

$$\chi \approx \frac{1}{4\pi\varepsilon^2}\int_\Omega(\phi^2-1)\phi\, dx. \tag{2.15}$$

The above formula can be compared with the formula (2.7) (with $a = -b = 1$) applied to the periodic boundary condition case:

$$\begin{aligned}\chi &= \frac{1}{4\pi}\int_\Omega \frac{\nabla|\nabla\phi|^2\cdot\nabla\phi}{2|\nabla\phi|^2}\, dx = \frac{1}{8\pi}\int_\Omega \nabla\ln|\nabla\phi|^2\cdot\nabla\phi\, dx \\ &= -\frac{1}{8\pi}\int_\Omega \triangle\phi\ln|\nabla\phi|^2\, dx, \end{aligned} \tag{2.16}$$

where we need only p being monotone from -1 to 1 in Ω instead of being a strictly tanh function.

The previous formula (2.15) has no derivative terms, while the latter one in (2.16) involves no bulk part.

We note that the approximations in (2.12), (2.13), and (2.15) are of spectral accuracy with $\epsilon \to 0$.

Remark 2.5. We note that it is possible to get simplified formulae of other types with the ansatz (2.10). For instance, in the 3-D case, let

$$\tilde{M}_{ij}(\phi) = \sqrt{\frac{35\epsilon}{64\sqrt{2}\pi}} \left((1-\phi^2)\nabla^2\phi + 2\phi\nabla_i\phi\nabla_j\phi\right). \tag{2.17}$$

Then, the Euler number of the surface Γ in the phase field formulation can also be given by

$$\frac{\chi}{2} = \lim_{\epsilon\to 0}\int_\Omega \Lambda(\tilde{M}(\phi))\,dx. \tag{2.18}$$

Note that the change of the constant factors in front of the matrices in (2.11) and (2.17) are due to the use of different weight functions $(1-\tanh^2(x))^2$ and $(1-\tanh^2(x))^4$ in the derivation.

2.6. Formulae for cylindrically symmetric membranes. In [19], we have used the energetic phase field models to compute 3-D vesicle membranes minimizing the bending elastic energy. The numerical examples are for cylindrically symmetric membranes with various different topologies. We now present the Euler number computation in such situations.

In the numerical simulations given in [19], where $\phi(x) \to \tanh(\frac{d(x,\Gamma)}{\sqrt{2}\epsilon})$ as $\epsilon \to 0$, the conventional cylindrical coordinates (r,θ,z) are used. Suppose that the membrane is rotationally invariant with respect to the z-axis, i.e., $\phi = \phi(z,r,\theta) = \phi(z,r)$; then

$$\nabla\phi = \begin{pmatrix} \partial_z\phi \\ \partial_r\phi \\ 0 \end{pmatrix}, \quad \nabla_i\phi\nabla_j\phi = \begin{pmatrix} (\partial_z\phi)^2 & \partial_z\phi\partial_r\phi & 0 \\ \partial_z\phi\partial_r\phi & (\partial_r\phi)^2 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{2.19}$$

and

$$\nabla^2\phi = \begin{pmatrix} \partial^2_{zz}\phi & \partial^2_{zr}\phi & 0 \\ \partial^2_{zr}\phi & \partial^2_{rr}\phi & 0 \\ 0 & 0 & \frac{1}{r}\partial_r\phi \end{pmatrix}. \tag{2.20}$$

If we substitute (2.19) and (2.20) into (2.11), we have that

$$M = \sqrt{\frac{3\epsilon}{8\sqrt{2}\pi}} \begin{pmatrix} \partial^2_{zz}\phi + \frac{2\phi}{1-\phi^2}(\partial_z\phi)^2 & \partial^2_{zr}\phi + \frac{2\phi}{1-\phi^2}\partial_z\phi\partial_r\phi & 0 \\ \partial^2_{zr}\phi + \frac{2\phi}{1-\phi^2}\partial_z\phi\partial_r\phi & \partial^2_{rr}\phi + \frac{2\phi}{1-\phi^2}(\partial_r\phi)^2 & 0 \\ 0 & 0 & \frac{1}{r}\partial_r\phi \end{pmatrix}.$$

Hence F, the sum of the determinants of all principal 2×2 minors, is equal to

$$F(r,z) = \frac{3\epsilon}{8\sqrt{2}\pi r}\partial_r\phi\left\{\partial^2_{zz}\phi + \partial^2_{rr}\phi + \frac{2\phi}{1-\phi^2}[(\partial_z\phi)^2 + (\partial_r\phi)^2]\right\}, \tag{2.21}$$

and

$$\begin{aligned}\frac{\chi}{2} &\approx \int\int 2\pi r F(r,z)\, drdz \\ &\approx \frac{3\epsilon}{4\sqrt{2}} \int\int \partial_r\phi \left\{ \partial^2_{zz}\phi + \partial^2_{rr}\phi + \frac{2\phi}{1-\phi^2}[(\partial_z\phi)^2 + (\partial_r\phi)^2] \right\} drdz \\ &\approx \frac{3\epsilon}{4\sqrt{2}} \int\int \partial_r\phi \left\{ \triangle\phi + \frac{2\phi}{1-\phi^2}|\nabla\phi|^2 \right\} drdz, \end{aligned} \tag{2.22}$$

where the operators ∇ and $\triangle$ are taken in the r-z plane.

If we do not take the ansatz $\phi = \tanh(\frac{d}{\sqrt{2}\epsilon})$, but simply take (2.19) and (2.20) into the formula (2.2), we get a similar formula in the more general case:

$$\frac{\chi}{2} = \frac{1}{4\pi(a-b)} \int_{\Omega(a,b)} \frac{1}{r} \frac{1}{|\nabla\phi|} \partial_r\phi \left\{ \partial^2_{zz}\phi + \partial^2_{rr}\phi - \frac{\nabla|\nabla\phi|^2 \cdot \nabla\phi}{2|\nabla\phi|^4}[(\partial_z\phi)^2 + (\partial_r\phi)^2] \right\} dx,$$

or

$$\frac{\chi}{2} = \frac{1}{2(a-b)} \int\int_{\Omega(a,b)} \frac{1}{|\nabla\phi|} \partial_r\phi \left\{ \triangle\phi - \frac{\nabla|\nabla\phi|^2 \cdot \nabla\phi}{2|\nabla\phi|^2} \right\} drdz. \tag{2.23}$$

We remark here that in the 3-D cylindrically symmetric case, the two curvatures can be calculated by

$$k_1 = \triangle\phi - \frac{\nabla|\nabla\phi|^2 \cdot \nabla\phi}{2|\nabla\phi|^2}, \tag{2.24}$$

$$k_2 = \frac{\partial_r\phi}{r|\nabla\phi|}, \tag{2.25}$$

respectively. One can also use them to derive the formula for the Euler number directly.

Using the partial derivatives, (2.23) can be written as

$$\frac{\chi}{2} = \int\int_{\Omega(a,b)} \frac{\partial_r\phi \left(\partial^2_{rr}\phi(\partial_z\phi)^2 + \partial^2_{zz}\phi(\partial_r\phi)^2 - 2\partial_r\phi\partial_z\phi\partial^2_{rz}\phi \right)}{2(a-b)((\partial_r\phi)^2 + (\partial_z\phi)^2)^{\frac{3}{2}}}\, drdz. \tag{2.26}$$

Using the difference approximation as described in [19], the above integrals can be readily evaluated numerically on a spatial grid.

3. Numerical realizations. In section 2, various formulae were presented for calculating the Euler number in two and three dimensions under different kinds of conditions. In general, we can apply finite difference or spectral methods directly to those formulae to calculate the Euler number. However, as discussed in section 2.3, the singular cases often happen in the process with topology changes, such as surface entanglement. When we apply the Euler number formulae to those singular cases, the numerical values at the singular vertices are always very large, which make the final results inaccurate.

Figure 3.1 illustrates a 3-D singular case. The right-most panel shows the function $F(r,z)$ in formula (2.21) with the r and z coordinates. From the right-most graph, we see the numerical value may even go up to $10,000$. If the Euler number is calculated directly by formula (2.23), although the theoretical value for this shape is 0.7500, we get 1.0671 which is close to the value 1, the Euler number of a sphere.

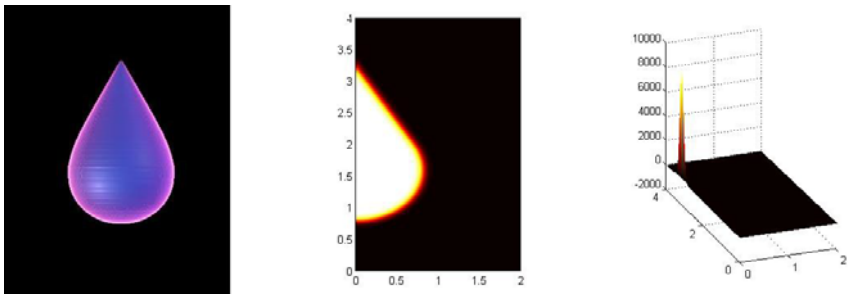


FIG. 3.1. *Singular values at the singular vertex. From left to right: the 3-D picture, the 2-D axis symmetrical r-z planar section, and the singular value around the singular vertex.*

This phenomenon is due to the so-called *singularity compensation* discussed earlier. Because of the finite discrete grid adopted in the numerical scheme, the curvature value at a singular vertex is always regularized from ∞ to a finite number. And in our case, such a change to the finite number provides certain *compensation* to the total Euler number, which can be viewed as putting a small elbow with a large curvature to complement the shape so that the singular shape changes to a regular shape.

A simple but effective way for getting rid of this singularity compensation is to avoid integrating over those singular vertices which are easy to detect, because the numerical values of the integrand at those locations are very big relative to other points. From the formula (2.4), we know that

$$\Lambda(M(x)) = \frac{1}{4\pi(a-b)} p'(d(x,\Gamma))K(x),$$

where $K(x)$ is the Gaussian curvature at a point x. If $p(x) = \tanh(x/(\sqrt{2}\epsilon))$, $p'(d(x,\Gamma)) = (1-p^2(d(x,\Gamma)))/(\sqrt{2}\epsilon)$, then

$$|\Lambda(M(x))| \le \frac{1}{4\sqrt{2}\pi\epsilon(a-b)} K(x).$$

In most cases, it is easy to estimate the possible largest Gaussian curvature value. In the phase field models, the radius of the smallest ball should be at least larger than the band width ϵ. Thus it is natural to choose $K < 1/\epsilon^2$ for the regular points and regard the singular points as those satisfying the condition $\Lambda(M(x)) > 1/(4\epsilon^3\sqrt{2}\pi(a-b))$.

We verify the above argument using the examples shown in Figure 3.2, which represent the same cylindrically symmetric membrane with a singularity at the upper tip whose Euler number can be calculated by formula (2.26). By excluding the singular points with $\Lambda(M(x)) > 200$, based on a 100×200 grid with $h = 0.01$, $\epsilon = 2h = 0.02$, and $\alpha = \frac{\pi}{6}$, the computed Euler number we get is 0.7498, which is a very good approximation of the theoretical value 0.7500 obtained from formula (2.9). For $\alpha = \frac{\pi}{4}$, the computed Euler number 0.8534 is again an accurate approximation of the theoretical value 0.8536.

In the following section, we apply this technique to all our experiments.

4. Applications to some phase field models. In this section, we present two applications of the Euler number developed in the earlier sections for general phase field models. The special examples include the deformation of vesicle membrane configurations minimizing the bending elastic energy and the coarsening of two

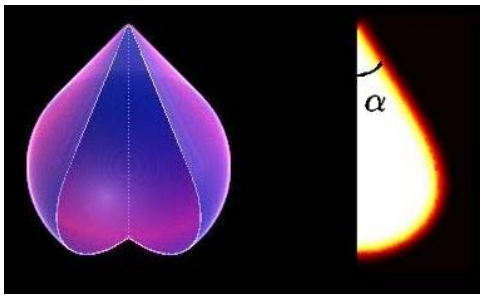


FIG. 3.2. *A special cylindrically symmetric membrane with a singular point.*

Newtonian bubbles in another Newtonian fluid. We expect that our formula can be equally applied to other more complicated and physical examples involving the phase field models as well.

4.1. The model for membranes minimizing the bending elastic energy. We now consider the problem of minimizing the bending elastic energy:

$$E_{\text{elastic}} = \int_\Gamma \frac{k}{2} H^2 \, ds, \tag{4.1}$$

with area and volume constraints. Here, $H = (k_1 + k_2)/2$ is the mean curvature of the membrane surface, with k_1 and k_2 as the principal curvatures. k is the bending rigidity, which can depend on the local heterogeneous concentration of the species (such as protein molecules on the blood cells).

The energetic phase field model studied in [19] is given by the solution of the following modified elastic energy:

$$W(\phi) = \int_\Omega \frac{k\epsilon}{2} \left| \Delta\phi - \frac{1}{\epsilon^2}(\phi^2 - 1)\phi \right|^2 dx. \tag{4.2}$$

As ϵ is taken asymptotically to zero, the minimum of the energy W approaches the original energy (4.1). Moreover, at least when the membrane Γ, viewed as a surface in Ω, is regular enough, we have

$$\phi(x) \tilde{=} \tanh\left(\frac{d(x,\Gamma)}{\sqrt{2}\epsilon}\right) + O(\epsilon^2) \tag{4.3}$$

approximately satisfied for all $x \in \Omega$. Here, $d = d(x,\Gamma)$ is the distance of the point $x \in \Omega$ to the surface Γ. Furthermore, the functional

$$A = \int_\Omega \phi(x) \, dx \tag{4.4}$$

goes to the difference of interior volume and exterior volume, and the function

$$B = \int_\Omega \left[\frac{\epsilon}{2} |\nabla\phi|^2 + \frac{1}{4\epsilon}(\phi^2 - 1)^2 \right] dx \tag{4.5}$$

approaches $2\sqrt{2}\,\text{area}(\Gamma)/3$, or about 0.94 times the area of Γ. We note that for energy minimizing functions,

$$\int_\Omega \frac{\epsilon}{2} |\nabla\phi|^2 dx \approx \int_\Omega \frac{1}{4\epsilon}(\phi^2 - 1)^2 \, dx.$$

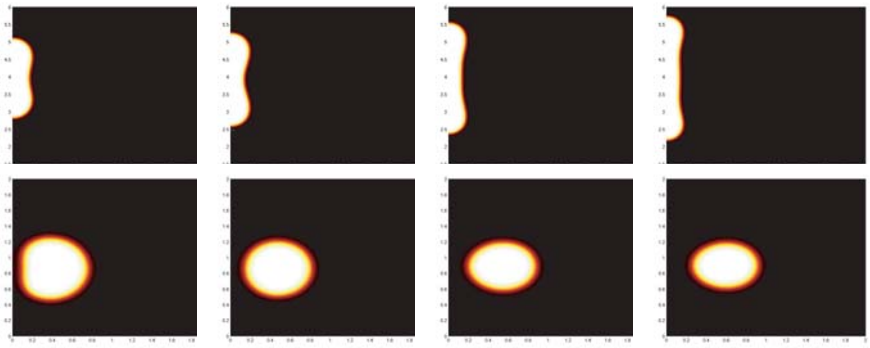


FIG. 4.1. *Deformation of a gourd (top) and a torus (bottom) with areas at* 5.834, 6.258, 6.682, *and* 7.000. *Pictures are displayed in different scales for best view.*

Thus, it is also convenient to take

$$B = \int_\Omega \epsilon |\nabla \phi|^2 dx$$

as the constraint for numerical purposes. For a more detailed asymptotic analysis and a more rigorous convergence analysis, including the convergence of the Euler Lagrange equation, we refer to our subsequent work [17].

It has come to our attention recently that the unconstrained variational approach of using (4.2) to approximate the 2-D Willmore functional (4.1) was also previously used, by the name of a relaxed formulation, in the application of image inpainting (see [23, 10, 32] and the references cited therein). This relaxed formulation was motivated by the Γ-convergence framework of De Giorgi [14] (see also [32]) for general variational problems, although the convergence of such a formulation remains to be rigorously justified.

The numerical simulations in [19] were aimed at the study of the minimizers of the elastic energy $W(\phi)$ under given surface area and volume. By scaling invariance, the volume can be fixed to be a constant, while the area is changed continuously to probe the energy landscape. A set of bubble shapes was discovered, and topological changes were observed along the way only when the configurations were visualized. Our objective here is to show that the Euler number formula (2.26) can be used as an effective tool to automatically detect the topological changes.

The detailed numerical algorithms and numerical simulations of the branches of membrane shape deformations have been presented in [19].

Figure 4.1 shows the deformation of a gourd (the first row, computed with a 40×375 grid and $\epsilon = 1.5h = 0.03$) and a torus (the second row, 100×100 grid, $\epsilon = 1.5h = 0.03$) with increasing surface areas, with volumes fixed at 1.1000. The gourd becomes thinner, and the torus moves further away from its axis. It is obvious that there is no topological change for both the gourd and the torus. This can be seen from the graph of the Euler numbers (Figure 4.2) with the value being kept at 1 for the gourd and 0 for the torus. On the other hand, in the same figure, the graphs of their energy are two intersecting curves, which illustrates the existence of two shapes with totally different topologies but the same energy.

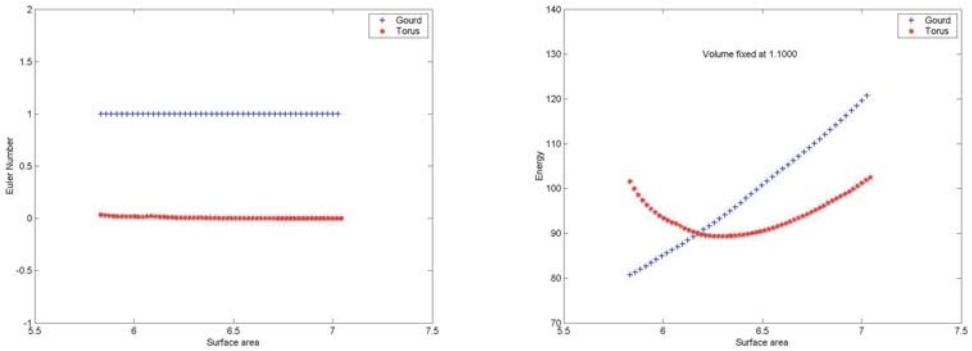


FIG. 4.2. *The energy (left) and the computed Euler number (right) of gourd and torus shapes.*

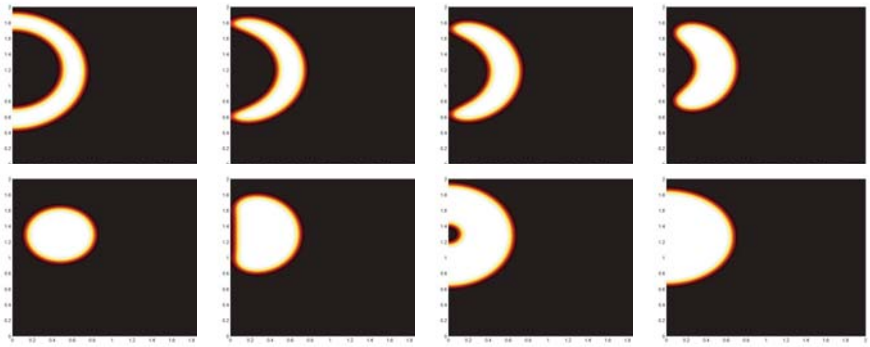


FIG. 4.3. *Deformation of a shell to a pitomba, a bangle, a torus, a longan and finally a ball (cross section view) with areas valued at* 9.811, 9.546, 8.485, 6.894, 6.364, 5.515, 5.303, *and* 5.091.

In the next set of numerical simulations, a rectangular domain with a 100×200 Cartesian grid, with $h = 0.01$, is used and $\epsilon = 2h = 0.02$. Figure 4.3 shows the shape deformations with a decreasing area from 11.2960 to 5.0912 while fixing the volume at 1.1000. The left picture in Figure 4.4 shows the corresponding change of the bending energy with different surface areas. Good resolutions of the interfaces based on the above choices of the computational grid and the parameter values have been demonstrated in [19] for such a solution branch.

In the whole energy minimizing process, the shape always jumps from one branch to another, which results in a discontinuous energy curve. Here the shape of the vesicle changes from a shell to a bangle (with no obvious energy jump), then to a torus (with a noticeable energy jump at the surface area $\beta = 6.6822$), then a longan (with an energy jump at around $\beta = 5.3563$), and finally to a spherical ball (an energy jump at around $\beta = 5.1442$). There are three topological changes during the shape deformation: (1) from shell to a bangle, (2) from torus to a longan, and (3) from longan to a spherical ball. The energy can hardly detect the change from shell to bangle, despite the occurrence of the topology change.

The right picture in Figure 4.4 shows the change of the computed Euler numbers. The graph has exactly three jumps corresponding to the three topological changes. We can see the corresponding Euler numbers 2, 0, 2, and 1 for shell, torus, longan,

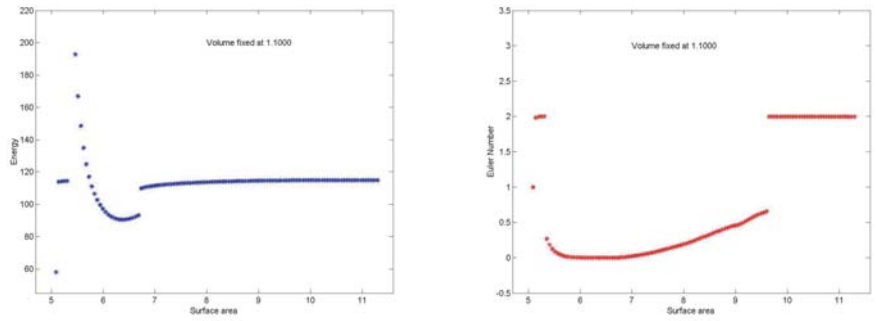


FIG. 4.4. *The energy (left) and the computed Euler number (right) for the deformation of a shell shape with decreasing area.*

and ball shapes, respectively.

The calculation of the Euler numbers used the autoselection algorithm for the best parameters a and b. Observing the second, third, and sixth graphs of Figure 4.3 carefully, it is not hard to understand why we need such an algorithm here, based on the theory stated in section 2.4. As different shapes have different best values of a and b, we thus have different integration areas for the Euler numbers. It may be noticeable that some values are not exactly 0 when the vesicle surface area belongs to the interval $[5.3, 9.8]$. Such small errors are mainly due to the approximation errors of the finite difference scheme and the integration scheme. Fortunately, those errors are always sufficiently small (less than 0.1 in this case) to be distinguishable from the Euler number jumps, which are at least 1. This makes our method very stable and sensitive in detecting the topology changes.

In summary, the above experiments demonstrate that our formula for the Euler number can be successfully used to retrieve topological information and to capture topological events. Of course, the Euler number alone does not completely determine the topology of the interface.

4.2. The phase field model of fluid bubble motion. In the study of the coarsening of the Newtonian bubbles in another Newtonian fluid, the following 2-D simple problem has been considered:

$$
\begin{cases}
\dfrac{\partial u}{\partial t} + (u \cdot \nabla)u - \nu \nabla \cdot D(u) + \nabla p + \lambda \nabla \cdot (\nabla \phi \odot \nabla \phi) = 0, & (x,t) \in Q_T, \\
\nabla \cdot u = 0, & (x,t) \in Q_T, \\
\dfrac{\partial \phi}{\partial t} + (u \cdot \nabla)\phi - \gamma \Delta(\Delta \phi + f(\phi)) = 0, & (x,t) \in Q_T,
\end{cases} \tag{4.6}
$$

with the initial values $u(x,0) = u_0(x), d(x,0) = d_0(x)$ and periodic boundary conditions. Here $f(\phi) = (1-\phi^2)\phi/\epsilon^2$, $D(u) = (\nabla u + (\nabla u)^T)/2$, $\nabla \phi \odot \nabla \phi$ is the induced elastic tensor, with the (j,k)th entry being $\partial_{x_j}\phi \partial_{x_k}\phi$. The above equations have been used in [28] to analyze the motion of bubbles in a Newtonian fluid. These phase field models are also very similar to the liquid crystal model studied in [20, 27, 29].

The above simple system was used to study the free interface motion under the surface tension in the mixture of two Newtonian fluids with the same density and viscosity constants [1, 2, 3, 4, 5, 15, 21, 25, 26, 28, 30, 31, 33, 36, 40]. The system

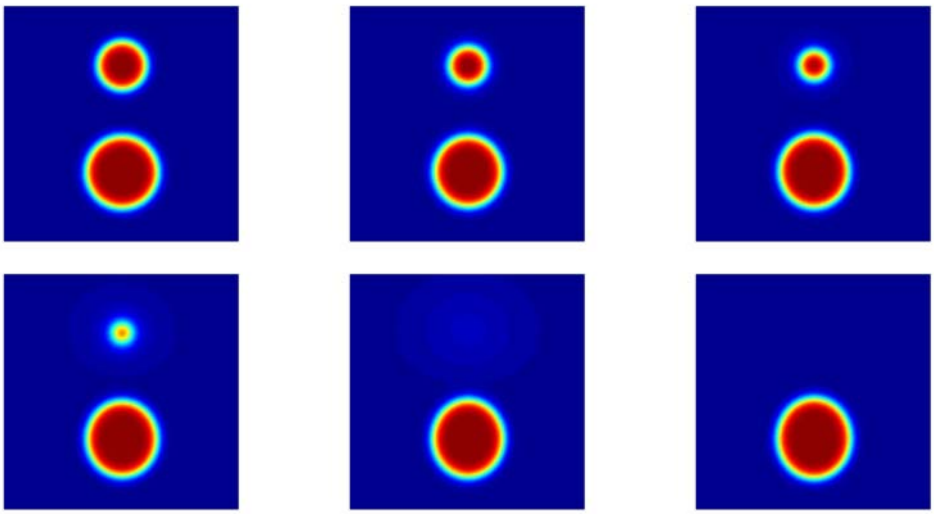


FIG. 4.5. *Deformation of two bubbles in a Newtonian fluid with the time valued at* 0.00, 0.10, 0.18, 0.22, 0.24, 0.28.

satisfies the following energy law:

(4.7)
$$\frac{d}{dt}\int_\Omega \left\{\frac{1}{2}|u|^2 + \frac{\lambda}{2}|\nabla\phi|^2 + \lambda F(\phi)\right\} dx = -\int_\Omega \{\nu|\nabla u|^2 + \gamma\lambda|\nabla(\Delta\phi - f(\phi))|^2\}\, dx.$$

Moreover, the whole system can be viewed as the approximation of the classical sharp interface model with the kinematic and traction-free boundary condition on the free interface [28]. As the transition width ϵ approaches zero, the induced bulk elastic stress term converges to the corresponding surface tension.

To test our formula for computing the Euler number, we solve the above system in two space dimensions via a spectral spatial discretization coupled with a second order semi-implicit-in-time scheme for ϕ and a semi-implicit projection scheme for the Navier–Stokes equations, such as those in [20, 28].

Figure 4.5 shows a special example of the deformation of two Newtonian bubbles in another Newtonian fluid. In this experiment, we take the 128×128 grid, period boundary condition on area $[0, 2\pi] \times [0, 2\pi]$, and $\epsilon = 2.5h = 0.1227$, $\lambda = 10.0$, $\gamma = 3.0$, $\nu = 1.0$. In the simulation, the larger bubble grows at the expense of the shrinkage of the smaller one. In fact, the smaller bubble dissolves into the fluid, while the bigger bubble absorbs from the fluid, similar to the well-known Oswald ripening effects (due to the Cahn–Hillard effect of the phase equation in (4.6)). The total volume of these two bubbles remains constant in time.

The topology change in this simulation can be characterized by the Euler number of the whole configuration. At the beginning, the Euler number of the configuration with the two bubbles is 2. And finally, after the smaller bubble is totally absorbed by the larger bubble, the Euler number of the bubble configuration becomes 1. Figure 4.6 shows the sharp change in the Euler number in this procedure using the formula (2.16).

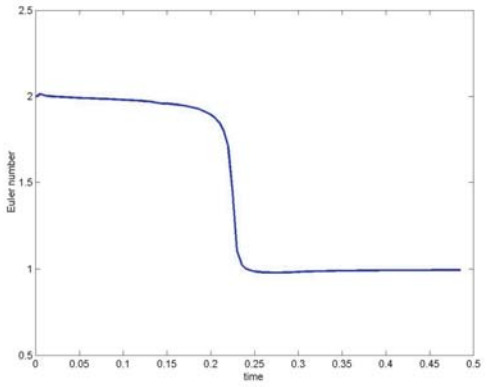


FIG. 4.6. *A plot of the Euler number in time with the annihilation of the small bubble.*

5. Conclusion. While an important advantage of phase field modeling of the interface variation and the interfacial motion is its ability to handle the change of interface topologies in natural and physically meaningful ways, it has also come to our attention that in many practical problems, useful topological information may be needed, and the effective control of the topological transformations may be important. In this paper, mechanisms to retrieve relevant topological information based on the phase field formulations are discussed. In particular, some robust formulae for computing a generalized Euler number of the interface are proposed based on the phase field order parameter ϕ. Using a special ansatz, we also get further simplified formulae. For smooth interfaces, our formulae give desired quantized characterization of the interface genus. When passing through singularities, they give fractional values that generalize the notion of the genus.

As a demonstration, numerical experiments are performed for the cases of a static deformation of a 3-D axial symmetric membrane as well as a time dependent annihilation of fluid bubbles in 2-D space. The experimental results show that the proposed methods for computing the Euler number are very effective and robust in detecting the topological changes. The ideas presented in this paper are very natural and easy to implement for other phase field models and may also be equally applicable to other simulation methods for free boundary and interface problems including the standard level set methods. Rigorous analysis of the formulae derived here based on formal asymptotic analysis are currently underway [17]. We are also working on the problem of taking the Euler number as a constraint within the phase field framework to study and analyze mechanisms in a physical system for controlling and preventing topological changes, should they become desirable.

REFERENCES

[1] D. M. ANDERSON AND G. B. MCFADDEN, *A diffuse-interface description of internal waves in a near-critical fluid*, Phys. Fluids, 9 (1997), pp. 1870–1879.

[2] D. M. ANDERSON, G. B. MCFADDEN, AND A. A. WHEELER, *Diffuse-interface methods in fluid mechanics*, Ann. Rev. Fluid Mech., 30 (1998), pp. 139–165.

[3] T. BLESGEN, *A generalization of the Navier–Stokes equations to two phase flow*, J. Phys. D Appl. Phys., 32 (1999), pp. 1119–1123.

[4] F. Boyer, *Mathematical study of multi-phase flow under shear through order parameter formulation*, Asymptot. Anal., 20 (1999), pp. 175–212.
[5] F. Boyer, *A theoretical and numerical model for the study of incompressible mixture flows*, Comput. & Fluids, 31 (2002), pp. 41–68.
[6] G. Caginalp, *An analysis of a phase field model of a free boundary*, Arch. Ration. Mech. Anal., 92 (1986), pp. 205–245.
[7] J. W. Cahn and S. M. Allen, *A microscopic theory for domain wall motion and its experimental verification in Fe-Al alloy domain growth kinetics*, J. Phys. Colloque, C7 (1977), pp. 51–54.
[8] J. W. Cahn, C. M. Elliott, and A. Novick-Cohen, *The Cahn–Hilliard equation with a concentration dependent mobility: Motion by minus the Laplacian of the mean curvature*, European J. Appl. Math. 7 (1996), pp. 287–301.
[9] J. W. Cahn and J. E. Hillard, *Free energy of a nonuniform system.* I. *Interfacial free energy*, J. Chem. Phys., 28 (1958), pp. 258–267.
[10] T. F. Chan, S. H. Kang, and J. Shen, *Euler's elastica and curvature-based inpainting*, SIAM J. Appl. Math., 63 (2002), pp. 564–592.
[11] Y. C. Chang, T. Y. Hou, B. Merriman, and S. Osher, *A level set formulation of Eulerian interface capturing methods for incompressible fluid flows.*, J. Comput. Phys., 124 (1996), pp. 449–464.
[12] L.-Q. Chen and Y.-Z. Wang, *The continuum field approach to modeling microstructural evolution*, J. Minerals Metals Materials Soc., 48 (1996), pp. 13–18.
[13] L.-Q. Chen, C. Wolverton, V. Vaithyananthan and Z.-K. Liu, *Modeling solid-state phase transformations and microstructure evolution*, MRS Bulletin, 26 (2001), pp. 197–202.
[14] E. De Giorgi, *Some remarks on Gamma-convergence and least squares method*, in Composite Media and Homogenization Theory (ICTP, Trieste, 1990), G. Dal Maso and G. F. Dell Antonio, eds., Birkhäuser Boston, Cambridge, MA, 1990, pp. 135–142.
[15] D. L. Denny and R. L. Pego, *Models of low-speed flow for near-critical fluids with gravitational and capillary effects*, Quart. Appl. Math., 58 (2000), pp. 103–125.
[16] M. P. do Carmo, *Differential Geometry of Curves and Surfaces*, Prentice–Hall, Englewood Cliffs, NJ, 1976.
[17] Q. Du, C. Liu, R. Ryham, and X. Wang, *A phase field formulation of the Willmore problem*, Nonlinearity, 18 (2005), pp. 1249–1267.
[18] Q. Du, C. Liu, R. Ryham, and X. Wang, *Phase field modeling of the spontaneous curvature effect in cell membranes*, Comm. Pure Appl. Anal., 4 (2005), pp. 537–548.
[19] Q. Du, C. Liu, and X. Wang, *A phase field approach in the numerical study of the elastic bending energy for vesicle membranes*, J. Comput. Phys., 198 (2004), pp. 450–468.
[20] Q. Du, B. Guo, and J. Shen, *Fourier spectral approximation to a dissipative system modeling the flow of liquid crystals*, SIAM J. Numer. Anal., 39 (2001), pp. 735–762.
[21] J. E. Dunn and J. Serrin, *On the thermomechanics of interstitial working*, Arch. Ration. Mech. Anal., 88 (1985), pp. 95–133.
[22] C. M. Elliott, *Approximation of curvature dependent interface motion*, in State of the Art in Numerical Analysis, I. Duff and G. A. Watson, eds., Clarendon Press, Oxford, UK, 1997, pp. 407–440.
[23] S. Esedoglu and J. H. Shen, *Digital inpainting based on the Mumford–Shah–Euler image model*, European J. Appl. Math., 13 (2002), pp. 353–370.
[24] W. George and J. Warren, *A Parallel 3D dendritic growth simulator using the phase-field method*, J. Comput. Phys., 177 (2002), pp. 264–283.
[25] M. E. Gurtin, D. Polignone, and J. Viñals, *Two-phase binary fluids and immiscible fluids described by an order parameter*, Math. Models Methods Appl. Sci., 6 (1996), pp. 815–831.
[26] D. Jacqmin, *Calculation of two-phase Navier–Stokes flows using phase-field modeling*, J. Comput. Phys., 155 (1999), pp. 96–127.
[27] F. H. Lin and C. Liu, *Nonparabolic dissipative systems, modeling the flow of liquid crystals*, Comm. Pure Appl. Math., 48 (1995), pp. 501–537.
[28] C. Liu and J. Shen, *A phase field model for the mixture of two incompressible fluids and its approximation by a Fourier-spectral method*, Phys. D, 179 (2003), pp. 211–228.
[29] C. Liu and N. J. Walkington, *Approximation of liquid crystal flows*, SIAM J. Numer. Anal., 37 (2000), pp. 725–741.
[30] C. Liu and N. J. Walkington, *An Eulerian description of fluids containing visco-hyperelastic particles*, Arch. Ration. Mech. Anal., 159 (2001), pp. 229–252.
[31] J. Lowengrub and L. Truskinovsky, *Quasi-incompressible Cahn–Hilliard fluids and topological transitions*, Roy. Soc. London Proc. Ser. A Math. Phys. Eng. Sci., 454 (1998), pp. 2617–2654.

[32] R. March and M. Dozio, *A variational method for the recovery of smooth boundaries*, Image & Vision Comput., 15 (1997), pp. 705–712.
[33] G. B. McFadden, A. A. Wheeler, and D. M. Anderson, *Thin interface asymptotics for an energy/entropy approach to phase-field models with unequal conductivities*, Phys. D, 144 (2000), pp. 154–168.
[34] G. B. McFadden, A. A. Wheeler, R. J. Braun, S. R. Coriell, and R. F. Sekerka, *Phase-field models for anisotropic interfaces*, Phys. Rev. E (3), 48 (1993), pp. 2016–2024.
[35] W. W. Mullins and R. F. Sekerka, *On the thermodynamics of crystalline solids*, J. Chem. Phys., 82 (1985), pp. 5192–5202.
[36] T. Qian, X. P. Wang, and P. Sheng, *Generalized Navier boundary condition for the moving contact line*, Comm. Math. Sci., 1 (2003), pp. 333–341.
[37] J. Rubinstein, P. Sternberg, and J. B. Keller, *Reaction-diffusion processes and evolution to harmonic maps*, SIAM J. Appl. Math., 49 (1989), pp. 1722–1733.
[38] J. E. Taylor and J. W. Cahn, *Linking anisotropic sharp and diffuse surface motion laws via gradient flows*, J. Statist. Phys., 77 (1994), pp. 183–197.
[39] P. Yu, S.-Y. Hu, L.-Q. Chen, and Q. Du, *An iterative-perturbation scheme for treating inhomogeneous elasticity in phase field models*, J. Comput. Phys., 208 (2005), pp. 34–50.
[40] P. Yue, J. Feng, C. Liu, and J. Shen, *A diffuse-interface method for simulating two-phase flows of complex fluids*, J. Fluid Mech., 515 (2005), pp. 293–317.

SIAM J. APPL. MATH.
Vol. 65, No. 6, pp. 1933–1951

THE FOCUS-CENTER-LIMIT CYCLE BIFURCATION IN SYMMETRIC 3D PIECEWISE LINEAR SYSTEMS*

EMILIO FREIRE†, ENRIQUE PONCE†, AND JAVIER ROS†

Abstract. The birth of limit cycles in 3D (three-dimensional) piecewise linear systems for the relevant case of symmetrical oscillators is considered. A technique already used by the authors in planar systems is extended to cope with 3D systems, where a greater complexity is involved.

Under some given nondegeneracy conditions, the corresponding theorem characterizing the bifurcation is stated. In terms of the deviation from the critical value of the bifurcation parameter, expressions in the form of power series for the period, amplitude, and the characteristic multipliers of the bifurcating limit cycle are also obtained.

The results are applied to accurately predict the birth of symmetrical periodic oscillations in a 3D electronic circuit genealogically related to the classical Van der Pol oscillator.

Key words. piecewise linear systems, bifurcation theory, limit cycles

AMS subject classifications. 37G15, 34C15

DOI. 10.1137/040606107

1. Introduction and main results. Piecewise linear modeling of nonlinear dynamical systems is especially successful in some engineering problems, such as the analysis and design of electronic oscillators or control systems (see, e.g., [CFPT02]). However, in the framework of piecewise linear systems, there are no general bifurcation results explaining the appearance or disappearance of self-sustained oscillations, as is the case for the Hopf bifurcation theorem in the context of differentiable systems. Thus, the authors gave in [FPR99] a complete characterization of the focus-center-limit cycle bifurcation for symmetric planar piecewise linear systems. Now we show how the corresponding result can be extended to the 3D case.

We consider a common situation in applications, namely, dynamical systems defined by piecewise continuous vector fields with three linear zones and two parallel frontiers. Furthermore, it is assumed that such systems show symmetry with respect to the origin; that is, if we put them in the form $d\mathbf{x}/d\tau = f(\mathbf{x})$ with $\mathbf{x} \in \mathbb{R}^3$, they satisfy $f(-\mathbf{x}) = -f(\mathbf{x})$. In particular, $f(0) = 0$, and so the origin is an equilibrium point for all values of the parameters. By means of a linear change of variables, it is always possible to suppose that the frontiers are the planes $\Sigma_1 = \{\mathbf{x} \in \mathbb{R}^3 : x_1 = 1\}$ and $\Sigma_{-1} = \{\mathbf{x} \in \mathbb{R}^3 : x_1 = -1\}$. We denote by L (left), C (central), and R (right) the regions of $\mathbb{R}^3$ at which $x_1 < -1$, $|x_1| \leq 1$, and $x_1 > 1$, respectively, hold.

To be more precise, we consider systems expressed as follows:

$$\dot{\mathbf{x}} = \begin{cases} A_L\mathbf{x} - \mathbf{b} & \text{if } x_1 < -1, \\ A_C\mathbf{x} & \text{if } |x_1| \leq 1, \\ A_L\mathbf{x} + \mathbf{b} & \text{if } x_1 > 1, \end{cases} \tag{1.1}$$

*Received by the editors April 1, 2004; accepted for publication (in revised form) February 2, 2005; published electronically August 3, 2005. This research was partially supported by grants DPI2000-1218-C04-04, BFM2001-2668, and BFM2003-00336 of the Spanish Ministry of Science and Technology. The authors were also supported by Junta de Andalucía, as members of the research group TIC-130, in the budget corresponding to 2003.

http://www.siam.org/journals/siap/65-6/60610.html

†Departamento Matemática Aplicada II, Universidad de Sevilla, Escuela Superior de Ingenierosm, Camino de los Descubrimientos s/n, 41092, SEVILLA, Spain (emilio@ma2.us.es, enrique@ma2.us.es, jros@platero.eup.us.es).

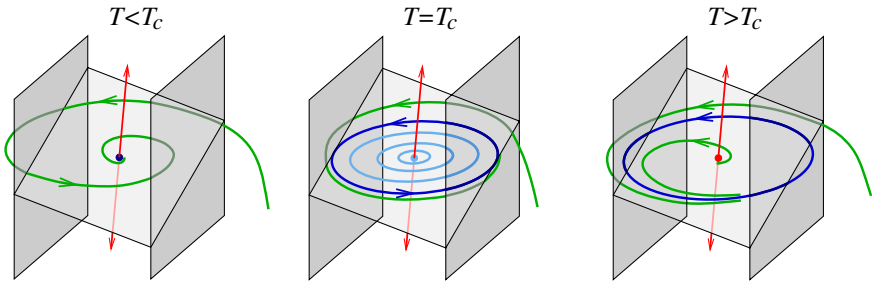


FIG. 1. *The focus-center-limit cycle bifurcation in the case $D > 0$, $\gamma > 0$. The focal plane and the complementary one-dimensional invariant manifold at the origin are shown, along with the two parallel planes which separate the three linear regions. In the situation sketched, as deduced from Theorem* 1.1*, the bifurcating limit cycle is of saddle type.*

where we have taken advantage of the continuity and symmetry of the vector field involved; in particular, the matrices A_L and A_C differ only in their first columns.

From Proposition 16 of [CFPT02], under the generic condition of observability, every system (1.1) can be written in the generalized Liénard form

$$\frac{d}{d\tau}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} t & -1 & 0 \\ m & 0 & -1 \\ d & 0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} T-t \\ M-m \\ D-d \end{bmatrix}\operatorname{sat}(x_1), \tag{1.2}$$

where $\operatorname{sat}(x_1)$ is the *normalized saturation*

$$\operatorname{sat}(x_1) = \begin{cases} -1, & x_1 \le -1, \\ x_1, & |x_1| < 1, \\ 1, & x_1 \ge 1, \end{cases}$$

so that, regarding system (1.1), we have

$$A_L = \begin{bmatrix} t & -1 & 0 \\ m & 0 & -1 \\ d & 0 & 0 \end{bmatrix}, \qquad A_C = \begin{bmatrix} T & -1 & 0 \\ M & 0 & -1 \\ D & 0 & 0 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} T-t \\ M-m \\ D-d \end{bmatrix}.$$

Note that system (1.2) is a particular instance of the more general Lur'e form

$$\frac{d\mathbf{x}}{d\tau} = A\mathbf{x} + \mathbf{b}\operatorname{sat}(\mathbf{c}^T\mathbf{x})$$

for the case $A = A_L$ and $\mathbf{c} = \mathbf{e_1}$, where $\mathbf{e}_1$ stands for the first vector of the canonical basis.

Clearly, the parameters t, m, d and T, M, D stand for the trace, the sum of principal minors of order two, and the determinant of each matrix, and they completely determine the dynamics of the system.

Choosing T as the bifurcation parameter, for the critical value $T_c = D/M$ with $M > 0$, system (1.2) has a linear center in the zone C (see Figure 1); that is, the matrix A_C has a pair of pure imaginary eigenvalues. We want to analyze whether a limit cycle bifurcates from this configuration as the bifurcation parameter T varies. Note the similarities with the classical Hopf bifurcation scenario.

It will be useful, in order to know the stability of such a limit cycle, to estimate the characteristic multipliers of the limit cycle, that is, the eigenvalues of the derivative of a Poincaré return map defined in an adequate section of the phase space. We will denote the logarithms of these characteristic multipliers by μ_r and μ_a, from radial and axial, respectively. Our main result is the following.

THEOREM 1.1. *Let us consider system* (1.2) *with* $M > 0$, $T_c = D/M$, *and* $\gamma = DM - Dm + dM - tM^2 \neq 0$. *For* $T = T_c$ *the system undergoes a focus-center-limit cycle bifurcation; that is, from the lineal center configuration in the central zone, which exists for* $T = T_c$, *one limit cycle appears for* $\gamma(T - T_c) > 0$ *and* $T - T_c$ *sufficiently small.*

The amplitude "a" (measured as the maximum of $|x_1|$*), the period* P_{er}, *and the logarithms of characteristic multipliers* μ_r *and* μ_a *of the periodic orbit are analytic functions at* 0, *in the variable* $(T - T_c)^{1/3}$; *namely,*

$$
\begin{aligned}
a &= 1 + \frac{(6\pi)^{2/3} M^{4/3}}{8\gamma^{2/3}} (T - T_c)^{2/3} + \frac{(6\pi^4)^{1/3} a_4}{960 M^{1/3} \gamma^{7/3}} (T - T_c)^{4/3} + O\left(T - T_c\right)^{5/3}, \\
P_{er} &= \frac{2\pi}{\sqrt{M}} + \frac{\pi(M - m)\sqrt{M}}{\gamma} (T - T_c) - \frac{6^{2/3} \pi^{5/3} M^{5/6} P_5}{20\gamma^{8/3}} (T - T_c)^{5/3} + O\left(T - T_c\right)^2, \\
\mu_r &= -\frac{(48\pi)^{1/3} M^{7/6} \gamma^{2/3}}{D^2 + M^3} (T - T_c)^{1/3} + O\left(T - T_c\right)^{2/3}, \\
\mu_a &= \frac{2\pi D}{M^{3/2}} + \frac{(48\pi)^{1/3}}{M^{5/6}} \left(\frac{Mt - D}{\gamma^{1/3}} + \frac{M^2 \gamma^{2/3}}{D^2 + M^3} \right) (T - T_c)^{1/3} + O\left(T - T_c\right)^{2/3},
\end{aligned}
$$

where

$$
\begin{aligned}
a_4 &= -120 t M^5 + \left(120D + 2t^3 + 21mt + 72d\right) M^4 \\
&\quad + \left[-\left(93m + 27t^2\right) D + \left(27m - 2t^2\right) d \right] M^3 + \left(2t^2 m + 25dt - 27m^2\right) DM^2 \\
&\quad + \left[25D^3 + 23\left(mt - d\right) D^2\right] M - 25mD^3, \\
P_5 &= \left[M\left(M - m\right)^2 + \left(Mt - d\right)^2\right] \left(Mt - D\right).
\end{aligned}
$$

In particular, if $\gamma > 0$ *and* $D < 0$, *then the limit cycle bifurcates for* $T > T_c$ *and is orbitally asymptotically stable.*

This theorem describes a codimension-one bifurcation, similar to the Hopf bifurcation of differentiable dynamics (see [CH82]), but some differences should be noted. In particular, the expressions characterizing the bifurcation are in terms of the parameter to the one third power instead of the one half power, and, more important, the limit cycle amplitude's leading order is $O(1)$. Thus, the stability change of the origin is accompanied by the abrupt appearance of a limit cycle of significant size. This comment also applies to the planar case, as appeared in [Kr87] and [FPR99].

When the coefficient γ is not equal to zero, it allows a complete characterization of the bifurcation criticality. Its role is analogous to the coefficient of the cubic term in the Poincaré–Andronov–Hopf normal form. When $\gamma = 0$, the bifurcation is of higher codimension, requiring a specific treatment that will appear elsewhere.

We want to remark that it is possible, with the same techniques, to obtain similar bifurcation results for the asymmetric case of single-sided saturation. Thus, the proposed methodology is able to cope with a wider class of piecewise linear systems.

The rest of the paper is structured as follows. In section 2, we show how the above result can be useful for accurately predicting the birth of symmetrical periodic

oscillations in a tridimensional electronic circuit, which can be built by taking a Van der Pol oscillator as starting point. The proof of Theorem 1.1 is given in section 3.

2. Predicting the onset of symmetrical periodic oscillations in a 3D electronic circuit. In this section, we consider the electronic circuit of Figure 2(a), genealogically related with the classical Van der Pol oscillator, in order to show the applicability of our results. Regarding this circuit, the nonlinear conductance NL is its active element, implemented by means of an operational amplifier with the feedback structure of Figure 2(b), and the current-voltage characteristic is shown in Figure 2(c). Note that we are dealing with a nonlinearity characteristic qualitatively similar to the cubic one appearing in the classical Rayleigh–Van der Pol oscillator. In fact, if we eliminate the capacitor C_2 and make $R = R_0 = 0$, then the resulting planar circuit could be thought of as a modern electronic realization of such classical oscillators; see [Kr87] and [FPR99].

Thus, the 3D circuit of Figure 2 can be built by adding the capacitance C_2 to a bidimensional oscillator circuit. In the context of chaotic circuits, such topology was originally proposed in [SYM81], and was studied afterwards in [FGA84] and [FRGP93] in the case $R_0 = 0$ and assuming a nonlinear positive conductance for the resistor R. With slight modifications, this circuit has been extensively studied in the last two decades; see [GK92] or [HBCJM91]. Taking $R_0 = 0$ and substituting the nonlinear element by the so-called Chua diode, many papers have also been written; see [CWHZ93], [Ma93], and references therein. Anyway, the onset of symmetrical periodic oscillations was never accurately predicted, since in most cases the circuit was analyzed by taking polynomial approximations. Thus, the rapid bifurcation for the limit cycle observed in practice was never justified.

It should be remarked that the characteristic of Chua's diode is qualitatively similar to the one presented in Figure 2(c) but the zone of negative slope is made up by three pieces with two different slopes. For that, at least two subcircuits with operational amplifiers like those shown in Figure 2(b) are needed. Thus, the Chua circuit characteristic has five linear segments instead of only three, as in our case. However, in modeling Chua's circuit, usually only the three innermost pieces are represented, since the two outermost pieces of positive slope are not physically used; see [Ke93].

As stated in [Kr87] and [FPR99], there exists an excellent agreement between the actual response of the nonlinear device NL in the circuit and its symmetric piecewise

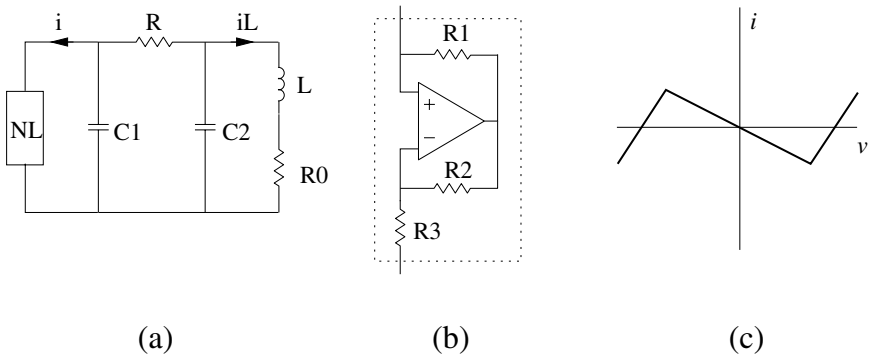


FIG. 2. (a) *The 3D electronic circuit.* (b) *Implementation of the nonlinear conductance NL.* (c) *Piecewise linear current-voltage characteristic of NL.*

linear mathematical model. Therefore, we are led to consider the piecewise linear dynamical system

$$\begin{aligned} C_1\frac{dv_1}{d\tau} &= \frac{v_2-v_1}{R}-i(v_1),\\ C_2\frac{dv_2}{d\tau} &= \frac{v_1-v_2}{R}-i_L,\\ L\frac{di_L}{d\tau} &= v_2-R_0i_L, \end{aligned} \tag{2.1}$$

where v_1 and v_2 are the voltages across the capacitors C_1 and C_2, respectively, while i_L is the current through the inductance. The nonlinear current-voltage characteristic is

$$i(v_1)=\frac{v_1-f(v_1)}{R_1} \quad \text{with} \quad f(v_1)=\begin{cases} E\,\text{sign}(v_1), & |v_1|>E/\sigma,\\ \sigma v_1, & |v_1|\le E/\sigma, \end{cases}$$

where

$$\sigma=1+\frac{R_2}{R_3}$$

is the gain of the operational amplifier configured (using feedback) as a noninverting amplifier and E is its saturation voltage.

With the following linear change of variables and time rescaling,

$$v_1=\frac{E}{\sigma}y_1, \quad v_2=\frac{E}{\sigma}y_2, \quad i_L=\frac{E}{\sigma}\sqrt{\frac{C_2}{L}}y_3, \quad \tau=RC_1\bar{\tau}, \tag{2.2}$$

and defining the following five nonnegative dimensionless parameters,

$$r=\frac{R}{R_1}, \quad c=\frac{C_1}{C_2}, \quad \mu=(\sigma-1)\frac{R}{R_1}=\frac{RR_2}{R_1R_3}, \quad \rho=\frac{R^2C_1^2}{LC_2}, \quad \kappa=\frac{RR_0C_1}{L}, \tag{2.3}$$

we can express system (2.1) as follows:

$$\frac{d}{d\bar{\tau}}\begin{bmatrix} y_1\\ y_2\\ y_3 \end{bmatrix}=\begin{bmatrix} -r-1 & 1 & 0\\ c & -c & -\sqrt{\rho}\\ 0 & \sqrt{\rho} & -\kappa \end{bmatrix}\begin{bmatrix} y_1\\ y_2\\ y_3 \end{bmatrix}+\begin{bmatrix} \mu+r\\ 0\\ 0 \end{bmatrix}\text{sat}(y_1). \tag{2.4}$$

For the subsequent analysis, we will choose μ and ρ as the main bifurcation parameters. In practice, to detect the bifurcation in a experimental way, it is better to tune the parameter μ by means of a variable resistor R_2, which is equivalent to varying the gain σ.

The observability matrix for system (2.1) is

$$\mathcal{O}=\begin{bmatrix} \mathbf{e}_1^T\\ \mathbf{e}_1^TA\\ \mathbf{e}_1^TA^2 \end{bmatrix}=\begin{bmatrix} 1 & 0 & 0\\ -r-1 & 1 & 0\\ (r+1)^2+c & -c-r-1 & -\sqrt{\rho} \end{bmatrix},$$

which has full rank for all the values of components of the circuit. From Proposition 16 of [CFPT02], system (2.1) can be expressed in Liénard's generalized form (1.2) with the following values:

$$\begin{aligned} &T=\mu-c-\kappa-1, && t=-r-c-\kappa-1<0,\\ &M=(c+1)\kappa-(c+\kappa)\mu+\rho, && m=(c+1)\kappa+(c+\kappa)r+\rho>0,\\ &D=(c\kappa+\rho)\mu-\rho, && d=-(c\kappa+\rho)r-\rho<0. \end{aligned} \tag{2.5}$$

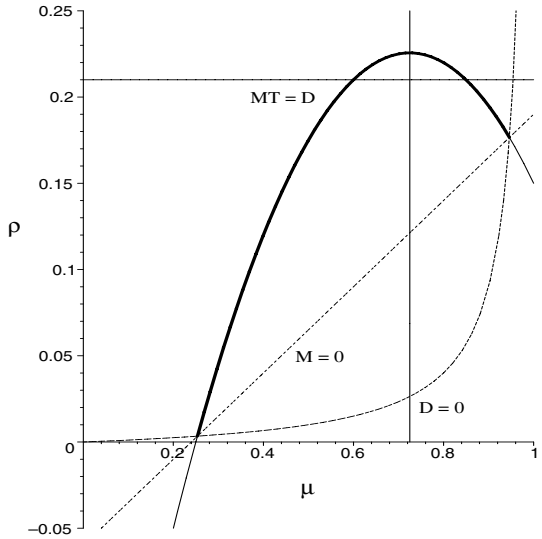


FIG. 3. *The parabolic arc (thick line) in the plane (μ, ρ) corresponding to the bifurcation locus of Proposition* 2.1 *for $c = 0.2$ and $\kappa = 0.05$. The horizontal line indicates the path followed as μ varies for a fixed value of ρ. The dashed line represents points with $D = 0$, so that above it we have $D < 0$. At the dotted straight line we have $M = 0$, and above this line we have $M > 0$. The vertical line corresponds to $\mu = \mu_*$.*

Note that these coefficients are the linear invariants of the two matrices involved, so that their computation is straightforward, and that it is not necessary to explicitly compute the linear change of variables required to get the Liénard form for applying Theorem 1.1.

The equation $MT - D = 0$ leads to

$$(c+\kappa)\mu^2 - \left[(c+\kappa)^2 + c + 2\kappa\right]\mu + \rho(c+\kappa) + \kappa\,(c+1)\,(c+\kappa+1) = 0, \tag{2.6}$$

which can be rewritten as $(\mu - \mu_*)^2 + \rho - \rho_* = 0$, where

$$\begin{aligned} \mu_* &= 1 + \frac{(c+\kappa)^2 - c}{2(c+\kappa)}, \\ \rho_* &= {\mu_*}^2 - \kappa(c+1)\left(1 + \frac{1}{c+\kappa}\right) \end{aligned} \tag{2.7}$$

represent the coordinates in the (μ, ρ)-plane of the vertex of the quadratic (2.6); see Figure 3. Now the application of Theorem 1.1 allows us to state the following result.

PROPOSITION 2.1. *Let us consider system* (2.4) *and assume that $c > 0$ and the parameter κ satisfies*

$$0 < \kappa < \kappa_{\max}(c) = \frac{\sqrt{c^2 + c} - c}{2}. \tag{2.8}$$

Then the system undergoes the focus-center-limit cycle bifurcation described in Theorem 1.1 *at the points of the (μ, ρ)-plane belonging to the parabolic arc defined by the quadratic equation*

$$(\mu - \mu_*)^2 + \rho - \rho_* = 0 \tag{2.9}$$

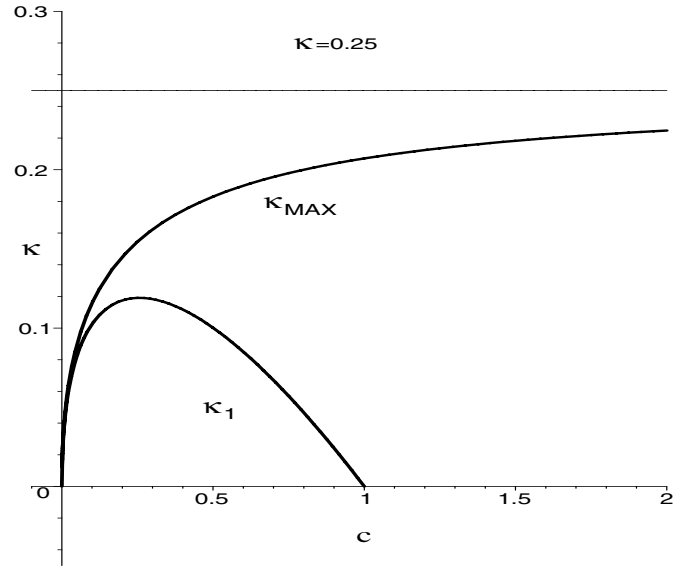


FIG. 4. *The graphs of the functions* $\kappa_{\max}(c)$ *and* $\kappa_1(c)$*, which determine different regions in the plane* (c,κ) *as described in statements* (a) *and* (b) *of Proposition* 2.1. *Note the horizontal asymptote at* $\kappa = 1/4$.

and satisfying

$$\rho > (c+\kappa)\mu - (c+1)\kappa. \tag{2.10}$$

The endpoints of the above parabolic arc are

$$(\mu_1,\rho_1) = \left(1 - \frac{c+\underline{c}}{2(c+\kappa)}, \frac{c(1-2\kappa)-\underline{c}}{2}\right), \quad (\mu_2,\rho_2) = \left(1 - \frac{c-\underline{c}}{2(c+\kappa)}, \frac{c(1-2\kappa)+\underline{c}}{2}\right),$$

where

$$\underline{c} = \sqrt{c^2(1-2\kappa)^2 - 4c(c+1)\kappa^2}.$$

In the points of the above parabolic arc, the inequality $D < 0$ *holds, and the following cases arise:*

(a) *If* $0 < c < 1$ *and* $0 < \kappa < \kappa_1$*, where* $\kappa_1 = \kappa_1(c)$ *is the only positive root of the quartic*

$$(c+\kappa)^4 + 4c\kappa(c+\kappa) - c^2 = 0, \tag{2.11}$$

then $\mu_1 < \mu_* < \mu_2$ *and two subcases appear; see Figure* 4.

(a.1) *If* $\mu_1 < \mu < \mu_*$*, then at the bifurcation points of the parabolic arc given by* (2.9)–(2.10) *one has* $\gamma > 0$*. Consequently, when* ρ *varies, the bifurcation is supercritical and the limit cycle is orbitally asymptotically stable.*

(a.2) *If* $\mu_* < \mu < \mu_2$*, then* $\gamma < 0$ *at the bifurcation points of the parabolic arc. Here, when* ρ *varies, the bifurcation is subcritical and the limit cycle is unstable.*

(b) *If* $0 < c < 1$ *and* $\kappa \geq \kappa_1$*, or* $c \geq 1$*, then all the bifurcation points of the parabolic arc* (2.9)–(2.10) *satisfy* $\gamma > 0$*. Therefore, the bifurcation is supercritical and the bifurcating limit cycle is orbitally asymptotically stable.*

TABLE 2.1
List of components for the circuit.

$C = C_1 = C_2$	100 nF
L	220 mH
R_1	10 kΩ
R_3	2200 Ω
R	1 kΩ
R_0	220 Ω

Proof. Conditions $T = T_c$ and $M > 0$ of Theorem 1.1 lead to $MT - D = 0$, which is equivalent to (2.9), and to (2.10). After some manipulations, we get the inequality

$$(c+\kappa)\mu^2 - (c+2\kappa)\mu + (c+1)\kappa < 0,$$

whose discriminant, namely $c^2 - 4c^2\kappa - 4c\kappa^2$, is positive due to (2.8). In fact, this expression coincides with $\underline{c}^2$. The endpoints of the parabolic arc can be obtained by solving the equation $M = 0$ and (2.9).

To show that $D < 0$ at the bifurcation values, as we are working at points where $MT - D = 0$ along with $M > 0$, it suffices to show that $T < 0$, which is a trivial task.

To prove statements (a) and (b), it is enough to study the sign of the coefficient γ in Theorem 1. Using the condition $MT - D = 0$, we have

$$\gamma = MT(M-m) + M(d - tM) = M[T(M-m) + d - tM],$$

and with $M > 0$ we get $\operatorname{sign}(\gamma) = \operatorname{sign}[T(M-m) + d - tM]$. Thus, using (2.5), (2.6), and canceling a factor $r + \mu > 0$, we conclude that

$$\operatorname{sign}(\gamma) = \operatorname{sign}\left[(c+\kappa)^2 + c + 2\kappa - 2(c+\kappa)\mu\right] = \operatorname{sign}(\mu_* - \mu). \tag{2.12}$$

Assume now that $0 < c < 1$ and $0 < \kappa < \kappa_1$. Thus, the left-hand side of (2.11) is negative, which implies $(c+\kappa)^2 < \underline{c}$. Then $\mu_1 < \mu_* < \mu_2$, and statement (a) follows.

When $c \le 1$ and $\kappa \ge \kappa_1$, we have $(c+\kappa)^2 \ge \underline{c}$. If $c > 1$, then we have $(c+\kappa)^2 > c > \underline{c}$. In both cases, we conclude that $\mu_* \ge \mu_2$, and statement (b) follows. ∎

For the sake of completeness, if we define, for $0 < c < 1$, the constants

$$q_1 = \sqrt[3]{27c + 26 + 3\sqrt{81c^2 + 156c + 75}} > 0, \qquad q_2 = q_1 + \frac{1}{q_1} - 4 > 0,$$

we obtain

$$\kappa_1(c) = \sqrt{\frac{c}{6}q_2} + \sqrt{c\sqrt{\frac{6c}{q_2}} - \frac{cq_2}{6} - 2c} - c,$$

which is represented for $0 < c < 1$ in Figure 4.

The above proposition enables us to design the electronic oscillator by choosing adequately the component values of the circuit. In particular, in order to minimize the signal distortion from the sinusoidal wave form, one must select parameters not far from the bifurcation curve where the onset of periodic oscillations has been predicted.

To assess the accuracy of piecewise linear modeling for this circuit, a SPICE implementation of the circuit was made; see [QNPS93]. The values chosen for the components are in Table 2.1, while the operational amplifier used was an LM324,

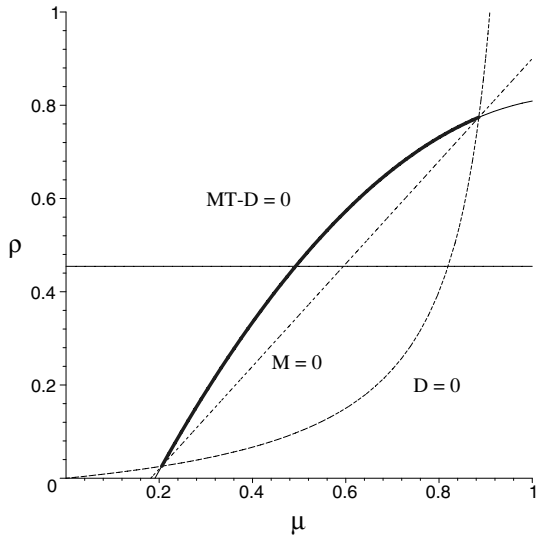


FIG. 5. *The parabolic arc (thick line) in plane* (μ, ρ) *predicted by Proposition* 2.1 *for* $\kappa = 0.1$. *The horizontal line indicates the path followed as* μ *varies for the fixed value of* ρ *used in the simulations. The dashed line represents points with* $D = 0$, *so that above it we have* $D < 0$. *The dotted straight line indicates points with* $M = 0$, *and above it we have* $M > 0$.

with a supply voltage of 9V and a measured saturation voltage of 8.5V (with slight variations around).

For these values, we have

$$c = \frac{C_1}{C_2} = 1, \quad \kappa = \frac{RR_0C}{L} = 0.1 < \frac{\sqrt{2}-1}{2} \approx 0.2071,$$

so that we can apply Proposition 2.1 and, in particular, its statement (b). Note that

$$\mu = \frac{RR_2}{R_1R_3} \approx \frac{R_2}{22000}, \quad \rho = \frac{R^2C}{L} \approx 0.4545,$$

so that by varying R_2 we move μ, describing a horizontal path that crosses the curve corresponding to the locus of bifurcation points, as shown in Figure 5. For the above value of ρ, the bifurcation takes place for the value $\bar{\mu} \approx 0.4924$, in accordance with (2.6), that corresponds with the value $R_2 \approx 10833\Omega$, and oscillations will appear by increasing R_2 above this critical value.

In Figures 6 and 7, we show the comparison between some experimental results taken from a SPICE simulation, once put into dimensionless form, and the predictions of Theorem 1.1 for the amplitude and the period of the bifurcating limit cycle. The excellent agreement achieved validates the piecewise linear model assumed for the operational amplifier nonlinear characteristic.

3. Proof of Theorem 1.1. In this section we provide the results necessary to prove Theorem 1.1.

For the critical value of the bifurcation parameter $T_c = D/M$, the matrix A_C has a pair of imaginary eigenvalues, so that for T in a neighborhood of T_c the eigenvalues of A_C will be $\alpha \pm i\beta$ and $\delta \in \mathbb{R}$. The characteristic polynomial of A_C is

$$p(\lambda) = \det(A_C - \lambda I) = -\lambda^3 + T\lambda^2 - M\lambda + D,$$

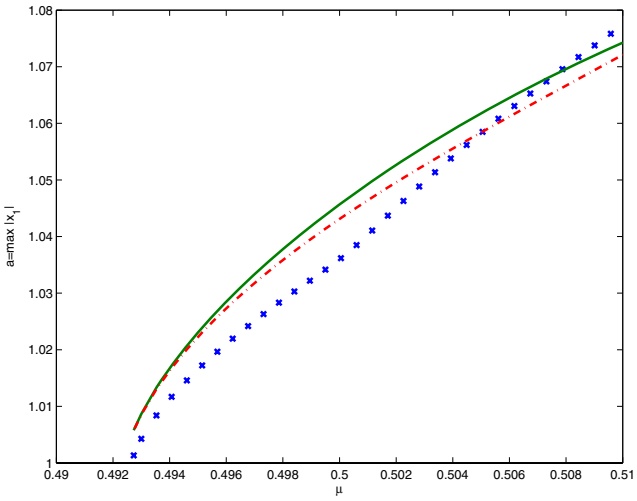


FIG. 6. *Comparison for amplitude between SPICE simulation data (× × ×), the expression corresponding to the two first non-null terms of Theorem* 1.1 *(—), and three non-null terms (-·-).*

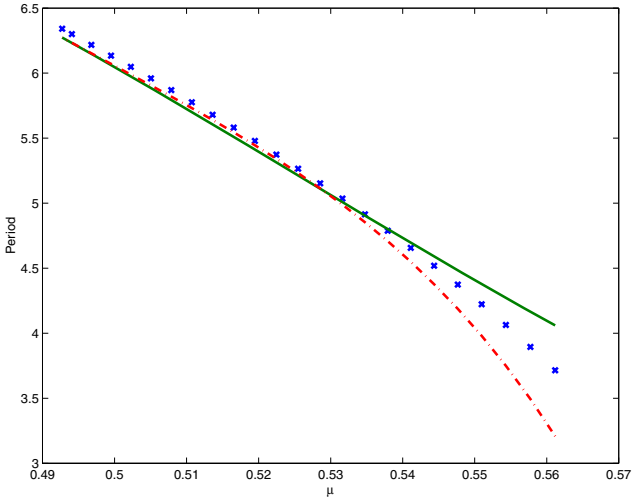


FIG. 7. *Comparison for period between SPICE simulation data (× × ×), the expression corresponding to the two first non-null terms of Theorem* 1.1 *(—), and three non-null terms (-·-).*

and thus

$$\begin{aligned} T &= \delta + 2\alpha, \\ M &= 2\alpha\delta + \alpha^2 + \beta^2, \\ D &= \delta(\alpha^2 + \beta^2). \end{aligned} \tag{3.1}$$

When $\alpha = 0$ and $\beta > 0$, or equivalently $D = MT$ and $M > 0$, system (1.2) has a linear center contained in an invariant plane given by $\delta^2 x_1 - \delta x_2 + x_3 = 0$. Additionally, the outermost periodic orbit of the center is tangent to the planes Σ_1 and Σ_{-1} at the points $[1, \delta, 0]^T$ and $[-1, -\delta, 0]^T$, respectively. Consequently, the time

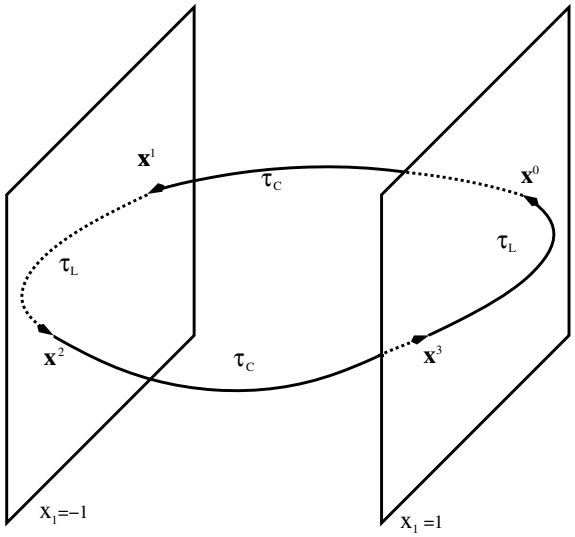


FIG. 8. *Sketch of a symmetrical periodic orbit which uses the three linear zones of system* (1.2).

spent by this orbit in going from $\mathbf{x}^0$ to $\mathbf{x}^1$ is $\tau_C = \pi/\beta$ in the zone C, and obviously $\tau_L = 0$ in the zone L.

We want to analyze the possible bifurcation of a limit cycle from the linear center in the zone C. (Obviously, it should be born from the outermost periodic orbit of the center.) As system (1.2) is linear in every zone, it is possible to obtain its solutions explicitly, and to identify symmetrical periodic solutions of the system living in the three zones with the solutions of the equations

$$\begin{aligned} e^{A_C\tau_C}\mathbf{x}^0 - \mathbf{x}^1 &= \mathbf{0}, \\ e^{A_L\tau_L}\mathbf{x}^1 - \int_0^{\tau_L} e^{A_L(\tau_L - s)}\mathbf{b}\, ds + \mathbf{x}^0 &= \mathbf{0}, \end{aligned} \tag{3.2}$$

where τ_C and τ_L are the times spent by the semiorbit in each zone, and

$$\mathbf{x}^0 = \begin{bmatrix} 1 \\ x_2^0 \\ x_3^0 \end{bmatrix}, \quad \mathbf{x}^1 = \begin{bmatrix} -1 \\ x_2^1 \\ x_3^1 \end{bmatrix},$$

are two intersection points of the orbit with the planes Σ_1 and Σ_{-1}, respectively (from the symmetry, there will be two more, $\mathbf{x}^2 = -\mathbf{x}^0$ and $\mathbf{x}^3 = -\mathbf{x}^1$); see Figure 8. We will refer to the system formed by (3.2) as the *closing equations*. The use of these equations goes back to Andronov and coworkers [AVK66], and it was exploited by Kriegsmann [Kr87] in the context of limit cycle bifurcations. This author studied the *rapid* bifurcation in the Wien bridge oscillator, later revisited in [FPR99].

Starting from the critical value $T = T_c$ and considering the outermost periodic orbit of the corresponding center configuration, we want to use the closing equations to analyze what happens with such periodic orbit as T varies, keeping M and D constant and always assuming $M > 0$. To achieve this goal, it is more convenient to vary the eigenvalues of A_C in a neighborhood of $(\alpha, \beta, \delta) = (0, \sqrt{M}, D/M)$, adding to the closing equations (3.2) the last two equations of (3.1), to impose that M and D are fixed.

The sorted set formed by (3.2) and the last two equations of (3.1) will be denoted by

$$\mathbf{F}(\mathbf{z}) = \mathbf{0}, \tag{3.3}$$

where $\mathbf{z} = (\alpha, \beta, \delta, \tau_C, \tau_L, x_2^0, x_3^0, x_2^1, x_3^1)$, which constitutes a nonlinear system of eight equations and nine unknowns, to be studied in a neighborhood of the point

$$\bar{\mathbf{z}} = \left(0, \sqrt{M}, \frac{D}{M}, \frac{\pi}{\sqrt{M}}, 0, \frac{D}{M}, 0, -\frac{D}{M}, 0\right).$$

Obviously, we are interested in a branch of solutions of (3.3) passing through $\bar{\mathbf{z}}$, and leading to positive values of τ_L. It turns out that system (3.3) has a trivial branch of solutions that passes through $\bar{\mathbf{z}}$ and can be parameterized as

$$\mathbf{z}(\mu) = \left(0, \sqrt{M}, \frac{D}{M}, \frac{\pi}{\sqrt{M}}, 0, \frac{D}{M} + \mu, \mu\frac{D}{M}, -\frac{D}{M} - \mu, -\mu\frac{D}{M}\right) \tag{3.4}$$

for every real μ. This trivial branch will be called the *spurious branch* because, for $\mu \neq 0$, these solutions do not correspond to periodic orbits of the system (1.2). The Jacobian matrix of $\mathbf{F}$ in $\bar{\mathbf{z}}$ does not have full rank; in fact, as the following result shows, the point $\bar{\mathbf{z}}$ is a branch point where two branches intersect each other. Moreover, we obtain a new set of equations for which $\bar{\mathbf{z}}$ is nonsingular.

LEMMA 3.1. *For the closing equations* (3.3) *with* $M > 0$*, the following statements hold:*

(a) *The fourth equation of* (3.3)*, namely*

$$F_4(\mathbf{z}) = 0,$$

is satisfied for every $\mathbf{z}$ *with* $\tau_L = 0$.

(b) *The function* $\widetilde{F}_4(\mathbf{z})$ *such that* $F_4(\mathbf{z}) = \tau_L \widetilde{F}_4(\mathbf{z})$ *is an analytic function in a neighborhood of* $\bar{\mathbf{z}}$.

(c) *If we define the modified closing equations*

$$\mathbf{G}(\mathbf{z}) = \mathbf{0}, \tag{3.5}$$

where $G_4 = \widetilde{F}_4$ *and* $G_i = F_i$ *for* $i \neq 4$*, then the solution set of* (3.5) *in a neighborhood of* $\bar{z}$ *is the solution set of* (3.3) *excepting the spurious branch* (3.4).

(d) *For system* (3.5) *the point* $\bar{z}$ *is a nonsingular point. Moreover, the solutions of* (3.5) *are analytic functions of* τ_L *at* 0*, and their corresponding Taylor series are*

$$\alpha = \frac{M^{5/2}\gamma}{12\pi(D^2 + M^3)}\tau_L^3 + \frac{M^{1/2}\xi_1}{720\ pi(D^2 + M^3)}\tau_L^5 + O\left(\tau_L^6\right), \tag{3.6}$$

$$\beta = \sqrt{M} - \frac{DM\gamma}{12\pi(D^2 + M^3)}\tau_L^3 - \frac{D\xi_1}{720\pi M(D^2 + M^3)}\tau_L^5 + O\left(\tau_L^6\right), \tag{3.7}$$

$$\delta = \frac{D}{M} + \frac{D^2\gamma}{6\pi M^{1/2}(D^2 + M^3)}\tau_L^3 + \frac{D^2\xi_1}{360\pi M^{5/2}(D^2 + M^3)}\tau_L^5 + O\left(\tau_L^6\right), \tag{3.8}$$

$$\tau_C = \frac{\pi}{\sqrt{M}} - \tau_L + \frac{M - m}{12}\tau_L^3 + \frac{\xi_2}{720\pi M^2}\tau_L^5 + O\left(\tau_L^6\right), \tag{3.9}$$

$$x_2^0 = \frac{D}{M} + \frac{M}{2}\tau_L + \frac{Mt-D}{12}\tau_L^2 + \frac{4M\gamma + \pi\sqrt{M}\xi_3}{24\pi M^{3/2}}\tau_L^3 - \frac{\xi_4}{720M^2}\tau_L^4$$
$$+ \left[\frac{D\xi_3 - M^3 d}{24M^2\left(e^{\pi D/M^{3/2}}+1\right)} - \frac{DM^{3/2}\gamma}{12\pi(D^2+M^3)}\right]\tau_L^4 + O\left(\tau_L^5\right), \tag{3.10}$$

$$x_2^1 = -\frac{D}{M} + \frac{M}{2}\tau_L - \frac{Mt-D}{12}\tau_L^2 - \frac{4M\gamma - \pi\sqrt{M}\xi_3}{24\pi M^{3/2}}\tau_L^3 + \frac{\xi_4}{720M^2}\tau_L^4$$
$$- \left[\frac{e^{\pi D/M^{3/2}}(D\xi_3 - M^3 d)}{24M^2\left(e^{\pi D/M^{3/2}}+1\right)} + \frac{DM^{3/2}\gamma}{12\pi(D^2+M^3)}\right]\tau_L^4 + O\left(\tau_L^5\right), \tag{3.11}$$

$$x_3^0 = \frac{D}{2}\tau_L + \frac{(Mt-D)\,D}{12M}\tau_L^2 + \left[\frac{DM^{3/2}\gamma}{6\pi(D^2+M^3)} + \frac{D\xi_3}{24M^2}\right]\tau_L^3$$
$$- \frac{D\xi_3 - M^3 d}{12M^2\left(e^{\pi D/M^{3/2}}+1\right)}\tau_L^3 - \frac{D\xi_4}{720M^3}\tau_L^4 + O\left(\tau_L^5\right), \tag{3.12}$$

$$x_3^1 = \frac{D}{2}\tau_L - \frac{(Mt-D)\,D}{12M}\tau_L^2 - \left[\frac{DM^{3/2}\gamma}{6\pi(D^2+M^3)} + \frac{D\xi_3}{24M^2}\right]\tau_L^3$$
$$+ \frac{D\xi_3 + e^{\pi D/M^{3/2}}M^3 d}{12M^2\left(e^{\pi D/M^{3/2}}+1\right)}\tau_L^3 + \frac{D\xi_4}{720M^3}\tau_L^4 + O\left(\tau_L^5\right), \tag{3.13}$$

where

$$\begin{aligned}
\gamma &= DM - Dm + dM - tM^2,\\
\xi_1 &= 5D^3M - 5D^3m - 15D^2M^2t + 11D^2Md + 4D^2Mmt - 15DM^4\\
&\quad + 21DM^3m + 9DM^3t^2 - 10DM^2dt - 6DM^2m^2 + DM^2mt^2 + 15M^5t\\
&\quad - 9M^4d - 12M^4mt + M^4t^3 + 6M^3dm - M^3dt^2,\\
\xi_2 &= 5D^2M - 5D^2m - 10DM^2t + 6DMd + 4DMmt - 9M^4\\
&\quad + 15M^3m + 5M^3t^2 - 6M^2dt - 6M^2m^2 + M^2mt^2,\\
\xi_3 &= D\,(D - Mt) + M^2m,\\
\xi_4 &= 15D^3 - 20D^2Mt + 16DM^2m + 4DM^2t^2 - 9M^3d - 7M^3mt + M^3t^3.
\end{aligned} \tag{3.14}$$

Proof. Statements (a) and (b) come from a direct inspection of (3.3).

Recalling (3.2), from statement (b) and (3.3) we conclude that

$$\widetilde{F}_4(\mathbf{z})\Big|_{\tau_L=0} = \lim_{\tau_L\to 0}\frac{1}{\tau_L}F_4(\mathbf{z}) = \mathbf{e}_1^T A_L \begin{bmatrix} -1 \\ x_2^1 \\ x_3^1 \end{bmatrix} + \mathbf{e}_1^T \begin{bmatrix} t-2\alpha-\delta \\ m-M \\ d-D \end{bmatrix} = -x_2^1 - 2\alpha - \delta.$$

The above computation shows that $\widetilde{F}_4(\mathbf{z}(\mu)) = \mu$, so that the spurious branch (3.4) does not belong to the solution set of $\widetilde{F}_4(\mathbf{z}) = 0$. Besides, every solution $\mathbf{z}$ of (3.3) with $\tau_L \neq 0$ is a solution of (3.5), and statement (c) is proven.

For the computation of the Jacobian matrix $\partial\mathbf{G}/\partial\mathbf{z}\big|_{\mathbf{z}=\bar{\mathbf{z}}}$ and the series (3.6)–(3.13), we have used the following approach. For the first three rows of the closing equations, we work with the equivalent expression

$$\mu_0 e^{\delta\tau_C}\mathbf{v} + e^{\alpha\tau_C}\cos(\beta\tau_C)\,\widehat{\mathbf{v}} + e^{\alpha\tau_C}\frac{\sin(\beta\tau_C)}{\beta}\,(A_C - \alpha I)\,\widehat{\mathbf{v}} - \mathbf{x}^1 = \mathbf{0},$$

where $\mathbf{v} = \left[1, 2\alpha, \alpha^2 + \beta^2\right]^T$ is a right eigenvector of A_C associated with the real eigenvalue δ, and

$$\widehat{\mathbf{v}} = \mathbf{x}^0 - \mu_0 \mathbf{v}$$

is the projection (following the direction of $\mathbf{v}$) of the vector $\mathbf{x}^0$ onto the invariant plane associated with the complex eigenvalues of A_C. Consequently, the coefficient μ_0 is

$$\mu_0 = \frac{\mathbf{w}^T \mathbf{x}^0}{\mathbf{w}^T \mathbf{v}} = \frac{\delta^2 - \delta x_2^0 + x_3^0}{(\delta - \alpha)^2 + \beta^2},$$

where $\mathbf{w}^T = \left[\delta^2, -\delta, 1\right]$ is a left eigenvector of A_C associated with the eigenvalue δ.

Regarding the next three rows of the closing equations, it is useful to write the matrix exponentials in series of τ_L. Then, in computing partial derivatives with respect to the variables other from τ_L, one only needs to consider the terms of degree zero in τ_L. This comment is also useful for obtaining $\widetilde{F}_4$ from F_4.

Thus, the Jacobian matrix $\partial \mathbf{G}/\partial \mathbf{z}\big|_{\mathbf{z}=\overline{\mathbf{z}}}$ is

$$\begin{bmatrix}
-\frac{\pi}{\sqrt{M}} & 0 & DMK & 0 & 0 & -DMK & M^2K & 0 & 0 \\
-\frac{\pi D}{M^{3/2}} & -\pi & 0 & -M & 0 & -1 & 0 & -1 & 0 \\
0 & -\frac{\pi D}{M} & DM^2K & -D & 0 & -DM^2K & M^3K - 1 & 0 & -1 \\
-2 & 0 & -1 & 0 & \frac{M}{2} & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & -M & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & -D & 0 & 1 & 0 & 1 \\
\frac{2D}{M} & 2\sqrt{M} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{2D}{\sqrt{M}} & M & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix},$$

where

$$K = \frac{e^{\pi D/M^{3/2}} + 1}{D^2 + M^3}. \tag{3.15}$$

If we remove the fifth column (corresponding to τ_L), the determinant of the resulting matrix is equal to

$$-2\pi M^2 \left(e^{\pi D/M^{3/2}} + 1\right) \neq 0,$$

and hence the matrix has full rank. From the implicit function theorem for analytic functions (see [CH82]) we obtain statement (d). All the computations of the above series expansions have been checked with the symbolic manipulator Maple; see [MGHLVM03]. □

In what follows, we give an auxiliary result to analyze the stability of the bifurcating limit cycle. First, we must study the behavior of the return map near a periodic orbit of three zones. Due to the symmetry, we need to use only the semiorbit that starts from $\mathbf{x}^0 \in \Sigma_1$, crosses Σ_{-1} at the point $\mathbf{x}^1$, and returns to this section at the point $\mathbf{x}^2 = -\mathbf{x}^0 \in \Sigma_{-1}$. We denote by $\mathbf{p}_0$, $\mathbf{p}_1 \in \mathbb{R}^2$, the coordinates of $\mathbf{x}^0$ and $\mathbf{x}^1$ restricted to their respective sections. From the transition maps associated with the flow, locally defined at the points $\mathbf{x}^0$ and $\mathbf{x}^1$, it is possible in adequate neighborhoods at the sections to define the functions providing the corresponding restricted coordinates. Let us denote by π_C, π_L such functions, satisfying $\pi_C(\mathbf{p}_0)$=$\mathbf{p}_1$, $\pi_L(\mathbf{p}_1) = -\mathbf{p}_0$,

and let $\pi_{LC} = \pi_L \circ \pi_C$. We will let $\tau_C(\mathbf{p}_0)$ and $\tau_L(\mathbf{p}_1)$ denote the times spent by the semiorbit in passing from $\mathbf{x}^0$ to $\mathbf{x}^1$ and from $\mathbf{x}^1$ to $\mathbf{x}^2$, respectively, and write $D_{\mathbf{p}}(\cdot)$ to indicate the derivative with respect to the restricted coordinates. The next result shows how to compute the derivative $D_{\mathbf{p}}\pi_{LC}$, intimately related to the monodromy matrix associated with the periodic orbit.

PROPOSITION 3.2. *Consider a symmetrical periodic orbit of system* (1.2) *that uses the three zones, starting from* $\mathbf{x}^0 \in \Sigma_1$ *with coordinates* $\mathbf{p}_0 \in \mathbb{R}^2$*, passing through* $\mathbf{x}^1 \in \Sigma_{-1}$ *with coordinates* $\mathbf{p}_1 \in \mathbb{R}^2$*, and transversal to both sections. Then, the product of the two matrices*

$$\begin{bmatrix} 1 & D_{\mathbf{p}}\tau_L(\mathbf{p}_1) \\ \mathbf{0} & D_{\mathbf{p}}\pi_L(\mathbf{p}_1) \end{bmatrix} \begin{bmatrix} -1 & D_{\mathbf{p}}\tau_C(\mathbf{p}_0) \\ \mathbf{0} & D_{\mathbf{p}}\pi_C(\mathbf{p}_0) \end{bmatrix} = \begin{bmatrix} -1 & D_{\mathbf{p}}\tau_C(\mathbf{p}_0) + D_{\mathbf{p}}\tau_L(\mathbf{p}_1)D_{\mathbf{p}}\pi_C(\mathbf{p}_0) \\ \mathbf{0} & D_{\mathbf{p}}\pi_{LC}(\mathbf{p}_0) \end{bmatrix}$$

is similar to

$$e^{A_L\tau_L(\mathbf{p}_1)}e^{A_C\tau_C(\mathbf{p}_0)}. \tag{3.16}$$

Proof. It is enough to use the explicit expressions of the solutions of system (1.2) at every zone and the continuity of the vector field; see [Ro03] for more details. ∎

The following lemma deals with a technical result that allows us to invert certain power series; see [FPR99] for a proof.

LEMMA 3.3. *Let be* $\eta = \xi^n\rho(\xi)$ *with* n *odd, where* ρ *is a real function analytic at* 0 *and such that* $\rho(0) = \rho_0 \neq 0$*. Then there exists a real function* χ *analytic at* 0*, with* $\chi(0) \neq 0$ *and such that* $\xi = \eta^{\frac{1}{n}}\chi(\eta^{\frac{1}{n}})$.

If we select only the solutions of the closing equations with $\tau_L > 0$ but sufficiently small, and $0 < \tau_C < \pi/\sqrt{M}$ but sufficiently close to $\pi/\sqrt{M}$, then we can assure that such solutions correspond to symmetrical and transversal periodic orbits; see [Ro03] for more details. Reciprocally, if we take a symmetrical periodic orbit that uses the three zones and is sufficiently close to the outermost periodic orbit of the center that exists for the critical values of parameters, then its corresponding values $\tau_C > 0$, $\tau_L > 0$, $\mathbf{x}^0$, $\mathbf{x}^1$, and remaining parameters determine a point $\mathbf{z}$ satisfying the closing equations. Therefore, we can establish with the above restrictions a correspondence between solutions $\mathbf{z}$ of closing equations and symmetrical periodic orbits. This correspondence, along with the uniqueness of the solution obtained in Lemma 3.1, ensures that the corresponding bifurcating periodic is an isolated periodic orbit, that is, a limit cycle.

Coming back to the statements of Theorem 1.1, we begin by using statement (d) of Lemma 3.1. We can compute $T(\tau_L)$ using that $T = 2\alpha + \delta$ and the corresponding expansions (3.6) and (3.8) for α and δ, obtaining

$$T = \frac{D}{M} + \frac{\gamma}{6\pi M^{1/2}}\tau_L^3 + \frac{\xi_1}{360\pi M^{5/2}}\tau_L^5 + O\left(\tau_L^6\right), \tag{3.17}$$

where γ and ξ_1 are given in the statement of Lemma 3.1. From (3.17) and taking into consideration that τ_L must be positive, it is obvious that $MT - D$ and γ have the same sign, and so the condition $\gamma(MT - D) > 0$ holds.

Now, if we apply Lemma 3.3 to (3.17), taking $n = 3$, $\eta = MT - D$, and $\xi = \tau_L$, we conclude that τ_L is an analytic function at the origin in the variable $(MT - D)^{1/3}$. A standard computation leads to the expansion

$$\tau_L = \frac{(6\pi)^{1/3}(MT - D)^{1/3}}{M^{1/6}\gamma^{1/3}} + \frac{\pi\xi_1}{30M^{5/2}\gamma^2}(MT - D) + O(MT - D)^{4/3}. \tag{3.18}$$

Due to the symmetry of the orbit, its period is equal to $2(\tau_C + \tau_L)$. Substituting expansion (3.18) into (3.9), and computing the above expression for the period, we get the expansion given for P_{er}.

We will now determine the amplitude of the periodic orbit. By using the variation of parameters formula, the solution of system (1.2) in zone R is

$$\mathbf{x}(\tau) = e^{A_L \tau} \mathbf{x}^3(\tau_L) + \int_0^\tau e^{A_L(\tau - s)} \mathbf{b}(\tau_L)\, ds, \tag{3.19}$$

so that its first component is

$$x_1(\tau) = \mathbf{e}_1^T \left\{ e^{A_L \tau} \begin{bmatrix} 1 \\ -x_2^1(\tau_L) \\ -x_3^1(\tau_L) \end{bmatrix} + \left(\sum_{i=0}^{\infty} A_L^i \frac{\tau^{i+1}}{(i+1)!} \right) \mathbf{b}(\tau_L) \right\}. \tag{3.20}$$

Let τ^* be the time when $|x_1|$ attains its maximum value in zone R. Taking derivatives with respect to τ in (3.20), and imposing that it must vanish at τ^*, we get

$$G(\tau_L, \tau^*) = \left. \frac{dx_1(\tau)}{d\tau} \right|_{\tau = \tau^*} = \mathbf{e}_1^T e^{A_L \tau^*} \begin{bmatrix} x_2^1(\tau_L) + T(\tau_L) \\ x_3^1(\tau_L) + M \\ D \end{bmatrix} = 0. \tag{3.21}$$

Now using expressions (3.11) and (3.13) and computing the power series of G in (τ_L, τ^*) at $(0,0)$, we obtain

$$G(\tau_L, \tau^*) = \frac{M}{2}\tau_L - M\tau^* + \frac{D - Mt}{12}\tau_L^2 + \frac{Mt - D}{2}\tau_L \tau^* + \frac{D - Mt}{2}\tau^{*2} + O(\tau_L, \tau^*)^3.$$

Hence, (3.21) defines implicitly in a neighborhood of $(0,0)$ a function $\tau^* = \psi(\tau_L)$ such that $G(\tau_L, \psi(\tau_L)) = 0$, namely,

$$\tau^* = \frac{1}{2}\tau_L + \frac{Mt - D}{24M}\tau_L^2 + O(\tau_L^4).$$

Substituting the above expansion together with (3.11), (3.13), and (3.17) into the expression (3.20), we get

$$a = x_1(\tau^*) = 1 + \frac{M}{8}\tau_L^2 + \frac{1}{1152M}(13D^2 - 11DMt + 15M^2 m - 2M^2 t^2)\tau_L^4 + O(\tau_L^5).$$

Using expression (3.18) for τ_L, we obtain the final expression for the amplitude a.

Let us now compute the characteristic multipliers of the bifurcating limit cycle. Due to the similarity relationship established in Proposition 3.2, we conclude that the product $\exp(A_L \tau_L) \cdot \exp(A_C \tau_C)$ corresponding to a solution of (3.5) has an eigenvalue equal to -1. We will denote by λ_r and λ_a the other two eigenvalues that correspond to the eigenvalues of the derivative $D_{\mathbf{p}} \pi_{LC}(\mathbf{p}_0)$ of the transition map associated with the semiorbit. The product of the three eigenvalues is then equal to

$$-\lambda_r \lambda_a = \det\left(e^{A_L \tau_L}\right) \det\left(e^{A_C \tau_C}\right).$$

Using that $\det(e^{A\tau}) = \exp(\tau \operatorname{trace}(A))$, we get

$$-\lambda_r \lambda_a = e^{\tau_L t + \tau_C T}. \tag{3.22}$$

The expansion of the product of exponentials in (3.16) leads to an expression of the form

$$e^{A_L\tau_L}e^{A_C(\tau_L)\tau_C(\tau_L)} = H_0 + \tau_L H_1 + \tau_L^2 H_2 + \cdots . \tag{3.23}$$

To compute the above matrices H_i, we write

$$\left(I + A_L\tau_L + A_L^2\frac{\tau_L^2}{2!} + \cdots\right) \times \left(e^{A_C(0)\tau_C(0)} + \tau_L \left.\frac{d}{d\tau_L}e^{A_C(\tau_L)\tau_C(\tau_L)}\right|_{\tau_L=0} + \frac{\tau_L^2}{2!}\left.\frac{d^2}{d\tau_L^2}e^{A_C(\tau_L)\tau_C(\tau_L)}\right|_{\tau_L=0} + \cdots\right).$$

From expansions (3.6)–(3.13), we obtain $\tau_C(0) = \pi/M^{1/2}$, $\tau_C'(0) = -1$, $\tau_C''(0) = 0$, $A_C'(0) = A_C''(0) = \mathbf{0}$, and using these values in the above expression, we finally get

$$H_0 = e^{A_C(0)\tau_C(0)} = \begin{bmatrix} D^2K - 1 & -DMK & M^2K \\ 0 & -1 & 0 \\ D^2MK & -DM^2K & M^3K - 1 \end{bmatrix},$$

$$H_1 = \begin{bmatrix} t - D/M \\ m - M \\ d - D \end{bmatrix} \begin{bmatrix} D^2K - 1 & -DMK & M^2K \end{bmatrix}$$

and

$$H_2 = \frac{Mt - D}{2M}H_1,$$

where K has been defined in (3.15), and it is emphasized that H_1 and H_2 are rank-one matrices.

The matrix H_0 has eigenvalues -1 (double) and $\lambda_0 = \exp(\pi D/M^{3/2})$. For the single eigenvalue λ_0, we select a right eigenvector $\mathbf{v}_0 = [1, 0, M]^T$ and a left eigenvector $\mathbf{w}_0^T = [D^2/M^2, -D/M, 1]$. We will denote by λ_a the eigenvalue of $D\pi_{LC}(\mathbf{p}_0)$ that for $\tau_L = 0$ is equal to λ_0. Since the eigenvalue λ_0 of H_0 is simple, we can apply perturbation theory (see section 2.8 of [Wi65]) to assure that the equality

$$\left(H_0 + \tau_L H_1 + \tau_L^2 H_2 + \cdots\right)\left(\mathbf{v}_0 + \tau_L\mathbf{v}_1 + \tau_L^2\mathbf{v}_2 + \cdots\right)$$

$$= \left(\lambda_0 + \tau_L\lambda_1 + \tau_L^2\lambda_2 + \cdots\right)\left(\mathbf{v}_0 + \tau_L\mathbf{v}_1 + \tau_L^2\mathbf{v}_2 + \cdots\right)$$

holds for certain vectors $\mathbf{v}_1$, $\mathbf{v}_2 \ldots$. As a consequence of Proposition 3.2 and (3.23), we get

$$\lambda_a = \lambda_0 + \tau_L\lambda_1 + \tau_L^2\lambda_2 + \cdots .$$

After some computations, we arrive at

$$\lambda_1 = \frac{\mathbf{w}_0^T H_1\mathbf{v}_0}{\mathbf{w}_0^T\mathbf{v}_0} = \left(\frac{Mt - D}{M} + \frac{\gamma M}{D^2 + M^3}\right)\lambda_0,$$

$$\lambda_2 = \frac{\mathbf{w}_0^T\left(H_2\mathbf{v}_0 + (H_1 - \lambda_1 I)\,\mathbf{v}_1\right)}{\mathbf{w}_0^T\mathbf{v}_0} = \frac{(Mt - D)\left(D^2 + M^3\right) + \gamma}{2M^2\left(D^2 + M^3\right)^2(\lambda_0 + 1)} \times \left[(Mt - D)\left[\left(D^2 + M^3\right)\lambda_0 + D^2 - M^3\right] + 2M^2(dM - Dm)\right]\lambda_0.$$

The logarithms μ_r and μ_a of characteristic multipliers of the complete periodic orbit must satisfy

$$e^{\mu_r} = \lambda_r^2, \qquad e^{\mu_a} = \lambda_a^2, \tag{3.24}$$

while from (3.22) we get the relationship

$$\mu_r + \mu_a = 2t\tau_L + 2T\tau_C. \tag{3.25}$$

From (3.24) and using the computed simple eigenvalue λ_a, we obtain

$$\mu_a = 2\log\left[\lambda_0 + \lambda_1\tau_L + \lambda_2\tau_L^2 + O(\tau_L^3)\right] = 2\log\lambda_0 + 2\log\left[1 + \frac{\lambda_1}{\lambda_0}\tau_L + \frac{\lambda_2}{\lambda_0}\tau_L^2 + O(\tau_L^3)\right]$$
$$= 2\lambda_0 + 2\frac{\lambda_1}{\lambda_0}\tau_L + \left[2\frac{\lambda_2}{\lambda_0} - \frac{\lambda_1^2}{\lambda_0^2}\right]\tau_L^2 + O(\tau_L^3).$$

Substituting here λ_1 and λ_2, and using expansion (3.18) of τ_L, we finally get the expression for μ_a that appears in Theorem 1.1. Using in (3.25) the expansions (3.9) for τ_C and (3.18) for τ_L, we compute μ_r.

Since the last assertion of Theorem 1.1 is a direct consequence of previous statements, its proof is now completed.

Acknowledgments. The authors sincerely appreciate the careful reading of the anonymous referees and their interesting suggestions that have notably improved the paper. They also want to acknowledge Jorge Galán for his invaluable suggestions and comments on a preliminary version of the manuscript, Manuel Román for his help with SPICE simulation of the circuit, and Fernando Fernández for helping with figures.

REFERENCES

[AVK66] A. A. Andronov, A. A. Vitt, and S. E. Khaikin, *Theory of Oscillators*, Dover, New York, 1966.

[CFPT02] V. Carmona, E. Freire, E. Ponce, and F. Torres, *On simplifying and classifying piecewise-linear systems*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 49 (2002), pp. 609–620.

[CH82] S. N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, New York, Berlin, 1982.

[CWHZ93] L. Chua, C. Wu, A. Huang, and G. Zhong, *A universal circuit for studying and generating chaos—Part* I: *Routes to chaos*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 40 (1993), pp. 732–744.

[FGA84] E. Freire, L. García-Franquelo, and J. Aracil, *Periodicity and chaos in an autonomous electronic oscillator*, IEEE Trans. Circuits Systems, 31 (1984), pp. 237–247.

[FRGP93] E. Freire, A. J. Rodríguez-Luis, E. Gamero, and E. Ponce, *A case of study for homoclinic chaos in an autonomous electronic circuit: A trip from Takens–Bogdanov to Hopf–Šil'nikov*, Phys. D, 62 (1993), pp. 230–253.

[FPR99] E. Freire, E. Ponce, and J. Ros, *Limit cycle bifurcation from center in symmetric piecewise-linear systems*, Internat. J. Bifur. Chaos, 9 (1999), pp. 895–907.

[GK92] M. G. M. Gomes and G. P. King, *Bistable chaos.* II. *Bifurcation analysis*, Phys. Rev. A, 46 (1992), pp. 3100–3110.

[HBCJM91] J. J. Healey, D. S. Broomhead, K. A. Cliffe, R. Jones, and T. Mullin, *The origins of chaos in a modified Van der Pol oscillator*, Phys. D, 48 (1991), pp. 322–339.

[Ke93] M. P. Kennedy, *Three steps to chaos—Part* II: *A Chua's circuit primer*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 40 (1993), pp. 657–674.

[Kr87] G. A. Kriegsmann, *The rapid bifurcation of the Wien bridge oscillator*, IEEE Trans. Circuits Systems, 34 (1987), pp. 1093–1096.

[Ma93] R. Madan, *Chua's Circuit: Paradigm for Chaos*, World Scientific, Singapore, 1993.

[MGHLVM03] M. B. Monagan, K. O. Geddes, K. M. Heal, G. Labahn, S. M. Vorkoetter, J. McCarron, and P. DeMarco, *Maple* 9 *Introductory Programming Guide*, Maplesoft, Waterloo, ON, 2003.

[QNPS93] T. Quarles, A. R. Newton, D. O. Pederson, and A. Sangiovanni-Vincentelli, *Spice*3 *Version* 3f3 *User's Manual*, Department of Electrical Engineering and Computer Sciences, University of California Berkeley, Berkeley, CA, 1993.

[Ro03] J. Ros, *Estudio del Comportamiento Dinámico de Sistemas Autónomos Tridimensionales Lineales a Trozos*, Ph.D. dissertation, Universidad de Sevilla, Seville, Spain, 2003 (in Spanish).

[SYM81] R. Shinriki, M. Yamamoto, and S. Mori, *Multimode oscillations in a modified Van der Pol oscillator containing a positive nonlinear conductance*, IEEE Proc., 69 (1981), pp. 394–395.

[Wi65] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, UK, 1965.

SIAM J. APPL. MATH.
Vol. 65, No. 6, pp. 1952–1980

ON A MATHEMATICAL MODEL OF THE PRODUCTIVITY INDEX OF A WELL FROM RESERVOIR ENGINEERING*

AKIF IBRAGIMOV†, DINARA KHALMANOVA‡, PETER P. VALKO§, AND JAY R. WALTON‡

Abstract. Motivated by the reservoir engineering concept of the productivity index of a producing oil well in an isolated reservoir, we analyze a time dependent functional, diffusive capacity, on the solutions to initial boundary value problems for a parabolic equation. Sufficient conditions providing for time independent diffusive capacity are given for different boundary conditions. The dependence of the constant diffusive capacity on the type of the boundary condition (Dirichlet, Neumann, or third boundary condition) is investigated using a known variational principle and confirmed numerically for various geometrical settings. An important comparison between two principal constant values of a diffusive capacity is made, leading to the establishment of criteria when the so-called pseudo-steady-state and boundary-dominated productivity indices of a well significantly differ from each other. The third boundary condition is shown to model the thin skin effect for the constant wellbore pressure production regime for a damaged well. The questions of stabilization and uniqueness of the time independent values of the diffusive capacity are addressed. The derived formulas are used in numerical study of evaluating the productivity index of a well in a general three-dimensional reservoir for a variety of well configurations.

Key words. parabolic equation, productivity index, diffusive capacity, time-invariant, skin

AMS subject classification. 76S05

DOI. 10.1137/040607654

1. Introduction.

In many applied problems, where the modeled processes are, in general, transient, it is important to define such functionals on the solutions, which are, in a sense, time invariant. Existence of such property is important from both practical and theoretical points of view. An important such example to petroleum reservoir engineering, the productivity index (PI), is studied here.

It was long ago observed by petroleum engineers that if a bounded reservoir is depleted by a well, then the ratio of the flow rate to the pressure drawdown (the pressure drop between the reservoir and the wellbore) stabilizes to a constant value. This constant value seems to depend only on the geometrical and hydrodynamical characteristics of the reservoir. In particular, it appears to be independent of the pressure drawdown in the reservoir or the flow rate from the well [23].

The first concise description of this fact was formulated in the classical book by Muskat [23]. The ratio of the rate of flow from the well to the difference between the average pressure on the wellbore and the average pressure in the reservoir is called the productivity index of the well [23]. There are two idealized production regimes considered most frequently for the purpose of analysis in engineering practice: the well can be produced either with a constant flow rate or with a constant wellbore

*Received by the editors April 30, 2004; accepted for publication (in revised form) October 19, 2004; published electronically August 3, 2005.
http://www.siam.org/journals/siap/65-6/60765.html
†Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX 79409-1042 (akif.ibraguimov@ttu.edu).
‡Department of Mathematics, Texas A&M University, MS 3368, College Station, TX 77840 (dkhalmanova@engr.psu.edu, jwalton@math.tamu.edu).
§Department of Petroleum Engineering, Texas A&M University, College Station, TX 77840 (p-valko@tamu.edu).

pressure. In a bounded reservoir depleted in either of the two regimes, the PI of a well stabilizes and remains constant in a long time asymptote.

To analyze the productivity of the well we consider three initial boundary value problems (IBVPs) that correspond to current engineering practice. However, while two of the formulated problems corresponding to the constant pressure production regime are well-posed, the problem modeling the regime with a constant rate of production is ill-posed in the sense of nonuniqueness of solution.

Field operations often reduce the permeability of the region adjacent to the wellbore—the so-called skin zone. Disregarding the skin effect leads to overestimation of the PI of the damaged well [30, 15]. One of the IBVPs considered in this article models the skin effect in the constant pressure production regime.

The objective for this paper is to build a rigorous mathematical frame for studying the PI. In this respect, it proves useful to introduce the concept of *diffusive capacity* for a well-reservoir system. The diffusive capacity is an integral type characteristic of the solution of an IVBP. To address the issue of nonuniqueness of solution of the ill-posed IBVP, we impose restrictions defining a class of solutions in which the diffusive capacity is unique. The inflicted restrictions are motivated by physical considerations as well as traditional engineering practice.

An important property of the PI to stabilize with time regardless of the production regime is then analyzed in terms of the diffusive capacity. Sufficient conditions for the diffusive capacity to be time independent are given for different boundary conditions; through a variational approach to studying the diffusive capacities, its dependence on different boundary conditions is revealed. The obtained theoretical results are then illustrated by numerical computations of the constant diffusive capacities for processes with different boundary conditions in various geometrical settings.

1.1. PI of a well in a bounded reservoir. Reservoir engineering approach: Shape factors. Consider a bounded hydrocarbon reservoir with a flowing fluid (oil) and a well produced with either constant wellbore pressure or constant production rate. The PI of a well is defined as [26]

$$PI(t) = \frac{q(t)}{p_w(t) - p_a(t)}, \tag{1}$$

where $q(t)$ is the rate of flow from the well, $p_w(t)$ is the flowing bottomhole pressure, and $p_a(t)$ is the average pressure of the fluid in the reservoir. When the well is produced with a constant wellbore pressure, its value is taken as $p_w(t)$ in (1). The concept of the PI of a well facilitates reservoir engineering methods of estimation of the available reserves and, consequently, helps to optimize the recovery efficiency.

About a century ago it was empirically observed that under either of the two recovery regimes, the PI of a well stabilizes and remains almost constant in a long time asymptote [26]. When the PI of a well is constant, the production regimes have traditionally accepted names: the production regime with the constant rate and constant PI is called a pseudo-steady-state (PSS), and the production regime with the constant wellbore pressure and the constant PI is called a boundary-dominated (BD) state.

The first analytical formula for representation of the PI of a well for a PSS regime was obtained by Muskat [23] for an isolated cylindrical reservoir and a given constant production rate on the fully penetrated vertical well. The IBVPs with the constant rate well boundary condition for a number of typical drainage shapes were first solved

by Matthews, Brons, and Hazebroek in [21] in connection to the analysis of the build-up wellbore pressure after well shut-in. Using the result of Matthews, Brons, and Hazebroek, an approximate formula for a PSS PI (with skin s) can be written as

$$J_{Dietz} = \frac{1}{\frac{1}{2}\ln\frac{4V}{\gamma C_A r_w^2} + s}, \tag{2}$$

where V is the area of the two-dimensional reservoir (a three-dimensional reservoir with a uniform thickness), r_w is the radius of the circular well, and γ is Euler's constant. Equation (2) uses the solution for the dimensionless PSS wellbore pressure first derived by Ramey and Cobb in [27]. The values of the so-called shape factor C_A were first presented in [6] and are usually referred to as Dietz's shape factors in the petroleum engineering literature. Positive skin captures the damage to the skin zone, while the negative skin was shown to model a stimulated well [12, 15, 4, 8, 14, 19].

The approximate formula (2) is also used to estimate the productivity of a well produced with a constant bottomhole pressure. However, it is known that the BD state PI of a well is, in general, different from the PSS PI. In particular, the empirical evidence is that the PSS PI is always greater than or equal to the BD PI.

In 1998 Wattenbarger and Helmy derived an algorithm and computed the values of shape factors in (2) for the typical shapes of the drainage area for BD state, using a method of images, Laplace transform, and a fundamental relationship between the images in Laplace space of the cumulative production and the production rate. The applicability of (2) is contingent on the method of images—a drainage area to which the method of images can be applied must be of a shape, which, when translated infinitely many times in all directions, can cover the entire two-dimensional plane.

Most solutions for evaluating the PI in three-dimensional reservoirs, i.e., for directionally drilled wells, follow the same principle as the two-dimensional methods in that they are based on a semianalytical solution for a particular case, from which one finds a convenient approximate formula which is then applied to similar reservoir/well configurations. The semianalytical solution is often based on the superposition of analytical solutions for a transient problem in an unbounded reservoir. For the solution of the problem to be unique, additional assumptions must be made. Usually the restrictions are imposed on the distribution of the pressure on the wellbore. Under one such restriction, the wellbore is assumed to have infinite conductivity, i.e., the wellbore pressure is assumed to be constant on the wellbore at each moment of time. Under another restriction the pressure flux through the wellbore surface is constant at all times.

The solution in a bounded reservoir is then expressed in terms of an infinite time dependent series, similar to the technique used in [21, 16]. Then a comprehensive computing procedure is applied to determine the stabilized values of the time dependent series in the obtained solution [20, 25, 29, 17, 3, 28].

In most cases the methods for computing the PI of a deviated or horizontal well in a three-dimensional reservoir are aimed at obtaining an appropriate value of a shape factor C_A and skin factor s in (2). The effects associated with the deviation of the well from a fully penetrated vertical one are included in the skin s. A vertical well is called fully penetrated if its penetration length is equal to the thickness of the reservoir. A vertical fully penetrated well corresponds to $s = 0$. The effects of the geometry of the external boundaries of the reservoir are included in the shape factor C_A [20, 7].

As seen from this brief review, the existing methods and techniques of evaluation of the PI impose serious restrictions on the geometry of the reservoir. In particular,

the vertical dimension of the reservoir has to be small in comparison to its lateral dimensions to allow one to neglect the flow in the vertical direction or include its effect in the geometrical skin, s_g. Another restriction is due to the use of the method of images, which requires the drainage area shape to be convex and suitable for covering the whole plane when translated infinitely many times.

One should also note that very little attention has been paid to methods for evaluating a BD PI. For instance, all works mentioned above are concerned only with evaluating the PSS PI in three-dimensional reservoirs. In practice, the BD PI values are taken to be equal to the PSS PI, although it has been shown that the difference between these two values of PI can be up to 10% even for horizontal flow in simple drainage shapes [13, 16].

2. Statement of the problem. Let a point in $\mathbb{R}^n$ be denoted by $x = (x_1, \dots, x_n)$, $n = 2, 3$. Let Ω be an open domain in $\mathbb{R}^n$ which is bounded by the two disjoint piecewise smooth surfaces Γ_w and Γ_e. Let $u(x,t)$, $t \in \mathbb{R}$, be a solution of the equation

$$\frac{\partial u}{\partial t} = Lu, \tag{3}$$

where $L = \nabla \cdot (A(x)\nabla)$, A is a symmetric positive definite matrix with smooth components and $\nabla = (\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n})$ is the usual gradient operator.

Let $u(x,t)$ be subject to the homogeneous Neumann boundary condition on Γ_e:

$$\frac{\partial u}{\partial \vec{\nu}} = (A(x)\nabla u) \cdot \vec{n} = 0, \tag{4}$$

where $\vec{n}$ is the outward normal to Γ_e. On the remaining part of the boundary, Γ_w, three types of boundary conditions will be considered:

(a) constant total flux $\int_{\Gamma_w} \frac{\partial u}{\partial \vec{\nu}} dS = -q$, q being a real positive constant;
(b) constant Dirichlet condition $u|_{\Gamma_w} = u_{w2}$, u_{w2} being a real positive constant;
(c) mixed boundary condition $\left((u - u_{w3})|_{\Gamma_w} + \alpha \frac{\partial u}{\partial \vec{\nu}}\right)|_{\Gamma_w} = 0$, where α and u_{w3} are real constants, $u_{w3} > 0$.

For simplicity, we assume that the components of the coefficient matrix A and the domain boundary are smooth, so solutions of the IBVPs I, II, and III (stated below) are understood in a classical sense. In (b), $u_{w2} > 0$ is a given constant; in (c), $u_{w3} > 0$ and α are given constants.

This leads to three IBVPs:

Problem I.

$$Lu = \frac{\partial u}{\partial t}, \quad x \in \Omega, \quad t > 0,$$
$$\frac{\partial u}{\partial \vec{\nu}}|_{\Gamma_e} = 0,$$
$$\int_{\Gamma_w} \frac{\partial u}{\partial \vec{\nu}} dS = -q,$$
$$u(x,0) = f_1(x).$$

Remark 1. As mentioned in the introduction, Problem I is ill-posed: there are infinitely many solutions. The PI will be modeled as an integral characteristic of a solution and hence will be lacking uniqueness of definition. Therefore, we will

consider two classes of solutions in each of which the solution is unique up to an additive constant. These two classes will be described in detail in sections 3 and 4. Each class has a clear physical meaning. The integral characteristic modeling the PI will be shown to be unique in each class.

Problem II.

$$\begin{aligned} Lu = \frac{\partial u}{\partial t}, \quad x \in \Omega, \quad t > 0, \\ \frac{\partial u}{\partial \vec{\nu}}|_{\Gamma_e} = 0, \\ u|_{\Gamma_w} = u_{w2}, \\ u(x,0) = f_2(x). \end{aligned}$$

Problem III.

$$\begin{aligned} Lu = \frac{\partial u}{\partial t}, \quad x \in \Omega, \quad t > 0, \\ \frac{\partial u}{\partial \vec{\nu}}|_{\Gamma_e} = 0, \\ \left(\alpha \frac{\partial u}{\partial \vec{\nu}} + (u - u_{w3})\right)|_{\Gamma_w} = 0, \\ u(x,0) = f_3(x). \end{aligned}$$

Remark 2. Physically, $u(x,t)$ is interpreted as the fluid pressure in the reservoir, and hence, we will restrict our attention only to positive solutions of Problems I, II, and III. Moreover, a solution to Problem I is not necessarily positive on Ω for all $t > 0$, even if the initial function $f_1(x)$ is positive on Ω. It will be shown that for positive q, there exists a solution to Problem I which is positive on Ω for $t \in (0, T)$ for some positive T.

Remark 3. The maximum principle for a parabolic equation implies that the solution of Problem II is unique and positive if the initial condition f_2 is positive on Ω [10]. The uniqueness, existence, and regularity of the solutions of Problem III with respect to the sign of the coefficient α in the boundary condition on Γ_w are discussed, for example, in [10]. Formally, Problem III is a generalization of Problem II. However, we consider Problem II separately in light of its importance for applications in the reservoir engineering.

Remark 4. The obtained results can be extended to a generalized Wiener solution of an IBVP in a locally smooth domain [18]. We will not present it in this work to preserve the original engineering statement of the problem.

2.1. Definition of diffusive capacity. Let us introduce the following notation. If v is a function defined on Ω, then let $\bar{v}_w$ and $\bar{v}_\Omega$ denote the average of v on Γ_w and Ω, respectively, defined by

$$\bar{v}_w = \frac{1}{W} \int_{\Gamma_w} v dS$$

and

$$\bar{v}_\Omega = \frac{1}{V} \int_{\Omega} v dx,$$

where $V = mes_n \Omega$, $W = mes_{n-1} \Gamma_w$.

DEFINITION 1. *Let $u(x,t)$ be a classical solution* [10] *of the parabolic equation $Lu = \frac{\partial u}{\partial t}$ in $\Omega \times (0,\infty)$ with boundary condition $\frac{\partial u}{\partial \vec{\nu}}|_{\Gamma_e} = 0$ and* (a), (b), *or* (c) *on Γ_w. Let $T > 0$ be such that $u(x,t) > 0$ for all $x \in \Omega$ and $t \in (0,T)$. The diffusive capacity of Γ_w with respect to Γ_e (or simply diffusive capacity) corresponding to the solution $u(x,t)$ is the ratio*

$$J(u,t) = \frac{\int_{\Gamma_w} \frac{\partial u}{\partial \vec{\nu}} dS}{\bar{u}_w - \bar{u}_\Omega}, \tag{5}$$

where $t \in (0,T)$.

Remark 5. For fixed boundary and initial conditions in Problem II (III), the diffusive capacity $J(u,t)$ corresponding to the solution u of Problem II (III) is a function of time only. However, for fixed boundary and initial conditions in Problem I, the diffusive capacity $J(u,t)$ is a time dependent functional on the set of solutions $\{u\}$ to Problem I.

Remark 6. The corresponding diffusive capacity corresponding to a solution of Problem III is defined as

$$J(u,t) = \frac{\int_{\Gamma_w} \frac{\partial u}{\partial \vec{\nu}} dS}{u_{w3} - \bar{u}_\Omega}. \tag{6}$$

Such correction to the general definition is based on the physical assumption that u_{w3} is an average wellbore pressure, measured inside the wellbore.

In our intended application, Ω represents a hydrocarbon reservoir with a flowing fluid (oil) with the outer boundary Γ_e and a well with boundary Γ_w. The outer boundary of the reservoir is assumed impermeable to the flowing fluid. It is assumed that the fluid is slightly compressible and its flow in the reservoir is governed by Darcy's law relating the gradient of pressure in the reservoir to the filtration velocity [23, 26]. Then $u(x,t)$ corresponds to the pressure in the reservoir and the three types of boundary conditions specified on the well Γ_w correspond to different recovery regimes. Boundary condition (a) models the recovery regime with constant production rate, (b) models the recovery regime with constant wellbore pressure, and (c) models the constant wellbore pressure regime of production from a well with nonzero skin [26]. The initial conditions f_1, f_2, and f_3 take on a meaning of the pressure distribution in the reservoir Ω; hence, we will require that $f_i \geq 0$ on Ω, $i = 1,2,3$. IBVP III will be discussed in greater detail in section 5. The diffusive capacity $J(u,t)$ takes on the meaning of the PI of the well.

3. Time independent diffusive capacity. In this section we show that for each of the IBVP (I, II, and III) there exist initial distributions $f_1(x)$, $f_2(x)$, and $f_3(x)$, respectively, such that the diffusive capacity with respect to the corresponding problem is constant [16]. For Problem I, we describe the class of solutions to IBVP I on which the diffusive capacity takes a unique value. In the last subsection, the time independent values of the diffusive capacity for Problems I, II, and III are compared to each other.

3.1. IBVP I. All solutions of Problem I, for which the diffusive capacity is independent of time, possess the following property.

Remark 7. If $u(x,t)$ is a solution of Problem I and $J(u,t) = J(u)$ is constant for all $t > 0$, then there exist real constants C and B such that

$$\bar{u}_w = \frac{1}{W} \int_{\Gamma_w} u dS = C + Bt. \tag{7}$$

This can be seen from the following argument. From the definition of the diffusive capacity (5), it follows that $\bar{u}_w = -\frac{q}{J(u)} + \bar{u}_\Omega$. Hence, $\frac{\partial \bar{u}_w}{\partial t} = \frac{\partial \bar{u}_\Omega}{\partial t}$. The divergence theorem implies that

$$\frac{\partial \bar{u}_\Omega}{\partial t} = \frac{1}{V}\int_\Omega Ludx = \frac{1}{V}\int_{\Gamma_w} \frac{\partial u}{\partial \vec{\nu}} dS. \tag{8}$$

Consequently,

$$\frac{\partial \bar{u}_w}{\partial t} = -\frac{q}{V}, \tag{9}$$

from which (7) easily follows.

The "infinite conductivity of the well" assumption asserts that at each instant of time, the pressure on the wellbore is constant. Together with the latter remark, this motivated us to study the diffusive capacity on the class of solutions of Problem I, defined by $\Upsilon = \{\, u \mid \exists C$ and B are constants, such that $u(x,t) = C + Bt$ for $x \in \Gamma_w$ and for $t \geq 0\,\}$.

PROPOSITION 1. *Problem* I *has a unique solution in class* Υ.

Proof. Assume that $u \in \Upsilon$ and $v \in \Upsilon$ are solutions of Problem I. Let C_1, B_1, C_2, and B_2 be such that for $t > 0$,

$$u(x,t)|_{\Gamma_w} = C_1 + B_1 t$$

and

$$v(x,t)|_{\Gamma_w} = C_2 + B_2 t.$$

Then the difference $g(x,t) = u(x,t) - v(x,t)$ is the solution of the following IBVP:

$$Lg = \frac{\partial g}{\partial t}, \quad x \in \Omega, \quad t > 0, \tag{10}$$

$$\frac{\partial g}{\partial \vec{\nu}}|_{\Gamma_e} = 0, \tag{11}$$

$$g|_{\Gamma_w} = (C_1 - C_2) + (B_1 - B_2)t, \tag{12}$$

$$g(x,0) = 0. \tag{13}$$

In addition,

$$\int_{\Gamma_w} \frac{\partial g}{\partial \vec{\nu}} = 0. \tag{14}$$

Condition (13) immediately implies that $C_1 = C_2$.

The function $h = \frac{\partial g}{\partial t}$ is a solution of the following problem:

$$Lh = \frac{\partial h}{\partial t}, \quad x \in \Omega, \quad t > 0, \tag{15}$$

$$\frac{\partial h}{\partial \vec{\nu}}|_{\Gamma_e} = 0, \tag{16}$$

$$h|_{\Gamma_w} = B_1 - B_2, \tag{17}$$

$$h(x,0) = \frac{\partial g}{\partial t}(x,0). \tag{18}$$

In addition, from the boundary condition on Γ_w of Problem I and the divergence theorem it follows that for $t > 0$,

$$\int_\Omega h dx = \frac{\partial}{\partial t}\int_\Omega g dx = \int_\Omega L g dx \equiv 0. \tag{19}$$

As a solution of the parabolic equation (15) with the Dirichlet condition (17) on one part of the boundary $\partial\Omega$ and Neumann condition (16) on the remaining part of $\partial\Omega$, h will converge to a constant $B_1 - B_2$ on Ω as $t \to \infty$ [18]. Together with condition (19) this implies that

$$(B_1 - B_2)V = \lim_{t\to\infty}\int_\Omega h(x,t)dx = 0. \tag{20}$$

Thus, $u = v$. □

For purposes that will become clear from Proposition 2, let us introduce the following auxiliary steady-state boundary value problem. Let $u_1(x)$ be such that

$$Lu_1 = -\frac{1}{V}, \tag{21}$$

$$u_1|_{\Gamma_w} = 0, \tag{22}$$

$$\frac{\partial u_1}{\partial \vec{\nu}}|_{\Gamma_e} = 0. \tag{23}$$

Then the following proposition gives a sufficient condition providing for time independent unique diffusive capacity $J(u,t) = J(u)$.

PROPOSITION 2. *If the initial condition in Problem* I *is given by* $f_1(x) = qu_1(x) + C$ *where* u_1 *is the solution of* (21)–(23) *and* C *is an arbitrary constant such that* $f_1(x) > 0$ *for all* $x \in \Omega$, *then the diffusive capacity corresponding to a solution* $u \in \Upsilon$ *of Problem* I *is independent of time and determined by*

$$J_I := J(u,t) = \frac{V}{\int_\Omega u_1(x)dx}. \tag{24}$$

Proof. Let the initial condition in Problem I be $f_1(x) = qu_1(x)$ and

$$u(x,t) = qu_1(x) - \frac{q}{V}t. \tag{25}$$

By virtue of the divergence theorem,

$$\int_{\Gamma_w}\frac{\partial u}{\partial \vec{\nu}}dS = -q.$$

Consequently, u is a solution of IBVP I with the initial distribution $f_1(x) = qu_1(x)$.

Note that u, defined by (25), belongs to class Υ. In addition, it is clear that the diffusive capacity $J(u,t)$ on $u(x,t)$ is constant and is given by

$$J(u,t) = \frac{V}{\int_\Omega u_1(x)dx} = J_{\mathrm{I}}. \qquad \square$$

Remark 8. Function u, defined by (25), is positive on Ω only for $t \in (0,T)$, where

$$T = \frac{\min_{x\in\Omega} u_1(x)}{V}. \tag{26}$$

Solutions of Problem I represent the pressure distribution in the reservoir at time t; hence, we are interested in the positive on Ω solutions only. Therefore, the diffusive capacity (as a model of a PSS PI) $J(u,t) = J_{\mathrm{I}}$ is defined only for $t \in (0,T)$, where T is given by (26).

The necessary condition for the time independent diffusive capacity on the solutions of Problem I in class Υ is given by the following proposition.

PROPOSITION 3. *If the diffusive capacity $J(u,t)$, corresponding to a solution $u \in \Upsilon$ of Problem* I, *is constant for all $t > 0$, then*

$$\int_\Omega (u(x,0) - qu_1(x))dx + C^* = 0, \tag{27}$$

where constant C^* is independent of q and u_1 is the solution of the problem (21)–(23).

Proof. Let $u \in \Upsilon$ be a solution of Problem I such that $J(u,t) = J(u)$ is constant for all $t > 0$. Let

$$g(x,t) = u(x,t) - \left(qu_1(x) - \frac{q}{V}t\right). \tag{28}$$

There exist constants C and B such that $u|_{\Gamma_w} = C + Bt$. Moreover, by (9), $B = -\frac{q}{V}$. Hence, g is a solution of the problem

$$Lg = \frac{\partial g}{\partial t}, \quad x \in \Omega, \quad t > 0, \tag{29}$$

$$\frac{\partial g}{\partial \vec{\nu}}|_{\Gamma_e} = 0, \tag{30}$$

$$g|_{\Gamma_w} = C, \tag{31}$$

$$g(x,0) = u(x,0) - qu_1(x). \tag{32}$$

In addition, g is subject to the following condition:

$$\int_{\Gamma_w} \frac{\partial g}{\partial \vec{\nu}} dS = 0. \tag{33}$$

As a solution of the parabolic equation (29) with the boundary conditions (30) and (31), $g(x,t) \to C$ as $t \to \infty$. Together with (33), the latter implies that $\bar{g}_\Omega = C$ for all $t > 0$. Therefore, $\int_\Omega g(x,0)dx = \int_\Omega (u(x,0) - qu_1(x))dx = CV = C^*$. $\square$

Remark 9. By Proposition 3, the initial distribution providing for the time independent diffusive capacity is unique up to an additive function of zero average on Ω and an additive constant independent of the geometry of the domain or boundary conditions.

Remark 10. The integral of the solution of Problem I at $t = 0$ represents the initial reserves in the reservoir [9, 2]. The main physical consequence of Proposition 3 is that the diffusive capacity as a model of the PI uniquely determines the average initial amount of the reserves in the reservoir.

3.2. IBVP II. Let

$$\frac{\partial u_2}{\partial t} = Lu_2, \tag{34}$$

$$\frac{\partial u_2}{\partial \vec{\nu}}|_{\Gamma_e} = 0, \tag{35}$$

$$u_2|_{\Gamma_w} = 0, \tag{36}$$

$$u_2(x,0) = f_2(x) - u_{w2}. \tag{37}$$

Obviously, $u(x,t) = u_2(x,t) + u_{w2}$ solves Problem II. Then the diffusive capacity for Problem II can be expressed in terms of $u_2(x,t)$, namely,

$$J(u,t) := J(u_2,t) = \frac{\int_{\Gamma_w} \frac{\partial u_2}{\partial \vec{\nu}} dS}{-\frac{1}{V}\int_\Omega u_2(x,t)dx}. \tag{38}$$

Consider the related Sturm–Liouville problem for the elliptic operator L and the first eigenpair of the latter; i.e., let λ_0 and $\phi_0(x)$ be the first eigenvalue and first eigenfunction, respectively, of the problem

$$L\phi_0 = -\lambda_0\phi_0, \tag{39}$$

$$\phi_0|_{\Gamma_w} = 0, \tag{40}$$

$$\frac{\partial \phi_0}{\partial \vec{\nu}}|_{\Gamma_e} = 0. \tag{41}$$

Let $u_2(x,t)$ be a solution of the IBVP (34)–(37) with the initial distribution $u_2(x,0)$ equal to $\phi_0(x)$. Then $u_2(x,t) = \phi_0(x)e^{-\lambda_0 t}$ is a solution of the IBVP (34)–(37). The diffusive capacity is constant and is equal to

$$J_{\mathrm{II}} := J(u_2,t) = \frac{\lambda_0 \int_\Omega \phi_0(x)dx e^{-\lambda_0 t}}{\frac{1}{V}\int_\Omega \phi_0(x)dx e^{\lambda_0 t}} = \lambda_0 V. \tag{42}$$

This leads to the next proposition.

PROPOSITION 4. *If the initial condition of Problem* II *is given by* $f_2(x) = \phi_0(x) + u_{w2}$*, where* ϕ_0 *is the eigenfunction of problem* (39)–(41) *corresponding to the minimal eigenvalue* λ_0*, then the diffusive capacity on the solution* u *of Problem* II *is constant and is given by*

$$J(u,t) = J_{\mathrm{II}} = \lambda_0 V.$$

In fact, the diffusive capacity is constant provided that the initial distribution $u_2(x,0)$ is equal to any eigenfunction $\phi_i(x)$, $i = 1, 2 \ldots$. However, only the eigenfunction corresponding to the minimal eigenvalue does not change sign on Ω; therefore, in terms of the pressure distribution in the hydrocarbon reservoir, $\phi_0(x)$ is the only physically realistic initial distribution.

3.3. IBVP III. Let $u_3(x,t) = u(x,t) - u_{w3}$, where u solves (III) and u_{w3} is the given average value of u on Γ_w (see Remark 6). Then $u_3(x,t)$ is a solution of the reduced problem

$$Lu_3 = \frac{\partial u_3}{\partial t}, \tag{43}$$

$$\frac{\partial u_3}{\partial \vec{\nu}}|_{\Gamma_e} = 0, \tag{44}$$

$$\left(\alpha \frac{\partial u_3}{\partial \vec{\nu}} + u_3\right)|_{\Gamma_w} = 0, \tag{45}$$

$$u_3(x,0) = f_3(x) - u_{w3}. \tag{46}$$

Diffusive capacity $J(u,t)$ corresponding to Problem III is expressed in terms of $J(u_3,t)$ in the following way:

$$J(u,t) = J(u_3,t) = \frac{\int_{\Gamma_w} \frac{\partial u_3}{\partial \vec{\nu}} dS}{-\frac{1}{V}\int_\Omega u dx}. \tag{47}$$

Physically, the Robin boundary condition on Γ_w in Problem III corresponds to production from a well with a thin-skin zone with constant wellbore pressure (constant $\bar{u}_3|_{\Gamma_w}$) [26]. A sufficient condition for the diffusive capacity to be constant is similar to that for Problem II.

In particular, consider the related Sturm–Liouville problem. Let λ_k^α and $\phi_k^\alpha(x)$ be an eigenpair of the problem

$$L\phi_k^\alpha = -\lambda_k^\alpha \phi_k^\alpha, \tag{48}$$

$$\frac{\partial \phi_k^\alpha}{\partial \vec{\nu}}|_{\Gamma_e} = 0, \tag{49}$$

$$\phi_k^\alpha + \alpha \frac{\partial \phi_k^\alpha}{\partial \vec{\nu}}|_{\Gamma_w} = 0. \tag{50}$$

Here, the superscript α is intended to emphasize that the solution and, hence, the diffusive capacity of Problem III depend on the value of parameter α. This dependence will be analyzed in subsequent sections. Let $u_3(x,t)$ be a solution of the IBVP (43)–(46) with the initial distribution $u_3(x,0) = \phi_k^\alpha(x)$. Then $u_3(x,t) = \phi_k^\alpha(x) e^{-\lambda_k^\alpha t}$ solves (43)–(46) and the diffusive capacity is time independent.

When parameter α in Problem III is positive, then the minimal eigenvalue λ_0^α is positive and the corresponding eigenfunction $\phi_0^\alpha(x)$ does not change sign on Ω.

In section 5 we will show that the boundary condition on Γ_w of Problem III models skin effect for a damaged well produced with a constant wellbore pressure. As mentioned in section 1, the production from a stimulated well is modeled by a negative skin factor s; therefore, we will analyze the behavior of the diffusive capacity on the solutions of Problem III for negative values of parameter α. The latter case will be discussed in more detail in section 5. For the purposes of this section, it is

sufficient to note that when $\alpha < 0$, the minimal eigenvalue and hence the constant diffusive capacity may be negative. Negative PI is an indication of injection into the well; therefore, to avoid the contradiction, our attention will be restricted to positive eigenvalues only. The analysis of the first eigenfunction will be given in section 5.

Regardless of the sign of α, let λ_0^α be the first nonnegative eigenvalue. If the initial distribution in (43)–(46) is equal to the corresponding eigenfunction, the constant diffusive capacity is given by

$$J_{\mathrm{III}}(\alpha) := J(u_3, t) = \lambda_0^\alpha V. \tag{51}$$

Therefore, we have shown the following proposition.

PROPOSITION 5. *If the initial condition of Problem* III *is given by* $f_3(x) = \phi_0^\alpha(x) + u_{w3}$, *where* ϕ_0^α *is the eigenfunction of problem* (48)–(50) *corresponding to the minimal positive eigenvalue* λ_0^α, *then the diffusive capacity on the solution* u *of Problem* III *is constant and is given by*

$$J(u,t) = J_{\mathrm{III}}(\alpha) = \lambda_0^\alpha V.$$

3.4. Comparison of the time independent diffusive capacities for Problems I, II, and III. The steady-state auxiliary problem (21)–(23) introduced earlier has a convenient variational formulation which facilitates deriving an important relation between the time independent diffusive capacities of Γ_w with respect to Γ_e in Ω.

Assume that solutions of Problems I, II, and III satisfy the conditions in Propositions 2, 4, and 5, respectively. Then the diffusive capacities for Problems I, II, and III (J_{I}, J_{II}, and $J_{\mathrm{III}}(\alpha)$) are time independent and their values are given by (24), (42), and (51), respectively.

Let $H^{1,2}(\Omega)$ be the Sobolev space [1]. Denote by $\overset{\circ}{H}{}^{1,2}(\Omega, \Gamma_w)$ the closure in the $H^{1,2}(\Omega)$ norm of smooth functions that vanish on Γ_w, and denote by $\overset{\circ}{H}{}^{1,2}(\Omega, \Gamma_w, \alpha)$ the closure in the $H^{1,2}(\Omega)$ norm of smooth functions such that $(u + \alpha \frac{\partial u}{\partial \vec{\nu}})|_{\Gamma_w} = 0$ [1].

The following are well-known variational principles yielding the first eigenvalues λ_0 and λ_0^α of the problems (39)–(41) and (48)–(50), respectively (see [5]):

$$\lambda_0 = \inf_{u \in \overset{\circ}{H}{}^{1,2}(\Omega, \Gamma_e)} \frac{\int_\Omega A \nabla u \cdot \nabla u dx}{\int_\Omega u^2 dx}, \tag{52}$$

$$\lambda_0^\alpha = \inf_{u \in \overset{\circ}{H}{}^{1,2}(\Omega, \Gamma_w \alpha)} \frac{\int_\Omega A \nabla u \cdot \nabla u dx + \frac{1}{\alpha} \int_{\Gamma_w} u^2 dS}{\int_\Omega u^2 dx}. \tag{53}$$

These two principles imply that for any positive α_1 and α_2 such that (see [5]) $\alpha_1 > \alpha_2$, $\lambda_0^{\alpha_1} < \lambda_0^{\alpha_2}$. Moreover, $\lambda_0^\alpha \nearrow \lambda_0$ as $\alpha \searrow 0$. This leads to the next proposition.

PROPOSITION 6. *If the initial conditions in Problems* II *and* III *are such that* J_{II} *and* $J_{\mathrm{III}}(\alpha)$ *are time independent and* $\alpha \searrow 0$, *then* $J_{\mathrm{III}}(\alpha) \nearrow J_{\mathrm{II}}$.

Another important comparison can be made between the time independent capacities for Problems I and II.

THEOREM 1. *If the initial conditions in Problems* I *and* II *are such that the diffusive capacities* J_{I} *and* J_{II} *are time independent, then*

$$J_{\mathrm{II}} \le J_{\mathrm{I}} \le C_\Omega J_{\mathrm{II}}, \tag{54}$$

where $C_\Omega = \frac{\max_\Omega \phi_0}{\bar{\phi}_0}$.

Proof. Let $u_1 \in \overset{\circ}{H}{}_2^1(\Omega, \Gamma_w)$ be a solution of the problem (21)–(23). We need to show that

$$\frac{1}{\int_\Omega u_1(x)dx} \geq \lambda_0.$$

From (52) it follows that

$$\lambda_0 \leq \frac{\int_\Omega (\nabla u_1)\cdot(A\nabla u_1)dx}{\int_\Omega u_1^2 dx}. \tag{55}$$

Using the identity

$$\nabla\cdot(u_1 A\nabla u_1) = (\nabla u_1)\cdot(A\nabla u_1) - u_1\nabla\cdot(A\nabla u_1),$$

applying the divergence theorem to the numerator, and making use of (21)–(23), we obtain

$$\lambda_0 \leq \frac{1}{V}\frac{\int_\Omega u_1 dx}{\int_\Omega u_1^2 dx}. \tag{56}$$

The last inequality can be rewritten as

$$\lambda_0 \leq \frac{1}{V}\frac{\left(\int_\Omega u_1 dx\right)^2}{\int_\Omega u_1^2 dx}\frac{1}{\int_\Omega u_1 dx}. \tag{57}$$

The first part of (54) now follows from Hölder's inequality.

Let $u_1(x)$ be a solution of (21)–(23) and ϕ_0 of (39)–(41). After multiplication of both sides of (21) by ϕ_0, using the symmetry of A in the identity

$$(\nabla\cdot(A\nabla u_1))\phi_0 = \nabla\cdot(\phi_0 A\nabla u_1) - \nabla\cdot(u_1 A\nabla\phi_0) + \nabla\cdot(A\nabla\phi_0)u_1, \tag{58}$$

followed by integration over Ω, from the divergence theorem one concludes that

$$\lambda_0 \max_\Omega \phi_0 \int_\Omega u_1 dV \geq \lambda_0 \int_\Omega u_1\phi_0 dV = \frac{1}{V}\int_\Omega \phi_0 dV = \bar{\phi}_0. \tag{59}$$

The latter can be recast as the second part of (54), using the positivity of u_1 and ϕ_0. ∎

Remark 11. The constant C_Ω is a peak-to-average ratio and has a clear physical meaning [24].

4. Transient diffusive capacity. In section 3 it was shown that the PI of a well in a reservoir is constant for all $t > 0$ provided that the pressure distribution at $t = 0$ satisfies certain conditions. The PI is known to stabilize in a long time asymptote regardless of the initial pressure distribution. In this chapter we will consider a transient diffusive capacity and investigate questions related to its stabilization. Thus, we will analyze Problems I and II with arbitrary initial conditions. The only restriction that is imposed on the initial conditions f_1 and f_2 of Problems I and II, respectively, is motivated by physical considerations: we require that f_1 and f_2 be positive smooth functions on Ω.

4.1. IBVP I: Constant production rate regime. In section 1 it was mentioned that the constant rate regime is usually modeled with one of two assumptions: at each time $t > 0$ either the pressure or the pressure flux is assumed to be constant on the wellbore. Proposition 2 shows that the condition of a constant wellbore pressure at each time $t > 0$ (infinite conductivity condition) is equivalent to the conditions of the PSS, i.e., the PI of a well is time independent. In this section we will show that the diffusive capacity on the class Υ of solutions of Problem I (defined in section 3) is stable with respect to small perturbations of boundary conditions. Recall that Υ is the class of solutions u of Problem I such that at each time $t > 0$, u is constant on Γ_w. Then the stability of J is established by the following proposition.

PROPOSITION 7. *Let $v(x,t)$ be a solution of Problem* I *such that $v(x,t) = Bt + C$ for all $x \in \Gamma_w$. Let $u(x,t)$ be a solution of Problem* I *such that $u(x,t) = Bt + C + h(x,t)$ for all $x \in \Gamma_w$, where $h(x,t)$ is a smooth, bounded function. For any $\epsilon > 0$, there exists $\delta > 0$ such that if for all $t > 0$, $|h(x,t)| \le \delta$ for all $x \in \Gamma_w$, then $|J(u,t) - J(v,t)| \le \epsilon$ for all $t > 0$.*

Proof. Function $\tilde{v}(x,t) = u(x,t) - v(x,t)$ is a solution of the following problem:

$$L\tilde{v} = \frac{\partial \tilde{v}}{\partial t}, \quad x \in \Omega, \quad t > 0, \tag{60}$$

$$\frac{\partial \tilde{v}}{\partial \vec{\nu}}|_{\Gamma_e} = 0, \tag{61}$$

$$\tilde{v}|_{\Gamma_w} = h(x,t), \tag{62}$$

$$\tilde{v}(x,0) = 0. \tag{63}$$

The maximum principle for parabolic equation (60) implies that $|\tilde{v}(x,t)| \le \delta$ for all $x \in \Omega$ and $t \ge 0$. Since $\int_{\Gamma_w} \frac{\partial u}{\partial \nu} dS = \int_{\Gamma_w} \frac{\partial u}{\partial \nu} dS = -q$ for $t \ge 0$,

$$\left| \frac{1}{J(v,t)} - \frac{1}{J(u,t)} \right| \le \frac{1}{q} \left| \frac{1}{W} \int_{\Gamma_w} (u - v) dS + \frac{1}{V} \int_{\Omega} (u - v) dx \right|.$$

Hence, $|\frac{1}{J(v,t)} - \frac{1}{J(u,t)}| \le \delta$. ∎

One should note that J_{I} is shown to be a PSS PI of a well only for solutions of Problem I that belong to class Υ. The extent to which the assumption of the infinite conductivity of the well is realistic for various reservoir-well configurations will be discussed in more detail in section 7. Below we investigate the question of the uniqueness of the PSS PI. Recall that the PSS PI is a constant value of the diffusive capacity on the solutions to Problem I.

Remark 12. J_{I} is not necessarily a unique constant value of the diffusive capacity on the solutions to Problem I.

This is established by the following argument. Consider solutions to Problem I with a constant flux on Γ_w; i.e., let $u(x,t)$ be a solution of the following problem:

$$Lu = \frac{\partial u}{\partial t}, \quad x \in \Omega, \quad t > 0, \tag{64}$$

$$\frac{\partial u}{\partial \vec{\nu}}|_{\Gamma_e} = 0, \tag{65}$$

$$\frac{\partial u}{\partial \vec{\nu}}|_{\Gamma_w} = -\frac{q}{W}, \tag{66}$$

$$u(x,0) = f_1(x). \tag{67}$$

The solution to (64)–(67) is given (up to an additive constant) by $u(x,t) = qv - \frac{q}{V}t + h(x,t)$, where $v(x)$ is a solution of the steady-state problem

$$Lv = -\frac{1}{V}, \quad x \in \Omega, \tag{68}$$

$$\frac{\partial v}{\partial \vec{\nu}}|_{\Gamma_e} = 0, \tag{69}$$

$$\frac{\partial v}{\partial \vec{\nu}}|_{\Gamma_w} = -\frac{1}{W}, \tag{70}$$

and $h(x,t)$ is a solution of the corresponding problem with homogeneous boundary conditions:

$$Lh = \frac{\partial h}{\partial t}, \quad x \in \Omega, \quad t > 0, \tag{71}$$

$$\frac{\partial h}{\partial \vec{\nu}}|_{\Gamma_e} = 0, \tag{72}$$

$$\frac{\partial h}{\partial \vec{\nu}}|_{\Gamma_w} = 0, \tag{73}$$

$$h(x,0) = f_1(x) - qv(x). \tag{74}$$

The solution to (71)–(74) is given by $h(x,t) = \sum_{n=0}^{\infty} c_n \phi_n(x) e^{-\lambda_n t}$, where $\phi_n(x)$ and λ_n are solutions of the related Sturm–Liouville problem and c_n are the coefficients of the Fourier expansion of $h(x,0)$ in terms of ϕ_n. The diffusive capacity $J(u,t)$ is given by

$$J(u,t) = \frac{-q}{\bar{v}_w - \bar{v}_\Omega + \bar{h}_w - \bar{h}_\Omega}. \tag{75}$$

Note that $\bar{v}_w$ and $\bar{v}_\Omega$ are constant, while $\bar{h}_w$ and $\bar{h}_\Omega$ are functions of time. Clearly, the difference $\bar{h}_w - \bar{h}_\Omega = \sum_{n=0}^{\infty} c_n(\bar{\phi}_{nw} - \bar{\phi}_{n\Omega}) e^{-\lambda_n t}$ converges to a constant as $t \to \infty$. Therefore, $J(u,t)$ converges to a constant value $\hat{J}$ as $t \to \infty$. However, $\hat{J}$ is not necessarily equal to J_{I}.

Henceforth, we do not address the uniqueness of the constant diffusive capacity on the solutions of Problem I and, consequently, of the PSS PI. In the subsequent sections we will refer to J_{I} as the value of the PSS PI, thus implicitly assuming that the wellbore has an infinite conductivity.

4.2. IBVP II: Constant wellbore pressure regime. For simplicity, consider the following problem for a parabolic equation. Let $u(x,t)$ be a solution of

$$Lu = \frac{\partial u}{\partial t}, \quad x \in \Omega, \quad t \geq 0, \tag{76}$$

$$\frac{\partial u}{\partial \vec{\nu}}|_{\Gamma_e} = 0, \tag{77}$$

$$u|_{\Gamma_w} = 0, \tag{78}$$

$$u(x,0) = u_0(x), \tag{79}$$

where $u_0(x) > 0$. Then the diffusive capacity is simply

$$J(u,t) = V \frac{\int_{\Gamma_w} \frac{\partial u}{\partial \vec{\nu}} dS}{\int_\Omega u dx}. \tag{80}$$

Along with (76)–(79), consider the related Sturm–Liouville problem for the operator L,

$$L\phi_k = -\lambda_k \phi_k, \qquad x \in \Omega, \quad t \geq 0, \tag{81}$$

$$\frac{\partial \phi_k}{\partial \vec{\nu}}|_{\Gamma_e} = 0, \tag{82}$$

$$\phi_k|_{\Gamma_w} = 0. \tag{83}$$

Let $\{\phi_k(x)\}_{k=0}^\infty$ be an orthonormal family of solutions of (81)–(83) with respect to the usual inner product in $L^2(\Omega)$. Define $d_k = \int_\Omega \phi_k(x)dx$ and $c_k = \int_\Omega u_0(x)\phi_k(x)dx$. Then the diffusive capacity can be written as

$$J(u,t) = V \frac{\sum_{k=0}^\infty c_k \lambda_k d_k e^{-\lambda_k t}}{\sum_{k=0}^\infty c_k d_k e^{-\lambda_k t}}.$$

The latter can be recast into

$$J(u,t) = V\lambda_0 \left[1 + \frac{\sum_{k=1}^\infty \frac{c_k}{c_0}\frac{d_k}{d_0}\left(\frac{\lambda_k}{\lambda_0} - 1\right) e^{-(\lambda_k - \lambda_0)t}}{1 + \sum_{k=1}^\infty \frac{c_k}{c_0}\frac{d_k}{d_0} e^{-(\lambda_k - \lambda_0)t}} \right]. \tag{84}$$

Since $\lambda_0 < \lambda_1 < \lambda_3 < \cdots$, as $t \to \infty$, $J(u,t) \to \lambda_0 V$. This proves the following.

PROPOSITION 8. *If u is a solution of IBVP* II, *then the diffusive capacity $J(u,t)$ converges to the constant value J_{II} as $t \to \infty$ for any initial condition f_2.*

In terms of the PI, Proposition 8 can be rephrased in the following way: if a well is produced with a constant wellbore pressure, the PI stabilizes to constant value J_{II} as $t \to \infty$ regardless of the initial pressure distribution.

Note that since the initial condition $u_0(x)$ is positive on Ω, $c_0 > 0$ and $d_0 > 0$. From the maximum principle for parabolic equation (76) it follows that $u(x,t) \geq 0$

for all $t > 0$. Consequently, the denominator in (84), equal to $\int_\Omega u(x,t)dx/c_0d_0e^{-\lambda_0 t}$, is positive for all $t > 0$. Therefore, from (84) follows the next remark.

Remark 13. If in (84) $c_kd_k > 0$ for any k, then $J(u,t) \searrow \lambda_0 V$.

The last observation allows one to analyze several physically important examples of the transient PI in terms of the diffusive capacity on the solutions of the IBVP for a parabolic equation.

Example 1. Suppose that a well is produced with constant rate, the PI is constant, and the well has infinite conductivity. Then the pressure in the reservoir $u(x,t)$ is determined (up to an additive constant) by $u(x,t) = qu_1(x) - \frac{q}{V}t$ (see Proposition 2), where $u_1(x)$ is a solution of the auxiliary steady-state problem

$$Lu_1(x) = -\frac{1}{V}, \quad x \in \Omega,$$

$$\frac{\partial u_1}{\partial \vec{\nu}}|_{\Gamma_e} = 0,$$

$$u_1|_{\Gamma_w} = 0.$$

Suppose that at some time $t_0 > 0$, the production regime was changed to a constant wellbore pressure production. Then the pressure in the reservoir $u(x,t)$ for $t > t_0$ is defined by $u(x,t) = v(x,t-t_0) - \frac{q}{V}(t-t_0)$, where $v(x,t)$ is a solution of the problem

$$Lv(x) = -\frac{\partial v}{\partial t}, \quad x \in \Omega, \quad t > 0,$$

$$\frac{\partial v}{\partial \vec{\nu}}|_{\Gamma_e} = 0,$$

$$v|_{\Gamma_w} = 0,$$

$$v(x,0) = qu_1(x).$$

The diffusive capacity $J(u,t) = J(v,t)$, where $v(x,t)$ is defined by

$$v(x,t) = \sum_{n=0}^{\infty} c_k\phi_k(x)e^{-\lambda_k t},$$

where $c_k = q\int_\Omega u_1(x)\phi_k dx$. Using integration by parts, we obtain

$$\int_\Omega Lu_1\phi_k = \int_\Omega u_1 L\phi_k.$$

Hence,

$$\frac{1}{V}\int_\Omega \phi_k = \lambda_k \int_\Omega u_1\phi_k.$$

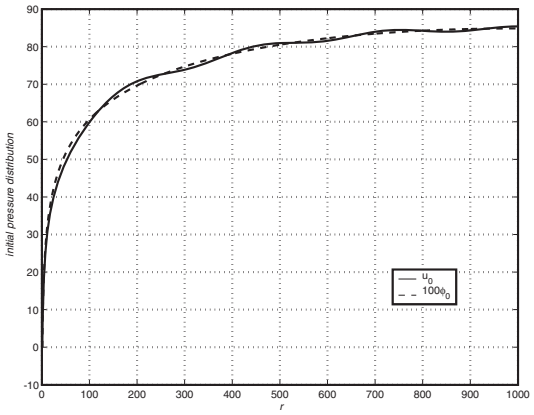


FIG. 1. *Radial profile of an initial distribution yielding small diffusive capacity.*

Thus, for any $k = 1, 2 \dots$, $d_k c_k > 0$ and (84) implies that $J(u,t) \searrow J_{\text{II}}$. In other words, when the regime of production changes from PSS, i.e., constant flow rate, to constant wellbore pressure, the PI monotonically decreases to the BD PI.

Example 2. For the purpose of analysis it is frequently assumed that at $t = 0$ the pressure in the reservoir is distributed uniformly, i.e., $u_0(x) = u_i$, where u_i is a positive constant. Then $c_k = u_i d_k$ and the PI is monotonically decreasing to the BD PI.

Finally, consider an example of the initial pressure distribution yielding the PI which is less than the BD PI.

Example 3. Let $u_0(x) = 100\phi_0(x) - 3\phi_1(x)$. Then the diffusive capacity $J(u,t) < \lambda_0 V$.

An example of such initial distribution for an ideal cylindrical reservoir with vertical fully penetrated well is given in Figure 1, where the radial profile of $u_0(r)$ is given. The dimensionless radius of the reservoir is equal to $R_D = 1000$. Physically this example may be interpreted as follows. Assume that the reservoir has been depleted by a set of wells. Suppose that the old wells are shut down and a new well is drilled and produced. Then the PI of the new well will monotonically increase to the BD PI value.

5. Model of the skin effect. Stabilized production with constant rate is characterized by the PSS PI. When the well is damaged, the value of the PI is less than what is predicted by the model. As described in section 1, such effect is called thin-skin effect. To take into account the skin effect, the PSS PI is corrected according to the equation

$$PI_{PSS,skin} = \frac{1}{\frac{1}{PI_{PSS}} + s}, \tag{85}$$

where s is the so-called *skin factor* or simply *skin*. The skin factor concept was originally introduced to describe the behavior of damaged wells. Others have extended the idea to stimulated wells which have a higher PI than the PSS PI of an ideal well. In [15] it was shown that a negative skin s corresponds to a stimulated well.

All existing results on modeling the skin effect pertain to the constant rate production regime. In this section it will be shown that for the constant wellbore pressure

production regime, the skin effect can be modeled by a third boundary condition specified on the well boundary.

5.1. Diffusive capacity for IBVP III in an annulus. Let $u(r,t)$ be a solution of the problem

$$\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right)=\frac{\partial u}{\partial t},\quad 1<r<R_D,\quad t>0, \tag{86}$$

$$\frac{\partial u}{\partial r}|_{r=R_D}=0, \tag{87}$$

$$\left(u+\alpha\frac{\partial u}{\partial r}\right)|_{r=1}=0, \tag{88}$$

$$u(r,0)=u_0(r). \tag{89}$$

Problem (86)–(89) models the axisymmetric flow of oil in an ideal isolated circular reservoir with a perfect circular well situated in the center. Here, $u(r,t)$ is the dimensionless pressure in the reservoir, the dimensionless formation permeability is 1, and the dimensionless outer radius is equal to R_D. The dimensionless wellbore radius is equal to 1. Constant wellbore pressure production is assumed. The thin skin zone adjacent to the well has a permeability below than that of the formation.

We will call a production regime for a well with a thin skin zone characterized by a constant PI a *generalized BD state*. When $\alpha=0$ (no damaged zone around the well), it is a BD regime.

Along with problem (86)–(89), consider a related Sturm–Liouville problem:

$$\frac{\partial}{\partial r}\left(r\frac{\partial \phi_k^\alpha}{\partial r}\right)=-\lambda_k^\alpha \partial\phi_k^\alpha,\quad 1<r<R_D,\quad t>0, \tag{90}$$

$$\frac{\partial \phi_k^\alpha}{\partial r}|_{r=R_D}=0, \tag{91}$$

$$\left(\phi_k^\alpha+\alpha\frac{\partial \phi_k^\alpha}{\partial r}\right)|_{r=1}=0. \tag{92}$$

Let λ_0^α be the minimal nonnegative eigenvalue of the problem (90)–(92). If the initial condition $u_0(r)=\phi_0^\alpha$ is the eigenfunction corresponding to λ_0^α, then by Propositions 5 and 6 the generalized BD PI is determined by $J_{\mathrm{III}}(\alpha)=\lambda_0^\alpha V$ and $J_{\mathrm{III}}(0)=J_{\mathrm{II}}$.

In analogy to (85), we define the skin factor s by

$$s=s(\alpha):=\frac{1}{J_{\mathrm{III}}(\alpha)}-\frac{1}{J_{\mathrm{II}}}=\frac{1}{J_{\mathrm{III}}(\alpha)}-\frac{1}{J_{\mathrm{III}}(0)}. \tag{93}$$

Positive skin defined by (93) is evidence of a damaged well. By analogy, the generalized BD index of a stimulated well should be greater than the BD index, yielding negative skin s.

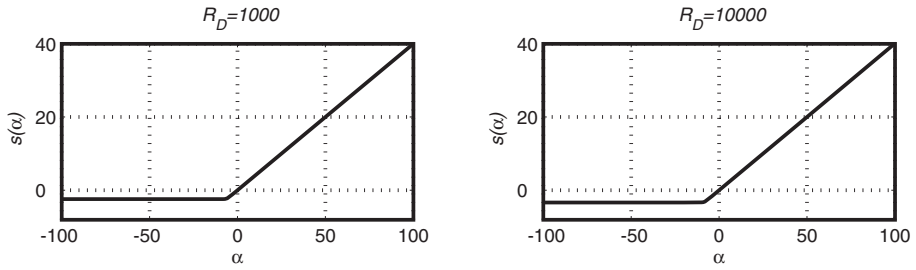


FIG. 2. *Graph of $s(\alpha)$ for $R_D = 1000$ (left panel) and for $R_D = 10,000$ (right panel).*

When $\alpha < 0$, λ_0^α is the first positive eigenvalue. The eigenpair solves known equations involving Bessel functions of the first and the second kind. Using known facts from the theory of Bessel functions, it is not hard to show the following.

PROPOSITION 9. *As $\alpha \to \infty$, $\lambda_0^\alpha \to 0$. As $\alpha \to -\infty$, $\lambda_0^\alpha \to \lambda_0^{(N)}$, where $\lambda_0^{(N)}$ is the minimal nontrivial eigenvalue of the following problem:*

$$\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) = \frac{\partial u}{\partial t}, \quad 1 < r < R_D, \quad t > 0, \tag{94}$$

$$\frac{\partial u}{\partial r}|_{r=R_D} = 0, \tag{95}$$

$$\frac{\partial u}{\partial r} = 0, \tag{96}$$

$$u(r,0) = u_0(r). \tag{97}$$

This implies, in particular, that $s(\alpha)$, defined by (93), is bounded from below, since $\lambda_0^{(N)}$ is bounded from above. The relation between s and α for $R_D = 1000$ and $R_D = 1000$ is shown in Figure 2 for a range of values of α. Figure 2 illustrates that when $\alpha > 0$, skin $s = \alpha$, i.e., the positive skin can be successfully modeled by the third boundary condition, in perfect agreement with the constant rate case. To analyze the case of $\alpha < 0$, additional considerations are necessary.

Eigenfunctions ϕ_0^α corresponding to the minimal positive eigenvalue λ_0^α of the problem (90)–(92) for two sample positive and negative values of α are pictured in Figure 3. As seen in Figure 3, for negative α the corresponding eigenfunction ϕ_0^α changes sign on the interval $1 < r < R_D$. Recall that the initial condition of the problem (86)–(89) u_0 is equal to ϕ_0^α. Consequently, the sufficient condition for the generalized BD state is such that the initial pressure distribution in the reservoir is not everywhere positive. Thus, a negative value of the skin factor s creates a physical contradiction, and problem (86)–(89) with $\alpha < 0$ cannot serve as an appropriate model for a stimulated well.

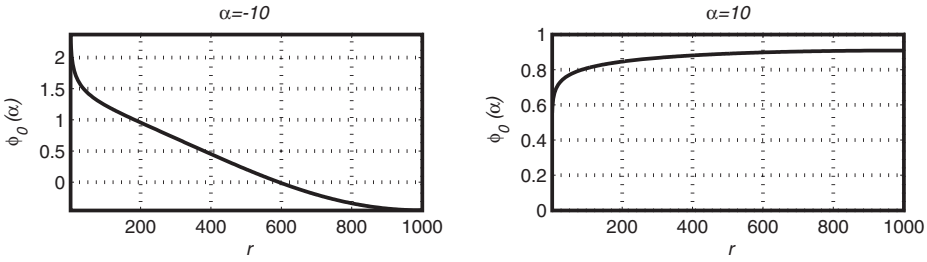


FIG. 3. *Eigenfunctions for negative α (left panel) and positive α (right panel).* $R_D = 1000$.

6. PI in a two-dimensional reservoir. In this chapter we present a numerical study of the diffusive capacity/PI in two-dimensional domains. We will restrict our attention to PSS and BD productivity indices only, that is, we will consider only IBVPs I and II.

If the thickness of the reservoir is uniform, then for a fully penetrated vertical well the three-dimensional problem reduces to a two-dimensional one. Since the radius of wellbore is small compared to the dimensions of the reservoir, we can assume that the pressure is uniformly distributed on the wellbore. Therefore, for a two-dimensional problem, the PSS PI is equal to J_{I} given by (24).

Under the assumption that the reservoir is ideal and the well is perfectly circular, vertical, and fully penetrated, the IBVPs I and II can be formulated in terms of dimensionless variables as follows. Let $\Omega \in \mathbb{R}^2$ be the horizontal cross-section of such a reservoir. Let $\{r, \theta\}$ be a polar coordinate system specified on Ω along with the Cartesian coordinate system $\{x, y\}$. The origins of both coordinate systems are located at the center of the well, which is represented by a circle with equation $r = 1$. Let R_D be the radius of the circle of the same area as Ω. Then the dimensionless area V of Ω is equal to $(R_D^2 - 1)/2$. As before, let Γ_e denote the exterior boundary of Ω. The auxiliary steady-state problem (21)–(23) and the Sturm–Liouville problem (39)–(41) can be written as

$$\frac{\partial^2 u_1}{\partial x^2} + \frac{\partial^2 u_1}{\partial y^2} = -\frac{1}{V}, \tag{98}$$

$$u_1|_{r=1} = 0, \tag{99}$$

$$\frac{\partial u_1}{\partial \vec{n}}|_{\Gamma_e} = 0, \tag{100}$$

and

$$\frac{\partial^2 \phi_0}{\partial x^2} + \frac{\partial^2 \phi_0}{\partial y^2} = -\lambda_0 \phi_0, \tag{101}$$

$$\phi_0|_{r=1} = 0, \tag{102}$$

$$\frac{\partial \phi_0}{\partial \vec{n}}|_{\Gamma_e} = 0, \tag{103}$$

respectively. By Propositions 2 and 4, the values of the PSS and BD PIs are given by the following equations, respectively:

$$J_{\mathrm{I}} = \frac{V}{\int_\Omega u_1 dx} \tag{104}$$

and

$$J_{\mathrm{II}} = \lambda_0 V. \tag{105}$$

As the first stage, J_{I} and J_{II} values were compared to the values obtained by Dietz's equation (2) for domains in which (2) can be applied, that is, for domains with polygonal exterior boundaries: rectangle, triangle, circle, romb, and hexagon. Value J_{I} was compared to the value of the PSS PI J_{PSS} computed by (2) with the shape factors C_A taken from [6] for every considered shape. The constant diffusive capacity J_{II}, given by (105), was compared to the PI J_{BD} computed by (2) with the BD shape factors C_A provided in [13]. The results were obtained for two values of the dimensionless radius R_D of the drainage area, $R_D = 1000$ and $R_D = 10,000$.

The obtained results are not presented here due to limited space, but (104) and (105) closely agree to the corresponding existing formulas. The largest difference between the corresponding values is the one between J_{I} and J_{II} in the drainage areas where the well is located far from the center of symmetry of the domain.

As noted, one of the disadvantages of (2) is that it cannot be applied to the drainage area shapes that do not satisfy the requirements of the method of images. On the other hand, (104) and (105) are valid for all drainage area shapes and can be applied to a general reservoir without the usual assumptions of the homogeneity and isotropy of the media. Below we exploit these useful features of the new formulas for PI to analyze its behavior in more complex geometries and for anisotropic media. Then, using the new method we will evaluate the diffusive capacity in domains with more complex geometry, revealing some geometric characteristics of the domain that lead to the nonnegligible difference between J_{I} and J_{II}.

6.1. PI in domains violating isoperimetric inequality. Theorem 1 of section 3 gives the means to investigate more deeply the effects on the difference between J_{I} and J_{II} of the shape of the exterior boundary of the domain. The difference between J_{I} and J_{II} is expected to be greater when the constant C_Ω on the right-hand side of inequality (54) is much greater than 1. The constant C_Ω is, in its turn, determined by the minimal eigenvalue λ_0 and the behavior of the corresponding eigenfunction ϕ_0 of the elliptic problem (101)–(103).

The first eigenpair of the problem is directly related to the geometry of the domain, namely, to the symmetry and curvature of the exterior boundary and the shape of the well boundary. To illustrate the effect of the curvature and the symmetry of the exterior boundary, consider domains in Figure 4 (A) and (B). If the domain does not satisfy the classical isoperimetric inequality, the first eigenvalue of the problem (101)–(103) can be small enough in comparison to C_Ω to make the difference between J_{I} and J_{II} significant [22]. It is not hard to show that for $0 < \epsilon < 1$, both domains pictured in Figure 4 violate the classical isoperimetric inequality [22]. For either shape, the domain parameters b and ϵ change so that the ratio of the area of the domain to the radius of the well is held constant and corresponds to $R = 1000$. The circular well is located in the center of the area. The results of the numerical investigation for domains violating the classical isoperimetric inequality [22] are collected in Table 1.

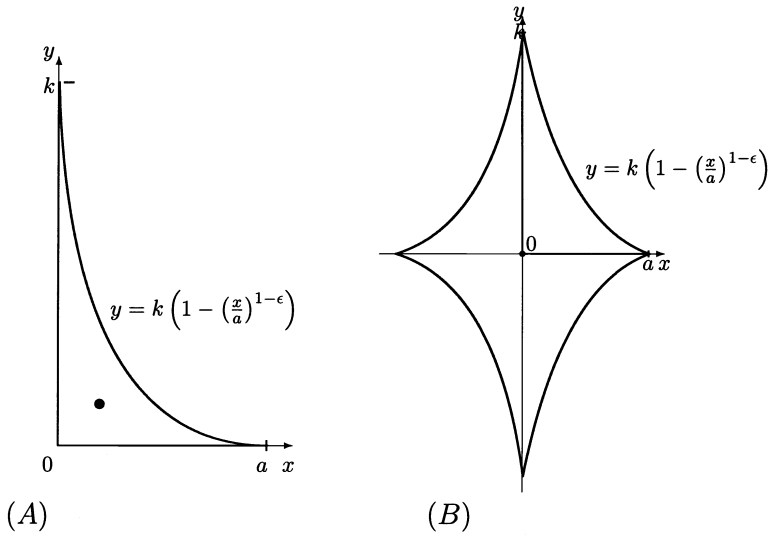


FIG. 4. *Domains violating isoperimetric inequality.*

TABLE 1
The difference between J_{I} and J_{II} in domains violating the isoperimetric inequality.

Shape	ϵ	J_{I}	J_{II}	$\left\lvert\frac{J_{\mathrm{I}}-J_{\mathrm{II}}}{J_{\mathrm{II}}}\right\rvert$, percent
	0.0	0.1227	0.1065	4.69
	0.4	0.0539	0.4370	23.34
	0.6	0.0137	0.0100	37.00
see Figure 4 (A)	0.8	0.0071	0.005	39.22
	0.8	0.0990	0.1222	19.00
see Figure 4 (B)	0.95	0.0056	0.0311	82.00

The symmetrical domain is presented to illustrate the importance of symmetry: the difference between J_{I} and J_{II} for a symmetrical domain is significantly less than for a nonsymmetrical domain with the same curvature of the exterior boundary.

7. PI in a three-dimensional reservoir. As described in the introduction, the existing methods for evaluating the PI have two major drawbacks. First, the evaluation of a PI requires solving a transient problem in a period long enough for the pressure to reach a PSS. When the well is not fully penetrated or directionally drilled (deviated or horizontal), the period necessary for the pressure to stabilize may become excessively long, creating difficulties for computational procedures. To address the problem of excessively long computations, some simplifying assumptions are made. Most of the methods are based on the assumption that the thickness of the reservoir is small enough to make the flow in the vertical direction negligible or so insignificant that its impact on the distribution of pressure can be included in a skin factor [20, 11]. With the restriction on the reservoir thickness, the problem reduces to a two-dimensional one. Then the techniques for two-dimensional reservoirs can be applied. The majority of such techniques utilize the method of images, creating the second drawback—restrictions on the geometry of the domain.

With this in mind, a number of numerical experiments were conducted for various

well configurations in three-dimensional domains. Here we illustrate the behavior of the PIs in a general homogeneous three-dimensional reservoir/well system. Equations (24) and (42) are convenient to use in such settings, since they require only solution of steady-state three-dimensional problems. Note that the use of (24) implies that in a constant rate of production regime, the pressure is uniformly distributed on the wellbore at each $t > 0$. One can argue that this assumption is physically realistic for horizontal wells of any length, if we assume that the wellbore has infinite conductivity so that the pressure of the fluid entering the wellbore instantly equalizes at every point of the wellbore. For vertical or slanted wells, the assumption of uniform pressure distribution on the wellbore at each $t > 0$ implies that we neglect gravity effects. Certainly, for long vertical or slanted wells, this assumption is not physically realistic.

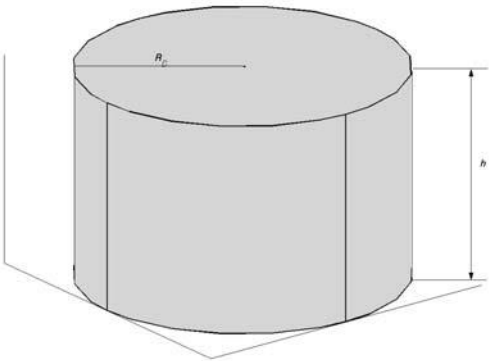


FIG. 5. *Schematic representation of domain* D_1.

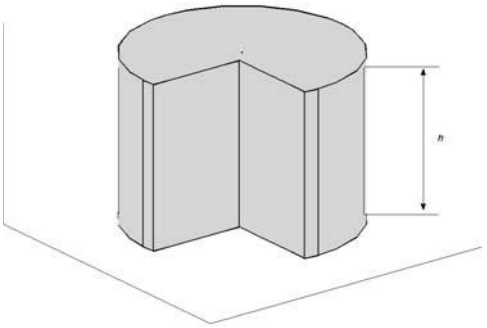


FIG. 6. *Schematic representation of domain* D_2.

Two domains modeling three-dimensional reservoirs that were considered for the numerical study are depicted in Figures 5, 6, and 7. Domain D_1 is a cylindrical reservoir of uniform thickness h and the dimensionless radius R_D. Analogously to the two-dimensional definition, R_D is defined as the ratio of the radius of the horizontal cross-section (in this case, circle) to the well radius. The value of R_D is set to 1000 for all settings. For consistency of comparisons made below, the radius of the circle

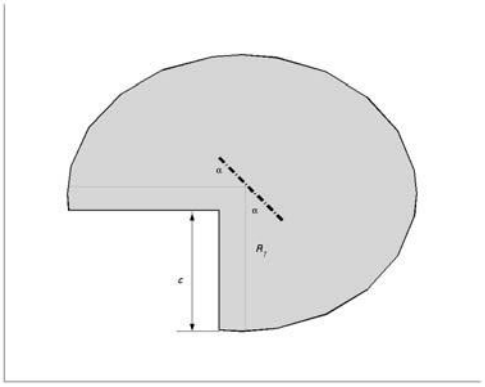


FIG. 7. *Schematic representation of horizontal projection of domain* D_2.

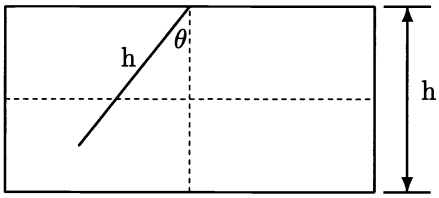


FIG. 8. *Schematic representation of the vertical cross-section for well configuration* (C).

of the cross-section of the domain D_2 is chosen so that the remaining area is equal to the area of the cross-section of domain D_1; i.e., the dimensionless radius associated with the horizontal cross-section of D_2 is $R_D = 1000$.

Two well configurations were considered for both reservoir models. For domain D_2, the direction of any considered well was such that its projection on the top of the reservoir corresponded to the schematic configuration shown in Figure 7. A well is modeled by a circular cylinder with the dimensionless radius $r_w = 1$. Then for both domains D_1 and D_2, the cross-section by the plane containing the well is a rectangle. Figures 8 and 9 show such cross-sections for every well configuration considered in the computational experiments. In configuration (E), the center of symmetry of the well coincides with the center of symmetry of the cross-section. In configuration (C), the well is drilled from the middle of the top side of the reservoir cross-section.

7.1. Directionally drilled wells. Effect of vertical flow. Productivity indices for well configuration (C) for domains D_2 and D_1 are given in Tables 2 and 3, respectively. In all cases, the penetration length of the well is equal to h so that for $\theta = 0$, the vertical well fully penetrates the reservoir. The graphs of J_{I} and J_{II} as functions of the angle θ of the well direction, shown in Figures 10 and 11, reveal that the optimal direction of a well of the fixed penetration length is not the vertical one. It is a clear indication of the effect of the vertical flow of fluid from the bottom of the reservoir toward the slanted well. This effect cannot be quantified by a reduced two-dimensional problem for a fully penetrated vertical well.

7.2. Horizontal well. Methods presented in [20, 11] rely heavily on the assumption that the vertical dimension of the reservoir is small compared to the penetration

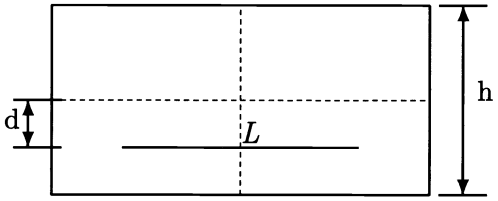


FIG. 9. *Schematic representation of the vertical cross-section for well configuration* (E).

TABLE 2
PIs for domains D_2, well configuration (C).

θ	0	15	30	45	60	75
J_{I}	0.1597	0.1714	0.1673	0.1634	0.1586	0.1529
J_{II}	0.1587	0.1704	0.1662	0.1623	0.1576	0.1520
$\left\lvert\frac{J_I - J_{\mathrm{II}}}{J_{\mathrm{II}}}\right\rvert$, percent	0.60	0.64	0.64	0.67	0.61	0.59

TABLE 3
PIs for domain D_1, well configuration (C).

	θ	0	8	15	30	45	60	75
$h = 100$	J_{I}	0.1629	0.1705	0.1765	0.1718	0.1691	0.1680	0.1662
	J_{II}	0.1623	0.1696	0.1758	0.1710	0.1683	0.1672	0.1655
	$\left\lvert\frac{J_I - J_{\mathrm{II}}}{J_{\mathrm{II}}}\right\rvert$, percent	0.36	0.50	0.37	0.50	0.48	0.46	0.47
$h = 200$	J_{I}	0.1629	0.1665	0.1697	0.1611	0.1426	0.1315	0.1199
	J_{II}	0.1623	0.1658	0.1689	0.1605	0.1422	0.1312	0.1196
	$\left\lvert\frac{J_I - J_{\mathrm{II}}}{J_{\mathrm{II}}}\right\rvert$, percent	0.36	0.41	0.43	0.38	0.30	0.27	0.28

length of the well. Moreover, as noted in [20], the precision of the evaluation of the PI for horizontal wells decreases drastically as the distance from the well to vertical boundaries of the reservoir becomes comparable to the distance to the top and/or the bottom of the reservoir, if the reduction to the two-dimensional problem is used. This section presents computational results for such settings when the assumption of the small reservoir thickness and the well being clearly inside the drainage area are relaxed.

The setting considered is a horizontal well with configuration (E), located at distance d below the plane of symmetry of domain D_1. The graphs of the computed PSS PI J_{I} as a function of distance d from the center of the reservoir for various penetration lengths L are shown in Figure 12.

For all practical purposes, one can conclude that the optimal location of a horizontal well in a cylindrical reservoir D_1 is in the horizontal plane of symmetry of the reservoir. Note that for long wells, however, the PSS PI slightly increases for small values of d. This may be an indication of an interesting feature of the diffusive capacity as a geometrical characteristic defined through the first eigenvalue λ_0. The latter is sensitive to the location of the well relative to the planes and lines of symmetry of the domain, as it is comprehensively illustrated in section 6. In three-dimensional

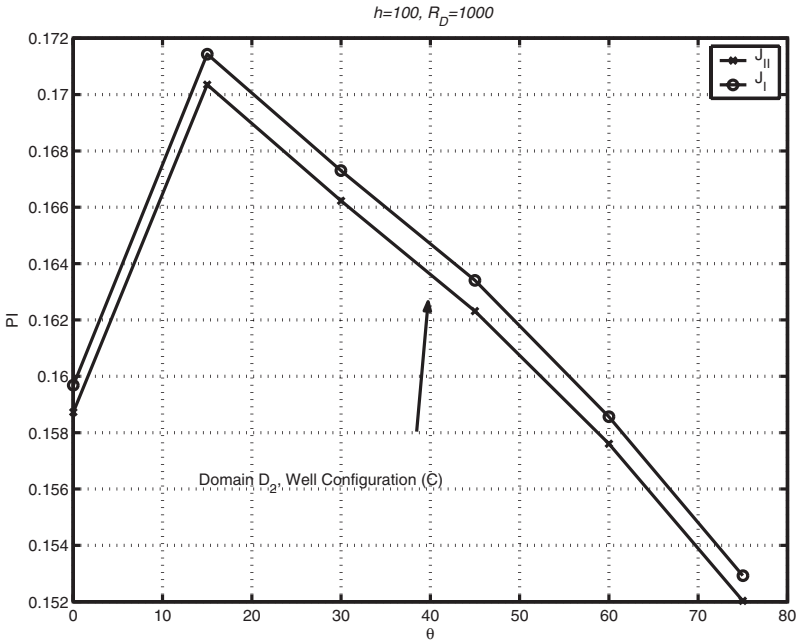


FIG. 10. *PIs for domain D_2, well configuration* (C).

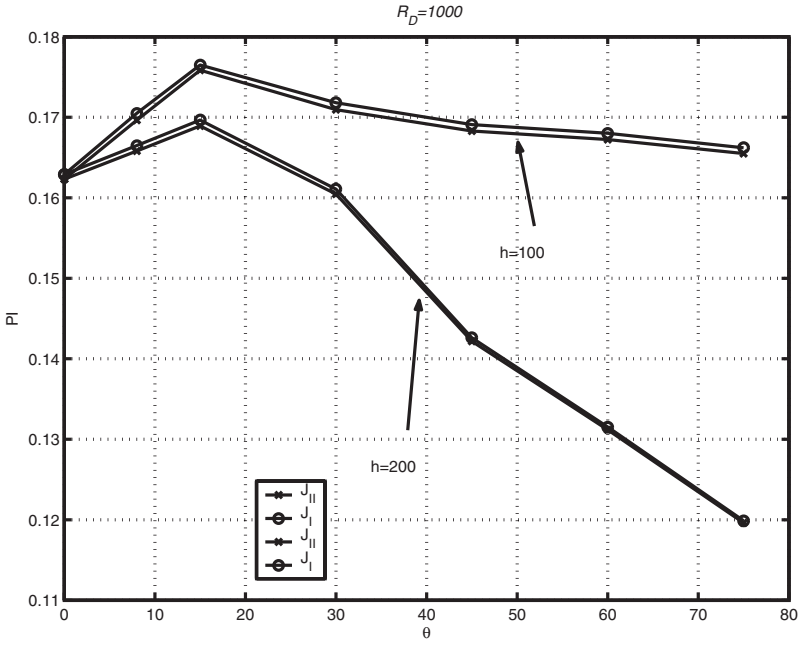


FIG. 11. *PIs for domain D_1, well configuration* (C).

domains, there are more such planes and lines of symmetry and, therefore, there may be several well configurations yielding maximal PI.

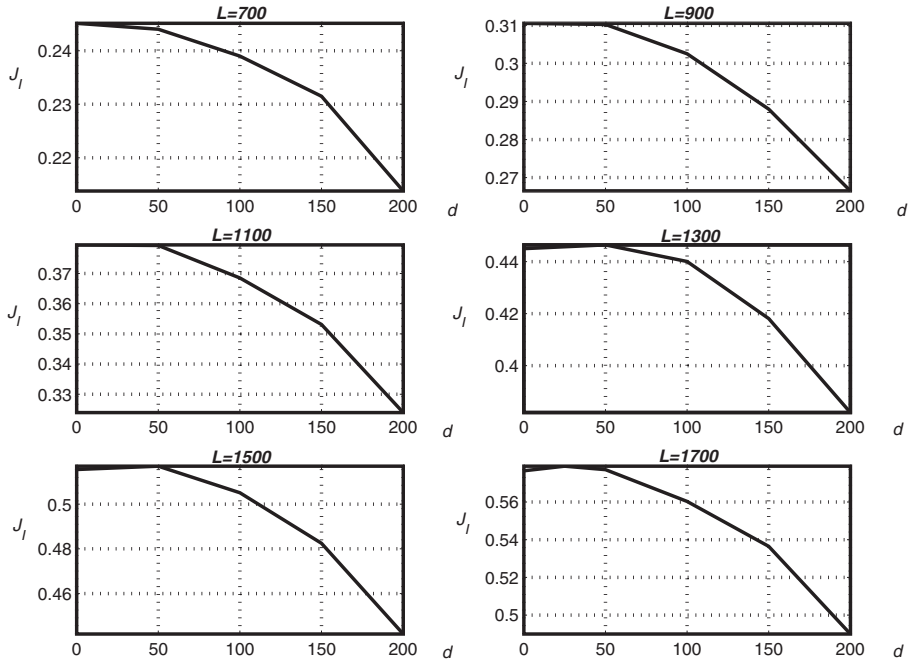


FIG. 12. *PSS PI for various values of d and L, well configuration* (E), $h = 500$.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] T. A. BLASINGAME, L. E. DOUBLET, AND P. P. VALKO, *Development and application of the multiwell productivity index (MPI)*, SPE J., 5 (2000), pp. 21–31.

[3] H. CINCO-LEY, H. J. RAMEY, JR., AND F. G. MILLER, *Pseudo-skin factors for partially penetrating directionally drilled wells*, in Proceedings of the 50th Annual Fall Meeting, SPE 5589, Dallas, TX, 1975.

[4] H. CINCO, F. G. MILLER, AND H. J. RAMEY, JR., *Unsteady-state pressure distribution created by a directionally drilled well*, J. Petroleum Technology, Nov. 1975, pp. 1392–1400.

[5] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics, Vol.* 1, Interscience Publishers, New York, 1953.

[6] D. N. DIETZ, *Determination of average reservoir pressure from build-up surveys*, J. Petroleum Technology, Aug. 1965, pp. 955–959.

[7] R. C. EARLOUGHER, JR., H. J. RAMEY, JR., F. G. MILLER, AND T. D. MUELLER, *Pressure distributions in rectangular reservoirs*, J. Petroleum Technology, Feb. 1968, pp. 200–208.

[8] M. J. FETKOVICH, *The isochronal testing of oil wells*, in Proceedings of the 48th Annual Fall Meeting, SPE 4529, Las Vegas, NV, 1973.

[9] M. J. FETKOVICH, E. J. FETKOVICH, AND M. D. FETKOVICH, *Useful concepts for decline—curve forecasting, reserve estimation and analysis*, SPE Reservoir Engrg., Feb. 1996, pp. 13–22.

[10] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1964.

[11] P. A. GOODE AND F. J. KUCHUK, *Inflow performance of horizontal wells*, SPE Reservoir Engrg., Aug. 1991, pp. 319–323.

[12] M. F. HAWKINS, JR., *A note on the skin effect*, Petr. Trans. AIME, 207 (1956), pp. 356–357.

[13] W. HELMY AND R. A. WATTENBARGER, *New shape factors for wells produced at constant pressure*, in Proceedings of the SPE Gas Technology Symposium, SPE 39970, Calgary, Canada, 1998.

[14] W. HURST, *Establishment of skin effect and its impediment to fluid flow into a well bore*, Petroleum Engineer, 25 (1953), pp. B-6–B-16.

[15] W. Hurst, J. D. Clark, and E. B. Brauer, *The skin effect in producing wells*, J. Petroleum Technology, Nov. 1969, pp. 1483–1489.
[16] A. Ibragimov and P. Valko, *On productivity index in pseudo-steady and boundary dominated flow regimes*, Technical report, Institute of Scientific Computations, 2000; also available online from http://www.isc.tamu.edu/iscpubs/iscreports.html.
[17] A. I. Ibragimov and M. N. Baganova, *Study of transient flow filtration towards a single horizontal well*, in Fundamental Bases of New Technologies in Oil and Gas Industry, Nauka, Moscow, 2000, pp. 192–198.
[18] A. I. Ibragimov and E. M. Landis, *On the behavior of the solution of the Zaremba problem in the neighborhood of the boundary point and at the infinity*, Dokl. Math., 57 (1998), pp. 185–186.
[19] R. F. Krueger, *An overview of formation damage and well productivity in oilfield operations*, J. Petroleum Technology, Feb. 1986, pp. 131–152.
[20] L. Larsen, *General productivity models for wells in homogeneous and layered reservoirs*, in Proceedings of the SPE Annual Conference and Exhibition, SPE 71613, New Orleans, LA, 2001.
[21] C. S. Matthews, F. Brons, and P. Hazebroek, *A method for determination of average pressure in a bounded reservoir*, Trans. AIME, 201 (1954), pp. 182–191.
[22] V. G. Maz'ya, *Differentiable Functions on Bad Domains*, World Scientific, Singapore, 1997.
[23] M. Muskat, *The Flow of Homogeneous Fluid Through Porous Media*, McGraw–Hill, New York, 1937.
[24] L. E. Payne and I. Stakgold, *On the mean value of the fundamental mode in the fixed membrane problem*, Appl. Anal., 3 (1973), pp. 225–303.
[25] J. K. Pucknell and P. J. Clifford, *Calculations of total skin factors*, in Proceedings of the Offshore Europe Conference, SPE 23100, Aberdeen, UK, 1991.
[26] R. Raghavan, *Well Test Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1991.
[27] H. J. Ramey, Jr., and W. M. Cobb, *A general pressure buildup theory for a well in a closed drainage area*, J. Petroleum Technology, Dec. 1971, pp. 1493–1505.
[28] D. J. Romero, P. P. Valko, and M. J. Economides, *Optimization of the productivity index and the fracture geometry of a stimulated well with fracture face and choke skins*, SPE Prod. Facilities, 18 (2003), pp. 57–63.
[29] T. Yildiz, *Assessment of total skin factors in perforated wells*, in Proceedings of the SPE European Formation Damage Conference, SPE 82249, Hague, The Netherlands, 2003.
[30] A. F. van Everdingen, *The skin effect and its influence on the productivity capacity of a well*, Petroleum Trans. AIME, 198 (1953), pp. 171–176.

SIAM J. Appl. Math.
Vol. 65, No. 6, pp. 1981–2004

WEBSTER'S HORN EQUATION REVISITED*

SJOERD W. RIENSTRA†

Abstract. The problem of low-frequency sound propagation in slowly varying ducts is systematically analyzed as a perturbation problem of slow variation. Webster's horn equation and variants in bent ducts, in ducts with nonuniform soundspeed, and in ducts with irrotational mean flow, with and without lining, are derived, and the entrance/exit plane boundary layer is given. It is shown why a varying lined duct in general does not have an (acoustic) solution.

Key words. Webster's horn equation, duct acoustics, sound propagation in lined ducts with mean flow, perturbation methods, method of slow variation

AMS subject classifications. 76Q05, 74J05, 35Q35, 35B20, 35P99

DOI. 10.1137/S0036139902413040

1. Introduction.

Sound of long wavelength, propagating in ducts of varying diameter like horns, is suitably described by an approximate equation, known as Webster's horn equation or just Webster's equation. This is an ordinary differential equation in the axial coordinate, and therefore forms a significant simplification of the problem [1, 2, 3].

The usual derivation is based on the assumption of a crosswise uniform acoustic pressure field, such that, by averaging over a duct cross section, the spatial dimensions of the problem are reduced from three to one.

Although it shows a remarkable evidence of ingenuity and physical insight, this derivation is mathematically unsatisfying. It is not clear (i) what exactly is the small parameter underlying the approximation, (ii) why the pressure may be assumed to be uniform, (iii) what the error is of the approximation, (iv) what the conditions are on the duct geometry and on the frequency of the field, (v) how to generalize to similar problems, (vi) how to generate higher order corrections, and (vii) what happens near the source or duct entrance or exit plane.

An asymptotically systematic derivation of the three-dimensional (3D) classic problem was given by Lesser and Crighton [4], extending the derivation of Lesser and Lewis in [5, 6]. They also showed for a number of 2D configurations how abrupt changes of the geometry (open end, slit in the wall) can be incorporated as boundary layer regions in a setting of matched asymptotic expansion. Their approach, based on introducing different longitudinal and lateral scales, is a special case of the method of slow variation put forward by Van Dyke [7]. Although only an asymptotically sound derivation is able to indicate the range of validity and the order of the error of the approximation, we found in the literature no variants of this problem (e.g., with mean flow [8, 9, 10, 11, 12]) that strictly follow that approach.

Particularly interesting would be an investigation of the related problems of lined ducts without and with flow, as this would form a natural long wavelength closure of the multiple scales theory of sound propagation in slowly varying ducts [13, 14, 15, 16].

*Received by the editors August 9, 2002; accepted for publication (in revised form) March 3, 2005; published electronically August 3, 2005.
http://www.siam.org/journals/siap/65-6/41304.html
†Department of Mathematics and Computing Science, Eindhoven University of Technology, Eindhoven, The Netherlands (s.w.rienstra@tue.nl).

Another problem of practical interest that is directly related to a systematic set-up is the entrance problem for a 3D duct of arbitrary cross section. The structure of the boundary layer was indicated by Lesser and Crighton [4], but they gave explicit examples only for 2D geometries.

All in all, while the problem of long wave sound propagation in slowly varying ducts, in various generalizations, is practically important, it still has a lot of open ends.

We will consider various cases in detail. First, we show how a systematic approach, known as the method of slow variation coupled with ideas of matched asymptotic expansions, leads to the classic Webster equation for hard-walled ducts with the entrance boundary layer. The small parameter ε is equal to the Helmholtz number, the ratio between a typical wavelength and the duct diameter, while a typical length scale of duct variation is of the same order of magnitude as the wavelength. Using similar results for the related problem of heat conduction [17], this entrance problem will be solved explicitly. It leads via matching conditions to conclusions about the way that the $\mathcal{O}(1)$ duct field error ($\mathcal{O}(\varepsilon)$ or $\mathcal{O}(\varepsilon^2)$) depends on the source.

Then we will show that our problem is not essentially different in other coordinate systems (like spherical coordinates), although special coordinates may be helpful in obtaining a more efficient approximation. Curved ducts, with a curvature radius of no more than the typical length scale of diameter variation, are shown to still produce the same equation.

The same type of analysis can be applied to ducts with lined walls of, say, impedance Z. It is found that for $Z = \mathcal{O}(1)$ only the trivial solution exists, while for $Z = \mathcal{O}(\varepsilon)$ there are only nontrivial solutions possible for certain geometry-dependent values of the wall impedance. As these impedance values vary along the duct, there are in general no solutions possible for the full duct. A subtle functional analytic result is used, due to Professor Jan de Graaf (TU Eindhoven), which is not available in the literature; therefore, Prof. de Graaf was kind enough to attach his derivation as an appendix to this paper.

We continue with more general analyses of the problem in a stagnant medium with slowly varying sound speed, and of sound in an irrotational isentropic mean flow, leading to generalized forms of Webster's horn equation.

We finish with the same problem with mean flow but now extended to ducts with lined walls. Using a recent result obtained for the related problem for high-frequency sound propagation in lined flow ducts [16], we are able to show for $Z = \mathcal{O}(1)$ that also here only a special hydrodynamic (nonacoustic) wave is possible.

2. The physical models.

2.1. The equations. In the acoustic realm of a perfect gas that we will consider, we have for pressure $\tilde{p}$, velocity $\tilde{\boldsymbol{v}}$, density $\tilde{\rho}$, entropy $\tilde{s}$, and soundspeed $\tilde{c}$

$$\begin{aligned} &\frac{\mathrm{d}\tilde{\rho}}{\mathrm{d}t} = -\tilde{\rho}\nabla\cdot\tilde{\boldsymbol{v}}, \quad \tilde{\rho}\frac{\mathrm{d}\tilde{\boldsymbol{v}}}{\mathrm{d}t} = -\nabla\tilde{p}, \quad \frac{\mathrm{d}\tilde{s}}{\mathrm{d}t} = 0, \\ &\mathrm{d}\tilde{s} = C_V\frac{\mathrm{d}\tilde{p}}{\tilde{p}} - C_P\frac{\mathrm{d}\tilde{\rho}}{\tilde{\rho}}, \quad \tilde{c}^2 = \frac{\gamma\tilde{p}}{\tilde{\rho}}, \quad \gamma = \frac{C_P}{C_V}, \end{aligned} \tag{1}$$

where γ, C_P, and C_V are gas constants. When the flow originates from a thermodynamically uniform state and consists of a stationary mean flow, with unsteady

time-harmonic perturbations of frequency ω given, in the usual complex notation, by

$$\tilde{\boldsymbol{v}} = \boldsymbol{V} + \mathrm{Re}(\boldsymbol{v}\,\mathrm{e}^{\mathrm{i}\omega t}), \quad \tilde{p} = P + \mathrm{Re}(p\,\mathrm{e}^{\mathrm{i}\omega t}), \quad \tilde{\rho} = D + \mathrm{Re}(\rho\,\mathrm{e}^{\mathrm{i}\omega t}), \quad \tilde{s} = S + \mathrm{Re}(s\,\mathrm{e}^{\mathrm{i}\omega t}) \tag{2}$$

($\omega > 0$), we obtain for the mean flow, upon linearization for small amplitude,

$$\begin{gathered} \nabla\cdot(D\boldsymbol{V}) = 0, \quad D(\boldsymbol{V}\cdot\nabla)\boldsymbol{V} = -\nabla P, \\ (\boldsymbol{V}\cdot\nabla)S = 0, \quad S = C_V \log P - C_P \log D, \quad C^2 = \frac{\gamma P}{D}, \end{gathered} \tag{3}$$

and for the perturbations

$$\mathrm{i}\omega\rho + \nabla\cdot(\boldsymbol{V}\rho + \boldsymbol{v}D) = 0, \tag{4a}$$

$$D\big(\mathrm{i}\omega + \boldsymbol{V}\cdot\nabla\big)\boldsymbol{v} + D\big(\boldsymbol{v}\cdot\nabla\big)\boldsymbol{V} + \rho(\boldsymbol{V}\cdot\nabla)\boldsymbol{V} = -\nabla p, \tag{4b}$$

$$(\mathrm{i}\omega + \boldsymbol{V}\cdot\nabla)s + \boldsymbol{v}\cdot\nabla S = 0, \tag{4c}$$

while

$$s = \frac{C_V}{P}p - \frac{C_P}{D}\rho = \frac{C_V}{P}\big(p - C^2\rho\big). \tag{4d}$$

Without mean flow, such that $\boldsymbol{V} = \nabla P = 0$, the equations may be reduced to (see section 8)

$$\nabla\cdot\big(C^2\nabla p\big) + \omega^2 p = 0. \tag{5}$$

If, in addition, the ambient medium is uniform, with a constant soundspeed C and density D, the acoustic field becomes isentropic and irrotational, and we may introduce a potential $\boldsymbol{v} = \nabla\phi$. Furthermore, (5) reduces to the Helmholtz equation. After introducing the free field wave number $k = \omega/C$, we have (see sections 3, 4, 6, 7)

$$\nabla^2\phi + k^2\phi = 0. \tag{6}$$

If the original flow field $\tilde{\boldsymbol{v}}$ is irrotational and isentropic everywhere (homentropic), we can introduce a potential for the velocity, where $\tilde{\boldsymbol{v}} = \nabla\tilde{\phi}$, and express $\tilde{p}$ as a function of $\tilde{\rho}$ only, such that we can integrate the momentum equation (Bernoulli's law, with constant E) to obtain for the mean flow

$$\tfrac{1}{2}V^2 + \frac{C^2}{\gamma - 1} = E, \quad \nabla\cdot(D\boldsymbol{V}) = 0, \quad \frac{P}{D^\gamma} = \text{constant}, \tag{7}$$

and for the acoustic perturbations

$$\big(\mathrm{i}\omega + \boldsymbol{V}\cdot\nabla\big)\rho + \rho\nabla\cdot\boldsymbol{V} + \nabla\cdot\big(D\nabla\phi\big) = 0, \quad D\big(\mathrm{i}\omega + \boldsymbol{V}\cdot\nabla\big)\phi + p = 0, \quad p = C^2\rho. \tag{8}$$

These last equations are further simplified (eliminate p and ρ and use the fact that $\nabla\cdot(D\boldsymbol{V}) = 0$) to the rather general convected wave equation (see section 9)

$$D^{-1}\nabla\cdot\big(D\nabla\phi\big) - \big(\mathrm{i}\omega + \boldsymbol{V}\cdot\nabla\big)\Big[C^{-2}\big(\mathrm{i}\omega + \boldsymbol{V}\cdot\nabla\big)\phi\Big] = 0. \tag{9}$$

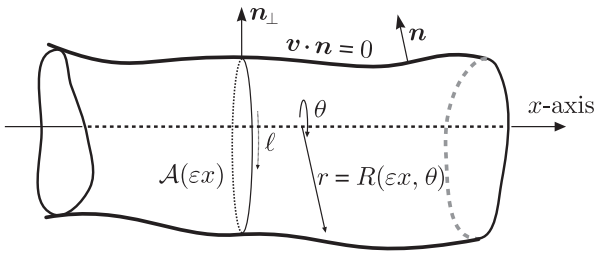


FIG. 1. *Sketch of geometry.*

2.2. Nondimensionalization. Without further change of notation, we will assume throughout this paper that the problem is made dimensionless: lengths on a typical duct radius, time on typical sound speed / typical duct radius, etc.

2.3. The geometry. The domain of interest consists of a duct $\mathcal{V}$ of arbitrary cross section, slowly varying in axial direction (see Figure 1). For definiteness, it is given by the function Σ in cylindrical coordinates, as follows:

$$\Sigma(X,r,\theta)=r-R(X,\theta)\leq 0, \tag{10}$$

where $X=\varepsilon x\geqslant 0$ is a so-called slow variable, while ε is small. A cross section $\mathcal{A}(X)$ at axial position X has surface area $A(X)$. Whenever relevant,[1] we assume lengths made dimensionless such that

$$A(0)=1.$$

At the duct surface $\Sigma=0$, the gradient $\nabla\Sigma$ is a vector normal to the surface (i.e., $\nabla\Sigma\propto\boldsymbol{n}$), while the transverse gradient $\nabla_\perp\Sigma$,

$$\nabla_\perp=\boldsymbol{e}_r\frac{\partial}{\partial r}+\boldsymbol{e}_\theta\frac{1}{r}\frac{\partial}{\partial\theta},\quad\text{with}\quad\nabla_\perp\Sigma=\boldsymbol{e}_r-\boldsymbol{e}_\theta\frac{1}{r}R_\theta, \tag{11}$$

(where R_θ denotes the partial derivative of R to θ) is directed in the plane of a cross section $\mathcal{A}(X)$ and normal to the duct circumference $\partial\mathcal{A}$. Thus if $\boldsymbol{n}_\perp$ is the component of the surface normal vector $\boldsymbol{n}$ in the plane of a cross section, we have $\nabla_\perp\Sigma\propto\boldsymbol{n}_\perp$.

2.4. Frequency. The frequencies considered are low, such that the corresponding typical wave number is of the same order of magnitude as the length scale of the duct variations, i.e., dimensionless $\mathcal{O}(\varepsilon^{-1})$. In order to quantify this, we will rescale $k=\varepsilon\kappa$ and $\omega=\varepsilon\Omega$.

3. The classical problem.

3.1. Equations and boundary conditions. The duct is semi-infinite and hard-walled. The solution is determined by a source at entrance plane $x=0$, and radiation conditions for $x\to\infty$. Other conditions, like a reflecting impedance plane at some exit plane $x=L$ (e.g., modeling a radiating open end [5] or a slit in the wall [4]), are also possible, but they do not essentially alter the present analysis.

Inside $\mathcal{V}$ we have for acoustic potential ϕ (see (6))

$$\nabla^2\phi+\varepsilon^2\kappa^2\phi=0\ \text{ if }\ \boldsymbol{x}\in\mathcal{V},\quad\text{with}\quad\nabla\phi\cdot\boldsymbol{n}=0\ \text{ at }\ \boldsymbol{x}\in\partial\mathcal{V}. \tag{12}$$

[1] In particular, in section 4.

At the entrance interface $x = 0$ we have a suitable boundary condition, say,

$$\phi(0, r, \theta) = F(r, \theta). \tag{13}$$

The boundary condition of hard walls at $r = R(X, \theta)$ may be given by

$$\nabla_\perp \phi \cdot \nabla_\perp \Sigma = \phi_r - \frac{R_\theta}{R^2}\phi_\theta = \varepsilon R_X \phi_x. \tag{14}$$

Except for the immediate neighborhood of the entrance plane, the typical axial variations of the acoustic field scale on the slow variable X, so we rewrite the equations and boundary conditions as

$$\begin{gathered} \varepsilon^2 \phi_{XX} + \nabla_\perp^2 \phi + \varepsilon^2 \kappa^2 \phi = 0, \\ \text{with} \quad \nabla\phi \cdot \nabla\Sigma = -\varepsilon^2 \phi_X R_X + \nabla_\perp \phi \cdot \nabla_\perp \Sigma = 0 \quad \text{at} \quad r = R. \end{gathered} \tag{15}$$

This rewriting in a slow variable is known as the method of slow variation [7]. Note that this equation has a small parameter multiplied by the highest derivative in the X-direction, suggesting a singular perturbation problem [4, 18, 19, 20] with boundary layers in X.

3.2. Asymptotic analysis: Outer solution. The following outer solution analysis will largely follow Lesser and Crighton [4], but we will give it in some detail for two reasons. First, we will have to define the solution for the inner solution at the entrance boundary layer to be discussed later. Second, it explicates the method of integration along a cross section that will be used in the various other configurations later.

Based on the observation that ε^2 is the only small parameter that occurs, we might be tempted to expand the solution in a Poincaré asymptotic power series in ε^2. However, we will see that this is not exactly true. Depending on the behavior of the solution near the entrance, the correction term should in general be $\mathcal{O}(\varepsilon)$ for matching. Nevertheless, the leading and first order equations will be equivalent. With the assumed Poincaré expansion of ϕ, expressed in X,

$$\phi(X, r, \theta; \varepsilon) = \phi_0(X, r, \theta) + \varepsilon\phi_1(X, r, \theta) + \varepsilon^2\phi_2(X, r, \theta) + \cdots, \tag{16}$$

we obtain to leading order

$$\nabla_\perp^2 \phi_0 = 0, \quad \text{with} \quad \nabla_\perp \phi_0 \cdot \boldsymbol{n}_\perp = 0 \quad \text{at} \quad r = R, \tag{17}$$

with a solution $\phi_0 = 0$. As the solution of a Neumann problem is unique up to a constant, $\phi_0 = \phi_0(X)$, a function to be determined. To first order we have

$$\nabla_\perp^2 \phi_1 = 0, \quad \text{with} \quad \nabla_\perp \phi_1 \cdot \boldsymbol{n}_\perp = 0 \quad \text{at} \quad r = R, \tag{18}$$

also with a constant solution, and so $\phi_1 = \phi_1(X)$, a function to be determined. To second order we now have

$$\nabla_\perp^2 \phi_2 + \phi_{0,XX} + \kappa^2 \phi_0 = 0, \qquad \text{with} \quad \nabla_\perp \phi_2 \cdot \boldsymbol{n}_\perp = \phi_{0,X} \frac{R R_X}{\sqrt{R^2 + R_\theta^2}} \quad \text{at} \quad r = R. \tag{19}$$

The assumption (16) that there exists a Poincaré expansion for ϕ, expressed in this slow variable X, is not trivial. (Poincaré expansions are critically dependent on the

variables chosen!) It requires certain solvability conditions for, e.g., ϕ_2, yielding an equation for ϕ_0. To obtain this, we integrate along a cross section $\mathcal{A}(X)$ and apply Gauss' theorem

$$\iint_{\mathcal{A}} \nabla_\perp^2 \phi_2 \,\mathrm{d}\sigma = \int_{\partial\mathcal{A}} \nabla_\perp \phi_2 \cdot \boldsymbol{n}_\perp \,\mathrm{d}\ell = \int_{\partial\mathcal{A}} \phi_{0,X} \frac{RR_X}{\sqrt{R^2+R_\theta^2}} \,\mathrm{d}\ell = \cdots .$$

Then we parametrize $\partial\mathcal{A}$ with θ such that $\mathrm{d}\ell = \sqrt{R^2+R_\theta^2}\,\mathrm{d}\theta$, and we continue

$$= \int_0^{2\pi} \phi_{0,X} RR_X \,\mathrm{d}\theta = \phi_{0,X} \int_0^{2\pi} RR_X \,\mathrm{d}\theta = \phi_{0,X} A_X. \tag{20}$$

On the other hand, we also have

$$\iint_{\mathcal{A}} \big[\phi_{0,XX} + \kappa^2\phi_0\big] \,\mathrm{d}\sigma = A\big(\phi_{0,XX} + \kappa^2\phi_0\big). \tag{21}$$

Altogether we have for ϕ_0 the equation

$$A^{-1}\big(A\phi_{0,X}\big)_X + \kappa^2\phi_0 = 0, \tag{22}$$

which is indeed Webster's horn equation [1, 2] in properly scaled coordinates.

Evidently, the first order solution follows the same pattern and also satisfies

$$A^{-1}\big(A\phi_{1,X}\big)_X + \kappa^2\phi_1 = 0. \tag{23}$$

For completeness we note from [21, 22, 23, 24, 3] that Webster's equation can be recast into a more transparent form by the transformation

$$A(X) = d(X)^2, \quad \phi = d^{-1}\psi, \tag{24}$$

leading to

$$\psi'' + \Big(\kappa^2 - \frac{d''}{d}\Big)\psi = 0. \tag{25}$$

Depending on the sign of $\kappa^2 - d''/d$, the solutions behave like propagating or exponentially decaying waves. Elementary solutions are readily found for geometries with $d''/d = m^2$, a constant, yielding Salmon's family of exponential and conical horns [21, 22].

3.3. Boundary conditions in X. The above equation for ϕ_0 and ϕ_1 is of second order, and therefore two boundary conditions are required to determine the solution. For $X \to \infty$ we have the condition of radiation. At $X = 0$ (Figure 2), ϕ_0 and ϕ_1 cannot satisfy the (r,θ)-dependent boundary condition (13). Indeed, as anticipated before, near $x = 0$ there is a boundary layer of $X = \mathcal{O}(\varepsilon)$, i.e., $x = \mathcal{O}(1)$, which determines the (outer) solutions ϕ_0 and ϕ_1 via conditions of matching. This will be considered in the next section.

4. Entrance boundary layer. Near the entrance, for $X = \mathcal{O}(\varepsilon)$, i.e., $x = \mathcal{O}(1)$, we have of course equation (12)

$$\nabla^2\phi + \varepsilon^2\kappa^2\phi = 0 \ \text{ if } \ \boldsymbol{x} \in \mathcal{V}, \qquad \text{with } \ \nabla_\perp\phi \cdot \boldsymbol{n} = 0 \ \text{ at } \ \boldsymbol{x} \in \partial\mathcal{V}. \tag{12}$$

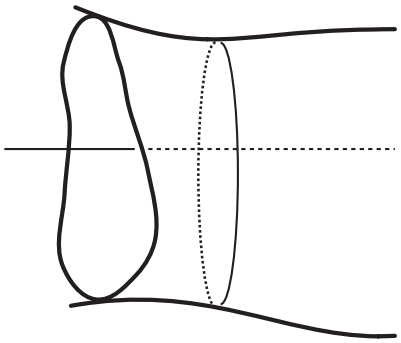


FIG. 2. *The entrance.*

Up to $\mathcal{O}(\varepsilon^2)$, this Helmholtz equation is equivalent to the Laplace equation. Therefore, the boundary layer analysis is essentially similar to that for the heat equation, discussed in Chandra [17]. Expand

$$\phi(X, r, \theta; \varepsilon) = \Phi_0(x, r, \theta) + \varepsilon\Phi_1(x, r, \theta) + \mathcal{O}(\varepsilon^2) \tag{26}$$

so that we have inside $\mathcal{V}$ to leading and first order,

$$\mathcal{O}(1): \quad \nabla^2\Phi_0 = 0, \tag{27a}$$

$$\mathcal{O}(\varepsilon): \quad \nabla^2\Phi_1 = 0. \tag{27b}$$

At $x = 0$ we have from (13) the initial conditions

$$\Phi_0(0, r, \theta) = F(r, \theta), \quad \Phi_1(0, r, \theta) = 0. \tag{28}$$

For $x \to \infty$ conditions of matching with the outer solution $\phi_0 + \varepsilon\phi_1$ apply. For the boundary condition at $r = R$ we have to expand $R(\varepsilon x, \theta)$. Note that for any function f

$$f(R(\varepsilon x); \varepsilon) = f(R + \varepsilon x R_X + \mathcal{O}(\varepsilon^2); \varepsilon) = f_0(R) + \varepsilon\big(f_1(R) + x f_{0,r}(R) R_X\big) + \mathcal{O}(\varepsilon^2), \tag{29}$$

where R without any argument denotes the value at $X = 0$. Furthermore, we have

$$\frac{R_\theta(X, \theta)}{R^2(X, \theta)} = \frac{R_\theta}{R^2} + \varepsilon x\Big(\frac{R_X}{R^2}\Big)_\theta + \mathcal{O}(\varepsilon^2). \tag{30}$$

Thus at the boundary

$$\begin{aligned}\nabla_\perp\phi \cdot \nabla_\perp\Sigma &= \phi_r - \frac{R_\theta}{R^2}\phi_\theta \\ &= \Phi_{0,r} - \frac{R_\theta}{R^2}\Phi_{0,\theta} + \varepsilon\left[\Phi_{1,r} - \frac{R_\theta}{R^2}\Phi_{1,\theta} + x\Phi_{0,rr}R_X - x\frac{R_\theta}{R^2}R_X\Phi_{0,r\theta} - x\Big(\frac{R_X}{R^2}\Big)_\theta\Phi_{0,\theta}\right] \\ &= \varepsilon R_X\Phi_{0,x},\end{aligned} \tag{31}$$

which means that at $r = R(0,\theta)$ for the leading and first order,

$$\nabla_\perp \Phi_0 \cdot \nabla_\perp \Sigma_0 = \Phi_{0,r} - \frac{R_\theta}{R^2}\Phi_{0,\theta} = 0, \tag{32a}$$

$$\begin{aligned}\nabla_\perp \Phi_1 \cdot \nabla_\perp \Sigma_0 &= \Phi_{1,r} - \frac{R_\theta}{R^2}\Phi_{1,\theta} \\ &= R_X\Phi_{0,x} - x\Phi_{0,rr}R_X + x\frac{R_\theta}{R^2}R_X\Phi_{0,r\theta} + x\Big(\frac{R_X}{R^2}\Big)_\theta\Phi_{0,\theta},\end{aligned} \tag{32b}$$

where $\Sigma_0 = \Sigma(0,r,\theta)$.

It is important for the subsequent matching to note that the solutions of (27) with (32) are defined only up to a linear term Kx. For Φ_0, however, this would result in terms of $\mathcal{O}(\varepsilon^{-1})$ if $x = \mathcal{O}(\varepsilon^{-1})$, which do not match with an outer solution $\phi_0 = \mathcal{O}(1)$. Therefore, we will not include this extra term. For Φ_1, on the other hand, we will have to retain the possibility, and in the end a linear term $K_1 x$ will be added, where K_1 must be determined by the matching.

From the identity at $r = R$,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\Phi_{0,\theta} = \Phi_{0,r\theta}R_\theta + \Phi_{0,\theta\theta}, \tag{33}$$

and with the defining equation applied at $r = R$ while using relation (32a),

$$-\Phi_{0,rr} = \frac{1}{R}\Phi_{0,r} + \frac{1}{R^2}\Phi_{0,\theta\theta} + \Phi_{0,xx} = \frac{R_\theta}{R^3}\Phi_{0,\theta} + \frac{1}{R^2}\Phi_{0,\theta\theta} + \Phi_{0,xx}, \tag{34}$$

it follows that (32b) is equivalent to

$$\begin{aligned}\nabla_\perp \Phi_1 \cdot \nabla_\perp \Sigma_0 &= \mathcal{Q}_0(x,\theta) \\ &\overset{\text{def}}{=\!=} R_X\Phi_{0,x}\big|_{r=R} + \frac{x}{R}\left\{RR_X\Phi_{0,xx}\big|_{r=R} + \frac{\mathrm{d}}{\mathrm{d}\theta}\Big(\frac{R_X}{R}\Phi_{0,\theta}\big|_{r=R}\Big)\right\}.\end{aligned} \tag{35}$$

4.1. Leading order. The right-running solution Φ_0 (only nonincreasing exponentials are allowed for matching) may be expressed by the eigenfunction expansion

$$\Phi_0(\boldsymbol{x}) = \sum_{n=0}^{\infty} F_n\psi_n(r,\theta)\,\mathrm{e}^{-\lambda_n x}, \tag{36}$$

where

$$\nabla_\perp^2\psi_n + \lambda_n^2\psi_n = 0, \quad \nabla_\perp\psi_n \cdot \nabla_\perp\Sigma_0 = 0, \tag{37}$$

with $\lambda_0 = 0$, ψ_0 a constant (normalized to 1), the other eigenvalues[2] λ_n real positive, and the eigenfunctions ψ_n real, orthogonal, and assumed normalized. In general these eigenfunctions are to be determined numerically. However, if the duct is cylindrical (i.e., R is independent of θ), we have

$$\psi_n(r,\theta) := \psi_{\nu\mu}(r,\theta) = \begin{cases} \dfrac{J_\nu\big(j'_{\nu\mu}r/R\big)}{\sqrt{\frac{\pi}{2}\Big(1-\frac{\nu^2}{{j'_{\nu\mu}}^2}\Big)}RJ_\nu(j'_{\nu\mu})}\begin{Bmatrix}\cos\nu\theta\\ \sin\nu\theta\end{Bmatrix} & \text{for } \nu \neq 0, \\ \dfrac{J_0\big(j'_{0\mu}r/R\big)}{\sqrt{\pi}RJ_0(j'_{0\mu})} & \text{for } \nu = 0, \end{cases} \tag{38}$$

[2]Strictly speaking, the numbers $-\lambda_n^2$ are the eigenvalues of operator $\nabla_\perp^2$.

where the index n is more practically changed into the double index $(\nu\mu)$. J_ν is the νth order ordinary Bessel function of the first kind [25], and $j'_{\nu\mu}$ is the μth (real-valued, positive) zero of J'_ν. The corresponding eigenvalue is thus $\lambda_n := j'_{\nu\mu}/R$.

The amplitudes are determined from the entrance interface $x = 0$ as follows:

$$F_n = \iint_{\mathcal{A}(0)} F(r,\theta)\psi_n(r,\theta)\,\mathrm{d}\sigma. \tag{39}$$

Note that, as ψ_n are orthonormal, the axial flux is, to leading order, proportional to the imaginary part of

$$\int_0^{2\pi}\int_0^R \Phi_0\Phi^*_{0,x} r\,\mathrm{d}r\mathrm{d}\theta = -\sum_{n=1}^{\infty}\lambda_n|F_n|^2\,\mathrm{e}^{-2\lambda_n x}\,. \tag{40}$$

As this expression is real, its imaginary part is zero, and thus the axial flux vanishes to leading order. Indeed, the outer solution is a slowly varying function of X, and therefore the flux, proportional to the axial derivative, is $\mathcal{O}(\varepsilon)$.

For $x \to \infty$, the exponential terms in $\Phi_0(\boldsymbol{x})$ vanish and we have

$$\Phi_0(\boldsymbol{x}) \simeq F_0. \tag{41}$$

4.2. First order. With the found expression for Φ_0, the right-hand side of (35), $\mathcal{Q}_0$, may be written as

$$\begin{aligned}\mathcal{Q}_0(x,\theta) &= \sum_{n=1}^{\infty} F_n\,\mathrm{e}^{-\lambda_n x}\left[-R_X\lambda_n\psi_n\big|_{r=R} + xR_X\lambda_n^2\psi_n\big|_{r=R} + \frac{x}{R}\frac{\mathrm{d}}{\mathrm{d}\theta}\Big(\frac{R_X}{R}\psi_{n,\theta}\big|_{r=R}\Big)\right]\\ &= R^{-1}\sum_{n=1}^{\infty} F_n\left[-\lambda_n RR_X\big(x\,\mathrm{e}^{-\lambda_n x}\big)_x\psi_n\big|_{r=R} + x\,\mathrm{e}^{-\lambda_n x}\,\frac{\mathrm{d}}{\mathrm{d}\theta}\Big(\frac{R_X}{R}\psi_{n,\theta}\big|_{r=R}\Big)\right].\end{aligned} \tag{42}$$

To solve the problem for Φ_1, we introduce a Green's function $G(\boldsymbol{x};\boldsymbol{\xi})$ with $\boldsymbol{x} = (x, r, \theta)$ and $\boldsymbol{\xi} = (\xi, \rho, \eta)$ satisfying

$$\begin{gathered}\nabla_\perp^2 G + \frac{\partial^2}{\partial x^2}G = -\delta(\boldsymbol{x}-\boldsymbol{\xi}), \quad \frac{\partial}{\partial n}G = 0 \ \text{ at } \ r = R(0,\theta), \quad G(\boldsymbol{x};\boldsymbol{\xi}) = 0 \ \text{ at } \ x = 0,\\ G(\boldsymbol{x};\boldsymbol{\xi}) \to \text{a constant} \ \text{ for } \ x\to\infty, \qquad x\frac{\partial}{\partial x}G(\boldsymbol{x};\boldsymbol{\xi}) \to\ 0 \ \text{ for } \ x\to\infty.\end{gathered} \tag{43}$$

We determine the Green's function by applying the Fourier sine transform[3] with respect to x $(x \to \alpha)$ to (43), to obtain

$$\nabla_\perp^2\hat{G} - \alpha^2\hat{G} = -\sqrt{\frac{2}{\pi}}\sin(\alpha\xi)\delta(\boldsymbol{x}_\perp - \boldsymbol{\xi}_\perp), \tag{44}$$

where $\boldsymbol{x}_\perp$ denotes the transverse component of $\boldsymbol{x}$, i.e., $\boldsymbol{x}_\perp = (r,\theta)$ (similarly for $\boldsymbol{\xi}_\perp$). We assume that the Green's function can be expanded by the same basis function as has been used for Φ_0,

$$\hat{G}(\alpha, r, \theta; \boldsymbol{\xi}) = \sum_{m=0}^{\infty} a_m(\alpha,\boldsymbol{\xi})\psi_m(r,\theta).$$

[3]Here $\hat{f}(\alpha) = \sqrt{\frac{2}{\pi}}\int_0^\infty \sin(\alpha x)f(x)\,\mathrm{d}x$, $f(x) = \sqrt{\frac{2}{\pi}}\int_0^\infty \sin(\alpha x)\hat{f}(\alpha)\,\mathrm{d}\alpha$.

Therefore

$$\nabla^2\hat{G} = -\sum_{m=0}^{\infty} a_m\lambda_m^2\psi_m(r,\theta).$$

Substituting this into (44) yields

$$\sum_{m=0}^{\infty} a_m\psi_m(\lambda_m^2+\alpha^2) = \sqrt{\frac{2}{\pi}}\sin(\alpha\xi)\delta(\boldsymbol{x}_\perp-\boldsymbol{\xi}_\perp). \tag{45}$$

Next, we multiply (45) with ψ_n and integrate over the cross section $\mathcal{A}(0)$ to obtain

$$\iint\limits_{\mathcal{A}(0)}\sum_{m=0}^{\infty} a_m\psi_n\psi_m(\lambda_m^2+\alpha^2)\,\mathrm{d}\sigma = \sqrt{\frac{2}{\pi}}\iint\limits_{\mathcal{A}(0)}\psi_n(r,\theta)\sin(\alpha\xi)\delta(\boldsymbol{x}_\perp-\boldsymbol{\xi}_\perp)\,\mathrm{d}\sigma. \tag{46}$$

Orthonormality of the basis functions yields

$$a_m = \sqrt{\frac{2}{\pi}}\left(\frac{\sin(\alpha\xi)}{\lambda_m^2+\alpha^2}\right)\psi_m(\rho,\eta). \tag{47}$$

Therefore,

$$\hat{G}(\alpha,r,\theta;\xi,\rho,\eta) = \sqrt{\frac{2}{\pi}}\sum_{m=0}^{\infty}\frac{\sin(\alpha\xi)}{\lambda_m^2+\alpha^2}\psi_m(\rho,\eta)\psi_m(r,\theta). \tag{48}$$

The inverse Fourier sine transform yields

$$G(\boldsymbol{x};\boldsymbol{\xi}) = \frac{2}{\pi}\sum_{m=0}^{\infty}\psi_m(\rho,\eta)\psi_m(r,\theta)\int_0^{\infty}\frac{\sin(\alpha x)\sin(\alpha\xi)}{\lambda_m^2+\alpha^2}\,\mathrm{d}\alpha, \tag{49}$$

where [25] for $\lambda_0 = 0$,

$$\int_0^{\infty}\frac{\sin(\alpha x)\sin(\alpha\xi)}{\alpha^2}\,\mathrm{d}\alpha = \frac{1}{2}\pi\min(x,\xi), \tag{50}$$

and for $\lambda_m > 0$,

$$\int_0^{\infty}\frac{\sin(\alpha x)\sin(\alpha\xi)}{\lambda_m^2+\alpha^2}\,\mathrm{d}\alpha = \frac{1}{2}\pi\,\mathrm{e}^{-\lambda_m\max(x,\xi)}\,\frac{1}{\lambda_m}\sinh(\lambda_m\min(x,\xi)). \tag{51}$$

Therefore, the $m=0$ term can be taken apart, and the Green's function becomes

$$G(\boldsymbol{x};\boldsymbol{\xi}) = x + \sum_{m=1}^{\infty}\psi_m(\rho,\eta)\psi_m(r,\theta)\,\mathrm{e}^{-\lambda_m\xi}\,\frac{\sinh(\lambda_m x)}{\lambda_m} \qquad \text{if } 0\le x\le\xi \tag{52a}$$

$$= \xi + \sum_{m=1}^{\infty}\psi_m(\rho,\eta)\psi_m(r,\theta)\,\mathrm{e}^{-\lambda_m x}\,\frac{\sinh(\lambda_m\xi)}{\lambda_m} \qquad \text{if } 0\le\xi\le x. \tag{52b}$$

Note that as $x\to\infty$, G tends to ξ and $\frac{\partial G}{\partial x}$ tends to zero exponentially.

Using this Green's function, we obtain for Φ_1 the following relation, to be integrated over domain $\mathcal{V}$:

$$\Phi_1\delta(\boldsymbol{x}-\boldsymbol{\xi}) = G\nabla^2\Phi_1 - \Phi_1\nabla^2 G. \tag{53}$$

However, since $\Phi_1 \sim K_1\xi$ for large ξ (see the remark below (32)), this yields a divergent integral as the domain here is a semi-infinite duct. Therefore, we consider a region $\mathcal{V}'$ with a finite length $0 \le x \le x_0$, where x_0 is small compared to ε^{-1} but large enough for all exponential terms to practically vanish. Integrate (53) along domain $\mathcal{V}'$ and by using Green's second identity we get

$$\Phi_1(\boldsymbol{\xi}) = \iiint_{\mathcal{V}'} \left(G\nabla^2\Phi_1 - \Phi_1\nabla^2 G\right) \mathrm{d}\boldsymbol{x} = \iint_{x=0} \left(-G\frac{\partial\Phi_1}{\partial x} + \Phi_1\frac{\partial G}{\partial x}\right) \mathrm{d}\sigma + \iint_{r=R(0,\eta)} \left(G\nabla_\perp\Phi_1 - \Phi_1\nabla_\perp G\right)\cdot\boldsymbol{n}_\perp \,\mathrm{d}\sigma + \iint_{x=x_0} \left(G\frac{\partial\Phi_1}{\partial x} - \Phi_1\frac{\partial G}{\partial x}\right) \mathrm{d}\sigma$$

$$= \iint_{r=R(0,\eta)} \frac{G\mathcal{Q}_0(x,\theta)}{|\nabla_\perp\Sigma|}\,\mathrm{d}\ell\mathrm{d}\xi + K_1\xi. \tag{54}$$

Since $|\nabla_\perp\Sigma| = \frac{1}{R}\sqrt{R^2+R_\theta^2}$ and $\mathrm{d}\ell = \sqrt{R^2+R_\theta^2}\,\mathrm{d}\theta$, we obtain

$$\Phi_1(\boldsymbol{\xi}) = \int_0^{2\pi}\int_0^\infty \mathcal{Q}_0(x,\theta)G(\boldsymbol{x};\boldsymbol{\xi})|_{r=R} R\,\mathrm{d}x\mathrm{d}\theta + K_1\xi. \tag{55}$$

As we have $\mathcal{Q}_0$ in the form of a series expansion, we can write

$$\Phi_1(\boldsymbol{\xi}) = K_1\xi + \sum_{n=1}^\infty F_n \int_0^{2\pi}\Bigg[-RR_X\lambda_n\psi_n\big|_{r=R}\int_0^\infty \mathrm{e}^{-\lambda_n x}\,G(\boldsymbol{x};\boldsymbol{\xi})\big|_{r=R}\,\mathrm{d}x + \left\{RR_X\lambda_n^2\psi_n\big|_{r=R} + \frac{\mathrm{d}}{\mathrm{d}\theta}\Big(\frac{R_X}{R}\psi_{n,\theta}\big|_{r=R}\Big)\right\}\int_0^\infty x\,\mathrm{e}^{-\lambda_n x}\,G(\boldsymbol{x};\boldsymbol{\xi})\big|_{r=R}\,\mathrm{d}x\Bigg]\mathrm{d}\theta. \tag{56}$$

As the series for $\mathcal{Q}_0$ converges uniformly for $x > 0$, we may exchange summation and integration. On the other hand, the fact that all basis functions have vanishing normal derivatives at the wall, i.e., $\nabla_\perp\psi_n\cdot\boldsymbol{n}_\perp = 0$, whereas $\nabla_\perp\Phi_1\cdot\boldsymbol{n}_\perp \neq 0$, suggests that this series does not converge uniformly near the wall.

The expression for Φ_1 is further specified by removing the x-integration:

$$\int_0^\infty \mathrm{e}^{-\lambda_n x}\,G(\boldsymbol{x};\boldsymbol{\xi})\Big|_{r=R}\,\mathrm{d}x = \frac{1-\mathrm{e}^{-\lambda_n\xi}}{\lambda_n^2} - \sum_{m=1}^\infty \psi_m(R,\theta)\psi_m(\rho,\eta)\frac{\mathrm{e}^{-\lambda_n\xi}-\mathrm{e}^{-\lambda_m\xi}}{\lambda_n^2-\lambda_m^2}, \tag{57}$$

$$\int_0^\infty x\,\mathrm{e}^{-\lambda_n x}\,G(\boldsymbol{x};\boldsymbol{\xi})\Big|_{r=R}\,\mathrm{d}x = \frac{2-(2+\lambda_n\xi)\,\mathrm{e}^{-\lambda_n\xi}}{\lambda_n^3} - \sum_{m=1}^\infty \psi_m(R,\theta)\psi_m(\rho,\eta)\frac{2\lambda_n(\mathrm{e}^{-\lambda_n\xi}-\mathrm{e}^{-\lambda_m\xi}) + \xi(\lambda_n^2-\lambda_m^2)\,\mathrm{e}^{-\lambda_n\xi}}{(\lambda_n^2-\lambda_m^2)^2}. \tag{58}$$

If $m = n$, the limit $\lambda_m \to \lambda_n$ should be taken. Now we are better able to recognize the nature of the nonuniform convergence. The dominating term is (we ignore for the moment the θ-integration)

$$\Phi_1(\boldsymbol{\xi}) \sim \sum_{m=1}^\infty \frac{\psi_m(R,\theta)\psi_m(\rho,\eta)}{\lambda_m^2}.$$

For a circular duct this may be compared, near $\rho = R$, to the prototype series

$$\sim \sum_{m=1}^{\infty} \frac{\cos(2\pi m\rho/R)}{m^2}.$$

The normal derivative yields the well-known saw-tooth function that vanishes (pointwise) at $\rho = R$ but converges to a finite nonzero value for any $\rho \neq R$.

For $x \to \infty$, the exponential terms in $\Phi_1(\boldsymbol{x})$ vanish and we have (we exchange the variables $\boldsymbol{x}$ and $\boldsymbol{\xi}$)

$$\Phi_1(\boldsymbol{x}) \simeq K_1 x + \sum_{n=1}^{\infty} F_n \int_0^{2\pi} \left[RR_X \lambda_n^{-1} \psi_n\big|_{\rho=R} + \frac{2}{\lambda_n} \frac{\mathrm{d}}{\mathrm{d}\eta} \Big(\frac{R_X}{R} \psi_{n,\eta}\big|_{\rho=R} \Big) \right] \mathrm{d}\eta.$$

By using the periodicity of ψ_n in its circumferential argument η, we have finally

$$\Phi_1(\boldsymbol{x}) \simeq K_1 x + \sum_{n=1}^{\infty} \frac{F_n}{\lambda_n} \int_0^{2\pi} RR_X \psi_n\big|_{\rho=R} \mathrm{d}\eta \qquad \text{for } \ x \to \infty. \tag{59}$$

4.3. Matching. Both the initial conditions for ϕ_0 and ϕ_1 and the constant K_1 are determined from matching with the outer solution. From (41) and (59) we have

$$\phi_0(0) + X\phi_{0,X}(0) + \varepsilon\phi_1(0) \sim F_0 + \varepsilon K_1 x + \varepsilon \sum_{n=1}^{\infty} \frac{F_n}{\lambda_n} \int_0^{2\pi} RR_X \psi_n\big|_{\rho=R} \mathrm{d}\eta, \tag{60}$$

and so we find

$$\begin{cases} \phi_0(0) = F_0, \\ \quad K_1 = \phi_{0,X}(0), \\ \phi_1(0) = \displaystyle\sum_{n=1}^{\infty} \frac{F_n}{\lambda_n} \int_0^{2\pi} RR_X \psi_n\big|_{\rho=R} \mathrm{d}\eta. \end{cases} \tag{61}$$

This determines the outer solution $\phi_0 + \varepsilon\phi_1$ (together with the radiation condition). It wouldn't be too difficult to guess that ϕ_0 depends on the average source excitation F_0, but the initial value for ϕ_1 is really subtle. The constant term in (59) is therefore probably the most important result of this tour de force to determine Φ_1.

An interesting question is then when ϕ_1 is present at all in the outer solution (or put in another way: what the error is if we only consider ϕ_0). For example, ϕ_1 is zero when the source consists of a simple piston with just $F(r,\theta) = F_0$, or when the duct entrance starts smoothly with $R_X = 0$, or when $RR_X\psi_n$ for all $n > 0$ are periodic along the circumference.

Although this last condition is not very likely to be possible, for a cylindrical duct at least the nonsymmetric modes vanish. In this case the eigenfunctions are given by (38). The integrals in (59) vanish for all $\nu \neq 0$. As a result we have

$$\phi_1(0) = 2\sqrt{\pi} RR_X \sum_{\mu=2}^{\infty} \frac{F_{0\mu}}{j'_{0\mu}}. \tag{62}$$

In other words, the first constant mode determines ϕ_0, while only the nonconstant symmetric modes determine ϕ_1. For example, a piston tilting along a diagonal like $F \sim r\sin\theta$ would produce a field vanishing to $\mathcal{O}(\varepsilon^2)$, while a "piston" that is symmetrically folded like $F \sim r^2$ would produce both $\mathcal{O}(1)$ and $\mathcal{O}(\varepsilon)$ terms.

5. Other coordinate systems. It was shown by Agullo, Barjau, and Keefe [26] that if the shape of the hard-walled duct is described in an orthogonal coordinate system (u, v, w) by the surface $\Sigma(v, w) = 0$, while the Helmholtz equation allows separable solutions of the form $\phi(u, v, w) = F(u)G(v, w)$, then there exist unidimensional (i.e., self-similar) waves in u of the type $\phi(u, v, w) = F(u)$. In this way it is possible to produce exact solutions of certain horn shapes, like the straight and exponential cone and others.

Although these solutions are interesting on their own, they have little to do with the present low k asymptotic problem, where the duct wall is never outside the lateral near field of the wave. Without this, there is no built-in mechanism that enforces the self-similarity, so any defect of symmetry in source or surface will produce deviations in the wave field that propagate without attenuation in other directions. Also the generalizations that will be discussed below are not possible at all or only in very limited form.

On the other hand, if the duct shape considered is close to one that allows such an exact solution, it may be advantageous, in terms of practical accuracy of the final result, to reformulate the problem in the other set of coordinates. The essence of the asymptotic problem remains the same.

We will illustrate this for spherical coordinates (r, θ, φ), where we temporarily redefine $x = r\cos\varphi$, $y = r\sin\varphi\cos\theta$, $z = r\sin\varphi\sin\theta$. (Note that we will use these coordinates *only in this section.*) A circular cone around the positive x-axis is given by $\varphi =$ constant, and a general cone of constant cross section by $\varphi = f(\theta)$.

In order to maintain the slender shape, necessary for the asymptotics, the duct will be long in r, compensated by a small opening angle in φ. We therefore introduce the scaled variables

$$\tau = \frac{2\sin\frac{1}{2}\varphi}{\varepsilon}, \qquad R = \varepsilon r \tag{63}$$

and write the general duct geometry as

$$\tilde{\Sigma}(R, \tau, \theta) = \tau - T(R, \theta) = 0, \tag{64}$$

where T is, by assumption, independent of ε. By this choice the surface area, $\tilde{A}(R)$ of any spherical cross section $R =$ constant is now exactly (i.e., independent of ε) equal to

$$\begin{aligned}\tilde{A}(R) &= \int_0^{2\pi}\int_0^{\varphi(R,\theta)} r^2\sin\varphi\,\mathrm{d}\varphi\mathrm{d}\theta = \int_0^{2\pi}\int_0^{T} r^2\varepsilon^2\tau\,\mathrm{d}\tau\mathrm{d}\theta \\ &= \frac{1}{2}R^2\int_0^{2\pi} T^2(R,\theta)\,\mathrm{d}\theta. \end{aligned} \tag{65}$$

Other choices for describing the duct shape are not essentially different, other than T, and therefore $\tilde{A}$, becoming dependent on ε. This gives complications in the form of extra asymptotic terms in the higher orders, which are irrelevant now.

The Helmholtz equation is given by

$$\frac{\varepsilon^2}{R^2}\frac{\partial}{\partial R}\Big(R^2\frac{\partial\phi}{\partial R}\Big) + \frac{1}{R^2\tau}\frac{\partial}{\partial\tau}\left(\tau\left(1-\frac{1}{4}\varepsilon^2\tau^2\right)\frac{\partial\phi}{\partial\tau}\right) + \frac{1}{R^2\tau^2(1-\frac{1}{4}\varepsilon^2\tau^2)}\frac{\partial^2\phi}{\partial\theta^2} + \varepsilon^2\kappa^2\phi = 0, \tag{66}$$

while the hard-wall boundary condition becomes

$$\nabla\phi\cdot\nabla\tilde{\Sigma} = \frac{1-\frac{1}{4}\varepsilon^2T^2}{R^2}\frac{\partial\phi}{\partial\tau} - \varepsilon^2\frac{\partial T}{\partial R}\frac{\partial\phi}{\partial R} - \frac{1}{R^2T^2(1-\frac{1}{4}\varepsilon^2T^2)}\frac{\partial T}{\partial\theta}\frac{\partial\phi}{\partial\theta} = 0. \tag{67}$$

We expand, as before,

$$\phi(R,\tau,\theta;\varepsilon) = \phi_0(R,\tau,\theta) + \varepsilon^2\phi_2(R,\tau,\theta) + \cdots$$

(skipping for now the $\mathcal{O}(\varepsilon)$-term) to obtain to leading order

$$\phi_{0,\tau\tau} + \frac{1}{\tau}\phi_{0,\tau} + \frac{1}{\tau^2}\phi_{0,\theta\theta} = 0, \quad \text{with} \quad \phi_{0,\tau} - \frac{T_\theta}{T^2}\phi_{0,\theta} = 0 \quad \text{at} \quad \tau = T. \tag{68}$$

If τ and θ are read as polar coordinates, this problem is qua form the same as (17), and thus we have the solution $\phi_0 = \phi_0(R)$ to be determined at the next order. We have

$$\phi_{2,\tau\tau} + \frac{1}{\tau}\phi_{2,\tau} + \frac{1}{\tau^2}\phi_{2,\theta\theta} + \big(R^2\phi_{0,R}\big)_R + R^2\kappa^2\phi_0 = 0,$$
$$\text{with} \quad \phi_{2,\tau} - \frac{T_\theta}{T^2}\phi_{2,\theta} = R^2T_R\phi_{0,R} \quad \text{at} \quad \tau = T.$$

This can be written as

$$\tilde{\nabla}^2\phi_2 + \big(R^2\phi_{0,R}\big)_R + R^2\kappa^2\phi_0 = 0, \quad \text{with} \quad \tilde{\nabla}\phi_2\cdot\tilde{\boldsymbol{n}} = R^2\phi_{0,R}\frac{TT_R}{\sqrt{T^2+T_\theta^2}}, \tag{69}$$

where $\tilde{\nabla}$ and $\tilde{\boldsymbol{n}}$ denote gradient and normal, respectively, in the (τ,θ)-plane. As a result we have virtually the same equation as (19), and after integration along a spherical surface $\tilde{A}(R)$ in (τ,θ) and using (65), we obtain

$$-\frac{\tilde{A}}{R^2}\Big(R^2\phi_{0,R}\Big)_R - \tilde{A}\kappa^2\phi_0 = \frac{1}{2}R^2\phi_{0,R}\frac{\mathrm{d}}{\mathrm{d}R}\int_0^{2\pi}T^2(R,\theta)\,\mathrm{d}\theta = R^2\phi_{0,R}\Big(\frac{\tilde{A}}{R^2}\Big)_R$$

or

$$\tilde{A}^{-1}\big(\tilde{A}\phi_{0,R}\big)_R + \kappa^2\phi_0 = 0. \tag{70}$$

We see that changing from the axial coordinate X to R and from the transverse cross section A to the spherical cross section $\tilde{A}$ leaves the final equation for ϕ_0 unchanged. Indeed, to the order considered, X and R and A and $\tilde{A}$ are the same.

6. Curved ducts. The present results remain valid for the slightly more general problem of curved ducts (like certain musical instruments) if the curvature of the duct axis (and its derivative) is $\mathcal{O}(\varepsilon)$. Together with the assumed slow variation in the axial coordinate, the associated orthogonal coordinate system (based on the tangent and, possibly, the normal and binormal of the curve that describes the duct axis) leave the Laplacian unchanged up to $\mathcal{O}(\varepsilon^3)$.

A simple example is the inside of a perturbed torus, described by a fixed torus radius ε^{-1} and slowly varying tube radius R. With local (polar-type) coordinates ξ, r, φ, we define

$$x = \varepsilon^{-1}(1+\varepsilon r\cos\theta)\cos(\varepsilon\xi), \quad y = \varepsilon^{-1}(1+\varepsilon r\cos\theta)\sin(\varepsilon\xi), \quad z = r\sin\theta, \tag{71}$$

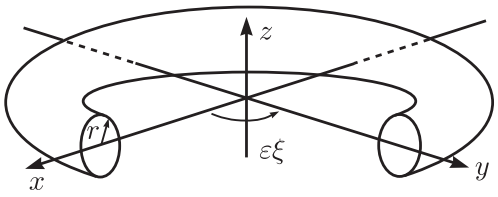


FIG. 3. *The torus coordinates.*

where $0 \le r \le R(\varepsilon\xi,\theta)$, $0 \le \theta < 2\pi$, $0 \le \varepsilon\xi < 2\pi$ (see Figure 3). If we write $X = \varepsilon\xi$, we get (cf. (6))

$$\nabla^2\phi + \varepsilon^2\kappa^2\phi$$
$$= \nabla_\perp^2\phi + \varepsilon^2\big(1+\varepsilon r\cos\theta\big)^{-2}\frac{\partial^2}{\partial X^2}\phi + \varepsilon\big(1+\varepsilon r\cos\theta\big)^{-1}\left[\cos\theta\frac{\partial}{\partial r}\phi - \frac{1}{r}\frac{\partial}{\partial\theta}\phi\right] + \varepsilon^2\kappa^2\phi = 0. \tag{72}$$

Boundary conditions at $\Sigma = r - R(X,\theta) = 0$ are

$$\nabla_\perp\phi\cdot\nabla_\perp\Sigma - \frac{\varepsilon^2 R_X\phi_X}{(1+\varepsilon r\cos\theta)^2} = 0. \tag{73}$$

If we expand $\phi = \phi_0 + \varepsilon\phi_1 + \varepsilon^2\phi_2 + \cdots$, we get to leading order

$$\nabla_\perp^2\phi_0 = 0, \qquad \text{with } \nabla_\perp\phi_0\cdot\boldsymbol{n}_\perp = 0, \tag{74}$$

and so $\phi_0 = \phi_0(X)$. Then $\frac{\partial}{\partial r}\phi_0 = \frac{\partial}{\partial\theta}\phi_0 = 0$, and we also have

$$\nabla_\perp^2\phi_1 = 0, \qquad \text{with } \nabla_\perp\phi_1\cdot\boldsymbol{n}_\perp = 0, \tag{75}$$

leading to $\phi_1 = \phi_1(X)$. Thus again $\frac{\partial}{\partial r}\phi_1 = \frac{\partial}{\partial\theta}\phi_1 = 0$, and we again obtain

$$\nabla_\perp^2\phi_2 + \phi_{0,XX} + \kappa^2\phi_0 = 0, \qquad \text{with } \nabla_\perp\phi_2\cdot\nabla_\perp\Sigma = \phi_{0,X}R_X,$$

yielding thus, after a similar argument as before, Webster's horn equation.

7. Impedance walls. If the duct walls is equipped with an impedance-type acoustic lining of complex impedance Z, we will in general (at least if $\mathrm{Re}(Z) > 0$) expect solutions that decay exponentially in the axial direction. Therefore, in the compressed variable X, only trivial (i.e., zero) solutions will exist. We will see that this is by and large the case, not only for dissipative walls with $\mathrm{Re}(Z) > 0$, but for any $|Z| < \infty$. Only for a purely imaginary impedance in a *straight* duct are there exceptions.

The impedance-wall boundary condition at $r = R$ is given by

$$\nabla\phi\cdot\boldsymbol{n} = -\frac{\mathrm{i}\varepsilon\kappa}{Z}\phi = \zeta\phi \tag{76}$$

with specific impedance Z. As before, we assume the Poincaré expansion $\phi = \phi_0 + \varepsilon\phi_1 + \varepsilon^2\phi_2 + \cdots$. First we note that it is easily verified that if $Z = 0$, only the trivial solutions $\phi_0 = \phi_1 = 0$ occur. Then we consider two possibilities: $Z = \mathcal{O}(1)$ and $Z = \mathcal{O}(\varepsilon)$.

7.1. $Z = \mathcal{O}(1)$. As $\zeta = \mathcal{O}(\varepsilon)$, we write $\zeta = \varepsilon\zeta_1$. In this case we have only trivial solutions. Expand equations and boundary conditions as before, to get to leading order

$$\nabla_\perp^2\phi_0 = 0, \quad \text{with} \quad \nabla_\perp\phi_0 \cdot \boldsymbol{n}_\perp = 0, \tag{77}$$

with solution $\phi_0 = \phi_0(X)$, a function to be determined. To first order we have

$$\nabla_\perp^2\phi_1 = 0, \quad \text{with} \quad \nabla_\perp\phi_1 \cdot \boldsymbol{n}_\perp = \zeta_1\phi_0. \tag{78}$$

Since

$$\iint_{\mathcal{A}} \nabla_\perp^2\phi_1 \,\mathrm{d}\sigma = \zeta_1\phi_0 \int_{\partial\mathcal{A}} \mathrm{d}\ell = 0, \tag{79}$$

we must have $\phi_0 = 0$, and so $\phi_1 = \phi_1(X)$. Nothing changes when we continue, and so all terms of the expansion vanish. Note that this is true for any Z.

7.2. $Z = \mathcal{O}(\varepsilon)$. Now we have $\zeta = \mathcal{O}(1)$, which changes the boundary condition expansion. To leading order we have

$$\nabla_\perp^2\phi_0 = 0 \ \text{ in } \ \mathcal{A}, \quad \text{with} \quad \nabla\phi_0 \cdot \boldsymbol{n}_\perp = \zeta\phi_0 \ \text{ at } \ \partial\mathcal{A}. \tag{80}$$

We would be tempted to assume that this problem has a solution or solutions for any given Z, but this is not true. Nontrivial solutions exist only for certain ζ. From Green's second identity applied to ϕ_0 and its complex conjugate, it can be deduced that any possible ζ is real. Furthermore, from Green's first identity applied to ϕ_0, it follows that any possible ζ is positive, and Z is thus negative imaginary.

But even with ζ real positive, there are only certain discrete values that allow a solution. This is best seen as follows. The problem described in (80) is an eigenvalue problem for the Dirichlet-to-Neumann operator Ξ: $f \mapsto g$, which maps a given Dirichlet boundary value f to the normal derivative g of f's harmonic extension into $\mathcal{A}$ (see [17]). In other words, $\Xi(f) = \frac{\partial}{\partial n}\psi\big|_{\partial\mathcal{A}}$, where ψ is the solution of

$$\nabla^2\psi = 0 \ \text{ in } \ \mathcal{A}, \quad \text{with} \quad \psi = f \ \text{ at } \ \partial\mathcal{A}. \tag{81}$$

As we are looking for $\Xi(\phi_0) = \zeta\phi_0$, equation (80) corresponds to the eigenvalue problem of Ξ. For the present discussion it is most relevant to know that this spectrum of eigenvalues of Ξ is discrete. As this result, due to Prof. Jan de Graaf, appears not to be available in the literature, it is considered concisely, but in great depth, in the appendix.

An example that illustrates this behavior explicitly is the circular duct $r = R(X)$, where

$$\phi_0 = f(X)\Big(\frac{r}{R(X)}\Big)^m \begin{Bmatrix}\cos m\theta\\ \sin m\theta\end{Bmatrix}, \quad \text{with} \quad \zeta = \frac{m}{R(X)}, \tag{82}$$

and m is a nonnegative integer. As the shape of the cross section $\mathcal{A}(X)$ changes with X, the discreteness of the spectrum of Ξ implies that the values of ζ that allow a solution also change with X, and in general there are no (nonzero) solutions possible along a varying duct for a fixed given ζ.

This is of course not true for a duct of constant cross section, $r = R(\theta)$, although now the asymptotics for small ε loses its meaning because there is no axial length

scale for the acoustic wave to be compared with. The problem simplifies further for the circular duct $r = R$, where (without approximation)

$$\phi(x, r, \theta) = J_m(\alpha r)\, \mathrm{e}^{-\mathrm{i}m\theta - \mathrm{i}\gamma x}, \quad \alpha^2 + \gamma^2 = k^2 \tag{83}$$

and the boundary condition requires

$$\frac{\alpha R J'_m(\alpha R)}{J_m(\alpha R)} = m - \frac{\alpha R J_{m+1}(\alpha R)}{J_m(\alpha R)} = \zeta R. \tag{84}$$

This equation has infinitely many solutions, but the wave is guaranteed unattenuated (γ real) if α is imaginary, say $\alpha = \mathrm{i}\tau$. Such solutions exist for real $\zeta \geq m/R$, because

$$\zeta R = m - \frac{\mathrm{i}\tau R J_{m+1}(\mathrm{i}\tau R)}{J_m(\mathrm{i}\tau R)} = m + \frac{\tau R I_{m+1}(\tau R)}{I_m(\tau R)} \geq m \tag{85}$$

(see [27]). Note that for small k, γ, α solutions we recover (82)

$$\zeta = \frac{m}{R} - \frac{\alpha^2 R}{2m+2} + \mathcal{O}(\alpha^4). \tag{86}$$

In other words, only solutions of this type exist near special values of ζ.

8. Variable mean soundspeed and density. If soundspeed $C = C(X, r, \theta)$ and mean density $D = D(X, r, \theta)$ are not uniformly constant, but vary in r, θ, and slowly in x, we have the reduced wave equation (5), rewritten in slowly varying coordinates as

$$\varepsilon^2 \frac{\partial}{\partial X}\big(C^2 p_X\big) + \nabla_\perp \cdot \big(C^2 \nabla_\perp p\big) + \varepsilon^2 \Omega^2 p = 0, \tag{87}$$

where the dimensionless frequency $\omega = \varepsilon\Omega$ is small. The hard-wall boundary condition is the same as (14). When we expand $p = p_0 + \varepsilon p_1 + \varepsilon^2 p_2 + \cdots$, we get to leading order

$$\nabla_\perp \cdot \big(C^2 \nabla_\perp p_0\big) = 0, \quad \text{with} \quad \nabla_\perp p_0 \cdot \boldsymbol{n}_\perp = 0, \tag{88}$$

which has a constant as the solution, so $p_0 = p_0(X)$, a function to be determined. We can derive the same equation for p_1, to get the same result $p_1 = p_1(X)$. For the second order we have

$$\nabla_\perp \cdot \big(C^2 \nabla_\perp p_2\big) + \frac{\partial}{\partial X}\big(C^2 p_{0,X}\big) + \Omega^2 p_0 = 0, \quad \text{with} \quad \nabla_\perp p_2 \cdot \boldsymbol{n}_\perp = p_{0,X} \frac{R R_X}{\sqrt{R^2 + R_\theta^2}}. \tag{89}$$

We go on to find a solvability condition for p_2 by integrating this equation along a cross section $\mathcal{A}$. Utilizing the following identity for any differentiable function f,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}X} \iint_{\mathcal{A}} f(\boldsymbol{X})\, \mathrm{d}\sigma &= \frac{\mathrm{d}}{\mathrm{d}X} \int_0^{2\pi}\!\!\int_0^R f(X, r, \theta) r\, \mathrm{d}r\theta \\ &= \int_0^{2\pi}\!\!\int_0^R f_X r\, \mathrm{d}r\mathrm{d}\theta + \int_0^{2\pi} f(X, R, \theta) R R_X\, \mathrm{d}\theta, \end{aligned} \tag{90}$$

we have

$$\iint_{\mathcal{A}} \nabla_\perp \cdot (C^2 \nabla_\perp p_2)\,\mathrm{d}\sigma = p_{0,X} \int_0^{2\pi} C^2 R R_X \mathrm{d}\theta$$

$$= p_{0,X}\left[\frac{\mathrm{d}}{\mathrm{d}X}\iint_{\mathcal{A}} C^2\,\mathrm{d}\sigma - \iint_{\mathcal{A}} \frac{\partial}{\partial X} C^2\,\mathrm{d}\sigma\right]. \tag{91a}$$

Furthermore, we have

$$\iint_{\mathcal{A}} \frac{\partial}{\partial X}\left(C^2 p_{0,X}\right)\mathrm{d}\sigma = p_{0,X}\iint_{\mathcal{A}} \frac{\partial}{\partial X} C^2\,\mathrm{d}\sigma + p_{0,XX}\iint_{\mathcal{A}} C^2\,\mathrm{d}\sigma,$$

$$\text{and} \quad \iint_{\mathcal{A}} \Omega^2 p_0\,\mathrm{d}\sigma = \Omega^2 p_0 A. \tag{91b}$$

Then, after introducing the cross-sectional averaged squared soundspeed

$$\overline{C^2} = \frac{1}{A}\iint_{\mathcal{A}} C^2\,\mathrm{d}\sigma, \tag{92}$$

a generalization of Webster's horn equation is obtained:

$$A^{-1}\left(A\overline{C^2}p_{0,X}\right)_X + \Omega^2 p_0 = 0. \tag{93}$$

This may be further simplified by the transformation

$$A(X)\overline{C^2}(X) = d(X)^2, \quad p_0 = d^{-1}\psi \tag{94}$$

into

$$\psi'' + \left(\frac{\Omega^2}{\overline{C^2}} - \frac{d''}{d}\right)\psi = 0. \tag{95}$$

9. Irrotational and isentropic mean flow. To analyze asymptotically low-frequency acoustic perturbations in a slowly varying duct with an irrotational isentropic mean flow, as described by (7) and (9), we need to approximate both mean flow and acoustic field to the same order of accuracy.

We start here with the mean flow. In the dimensionless variables used, we have $C^2 = D^{\gamma-1}$, so equations (7) simplify to

$$\tfrac{1}{2}V^2 + \frac{D^{\gamma-1}}{\gamma-1} = E, \qquad \nabla\cdot(D\boldsymbol{V}) = 0. \tag{96}$$

The mass flux at any cross section A is given by

$$\iint_{\mathcal{A}} DU\,\mathrm{d}\sigma = \mathcal{F}. \tag{97}$$

Due to the nondimensionalization, U, D, A, $\mathcal{F}$, and E are $\mathcal{O}(1)$. Introduce the slow variable $X = \varepsilon x$, and assume that $\boldsymbol{V}$ and D depend essentially on X, rather than x. We write the velocity as

$$\boldsymbol{V} = U\boldsymbol{e}_x + \boldsymbol{V}_\perp \tag{98}$$

to distinguish between axial and crosswise components. If flux $\mathcal{F}$ and thermodynamical constant E are given and independent of ε, we can expand $U = U_0 + \mathcal{O}(\varepsilon^2)$ and $D = D_0 + \mathcal{O}(\varepsilon^2)$. As the flow is a potential flow, we can derive, in the same way as in Rienstra [14, 16], that $D_0 = D_0(X)$, $U_0 = U_0(X)$, and $\boldsymbol{V}_\perp = \varepsilon \widetilde{\boldsymbol{V}}_{\perp 0} + \mathcal{O}(\varepsilon^3)$, satisfying the equations (to be solved numerically)

$$D_0 U_0 A = \mathcal{F}, \qquad \frac{\mathcal{F}^2}{2D_0^2 A^2} + \frac{D_0^{\gamma-1}}{\gamma - 1} = E. \tag{99}$$

9.1. Mean flow and hard walls. Next we consider the acoustic field. Using the above results for the mean flow, (9) becomes to leading order

$$\begin{aligned}&\nabla_\perp^2 \phi + \varepsilon^2 D_0^{-1} \big(D_0 \phi_X\big)_X \\ &\qquad = \varepsilon^2 \left(\mathrm{i}\Omega + U_0 \frac{\partial}{\partial X} + \widetilde{\boldsymbol{V}}_{\perp 0} \boldsymbol{\cdot} \nabla_\perp\right) \left[C_0^{-2} \left(\mathrm{i}\Omega + U_0 \frac{\partial}{\partial X} + \widetilde{\boldsymbol{V}}_{\perp 0} \boldsymbol{\cdot} \nabla_\perp\right) \phi\right],\end{aligned}$$

with hard wall boundary condition

$$\nabla \phi \boldsymbol{\cdot} \boldsymbol{n} = 0 \quad \text{at} \quad r = R.$$

We expand $\phi = \phi_0 + \varepsilon \phi_1 + \varepsilon^2 \phi_2 + \cdots$. To leading order we have

$$\nabla_\perp^2 \phi_0 = 0, \quad \nabla_\perp \phi_0 \boldsymbol{\cdot} \boldsymbol{n}_\perp = 0, \tag{100}$$

yielding the constant solution, i.e., $\phi_0 = \phi_0(X)$.

To first order we have the same equation. To second order we have

$$\begin{aligned}&\nabla_\perp^2 \phi_2 + D_0^{-1} \big(D_0 \phi_{0,X}\big)_X \\ &\qquad = \left(\mathrm{i}\Omega + U_0 \frac{\partial}{\partial X} + \widetilde{\boldsymbol{V}}_{\perp 0} \boldsymbol{\cdot} \nabla_\perp\right) \left[C_0^{-2} \left(\mathrm{i}\Omega + U_0 \frac{\partial}{\partial X} + \widetilde{\boldsymbol{V}}_{\perp 0} \boldsymbol{\cdot} \nabla_\perp\right) \phi_0\right],\end{aligned}$$

with boundary conditions given by (19). After integration across a cross section $\mathcal{A}(X)$, we obtain, similar to before, Webster's horn equation generalized for irrotational isentropic mean flow:

$$(D_0 A)^{-1} (D_0 A \phi_{0,X})_X = \left(\mathrm{i}\Omega + U_0 \frac{\partial}{\partial X}\right) \left[C_0^{-2} \left(\mathrm{i}\Omega + U_0 \frac{\partial}{\partial X}\right) \phi_0\right]. \tag{101}$$

This result seems to be equivalent to equations given by [8, 9, 10, 11, 12] and (apart from a factor $\frac{1}{2}$) [3, p. 422].

9.2. Mean flow and impedance walls. The problem with mean flow and an impedance wall is more intricate. Instead of the duct wall boundary condition given in (76), we have Myers' condition [28], rewritten (see [29, 30]) as follows:

$$\mathrm{i}\omega D\big(\boldsymbol{v} \boldsymbol{\cdot} \boldsymbol{n}\big) = \frac{\mathrm{i}\omega D p}{Z} + \mathcal{M}\Big(\frac{D \boldsymbol{V} p}{Z}\Big), \tag{102}$$

where impedance $Z = Z(X, \theta)$ may be function of position, and operator $\mathcal{M}$ is defined by

$$\mathcal{M}(\boldsymbol{F}) = \nabla \boldsymbol{\cdot} \boldsymbol{F} - \boldsymbol{n} \boldsymbol{\cdot} (\boldsymbol{n} \boldsymbol{\cdot} \nabla \boldsymbol{F}). \tag{103}$$

Since $\mathcal{M}(\frac{D\boldsymbol{V}p}{Z}) = \mathcal{O}(\varepsilon)$, we write $\mathcal{M}(\frac{D\boldsymbol{V}p}{Z}) = \varepsilon\widetilde{\mathcal{M}}(\frac{D\boldsymbol{V}p}{Z})$. After expanding $\phi = \phi_0 + \varepsilon\phi_1 + \cdots$ and $p = \varepsilon p_0 + \cdots$ with

$$p_0 = -D_0\left(\mathrm{i}\Omega + U_0\frac{\partial}{\partial X} + \widetilde{\boldsymbol{V}}_{\perp 0}\cdot\nabla_\perp\right)\phi_0, \tag{104}$$

we get

$$\mathrm{i}\Omega D_0\big(\nabla_\perp\phi_0\cdot\boldsymbol{n}_\perp\big) + \mathrm{i}\varepsilon\Omega D_0\big(\nabla_\perp\phi_1\cdot\boldsymbol{n}_\perp\big) = \varepsilon\frac{\mathrm{i}\Omega D_0 p_0}{Z} + \varepsilon\widetilde{\mathcal{M}}\Big(\frac{D_0\boldsymbol{V}_0 p_0}{Z}\Big) + \mathcal{O}(\varepsilon^2), \tag{105}$$

where $\boldsymbol{V}_0 = U_0\boldsymbol{e}_x + \varepsilon\widetilde{\boldsymbol{V}}_{\perp 0}$.

9.2.1. $Z = \mathcal{O}(1)$. As before, we get to leading order

$$\nabla_\perp^2\phi_0 = 0, \quad \text{with} \quad \nabla_\perp\phi_0\cdot\boldsymbol{n}_\perp = 0,$$

so $\phi_0 = \phi_0(X)$ and therefore $p_0 = p_0(X)$. To first order we have the same equation $\nabla_\perp^2\phi_1 = 0$ for ϕ_1, but the boundary condition is now

$$\mathrm{i}\Omega D_0\big(\nabla_\perp\phi_1\cdot\boldsymbol{n}_\perp\big) = \frac{\mathrm{i}\Omega D_0 p_0}{Z} + \widetilde{\mathcal{M}}\Big(\frac{D_0\boldsymbol{V}_0 p_0}{Z}\Big). \tag{106}$$

In order to continue, we need from [16] the following property of the operator $\mathcal{M}$.

For any sufficiently smooth vectorfield with $\boldsymbol{f}\cdot\boldsymbol{n} = 0$ at $r = R$, we have

$$\int_{\partial\mathcal{A}}\Big[\nabla\cdot\boldsymbol{f} - \boldsymbol{n}\cdot\big(\boldsymbol{n}\cdot\nabla\boldsymbol{f}\big)\Big]\left\|\frac{\partial\boldsymbol{r}}{\partial x}\times\frac{\partial\boldsymbol{r}}{\partial\ell}\right\|\mathrm{d}\ell = \frac{\mathrm{d}}{\mathrm{d}x}\int_{\partial\mathcal{A}}\big(\boldsymbol{f}\times\boldsymbol{n}\big)\cdot\mathrm{d}\boldsymbol{\ell},$$

where $(x,\ell)\mapsto\boldsymbol{r}(x,\ell)$ is a parameterization of the surface.

Since

$$\left\|\frac{\partial\boldsymbol{r}}{\partial x}\times\frac{\partial\boldsymbol{r}}{\partial\ell}\right\| = \sqrt{1+\varepsilon^2\frac{R^2R_X^2}{R^2+R_\theta^2}} = 1+\mathcal{O}(\varepsilon^2),$$

we have as a result

$$\int_{\partial\mathcal{A}}\widetilde{\mathcal{M}}\Big(\frac{D_0\boldsymbol{V}_0 p_0}{Z}\Big)\mathrm{d}\ell = \frac{\mathrm{d}}{\mathrm{d}X}\int_{\partial\mathcal{A}}\frac{D_0U_0p_0}{Z}\mathrm{d}\ell + \mathcal{O}(\varepsilon).$$

We apply this to the equation for ϕ_1, in order to obtain an equation for ϕ_0. From

$$\iint_{\mathcal{A}}\nabla_\perp^2\phi_1\,\mathrm{d}\sigma = \int_{\partial\mathcal{A}}\nabla_\perp\phi_1\cdot\boldsymbol{n}_\perp\,\mathrm{d}\ell = 0,$$

together with (106) and noting that most functions depend on X only, it follows that

$$\mathrm{i}\Omega D_0p_0\mathcal{L} + \frac{\mathrm{d}}{\mathrm{d}X}\big(U_0D_0p_0\mathcal{L}\big) = 0, \quad \text{where} \quad \mathcal{L}(X) = \int_{\partial\mathcal{A}}\frac{1}{Z}\,\mathrm{d}\ell$$

($\mathcal{L}$ may be interpreted as the "total admittance" at X), with solution

$$p_0 = \text{constant}\,\frac{1}{U_0D_0\mathcal{L}}\exp\left(-\mathrm{i}\int^X\frac{\Omega}{U_0(\xi)}\,\mathrm{d}\xi\right).$$

Here ϕ_0 follows from (104) but is more difficult to obtain in explicit form. Note that this pressure field is not an acoustic wave, but it is of hydrodynamic nature. It does not propagate with the soundspeed, but with the mean flow velocity.

9.2.2. $Z = \mathcal{O}(\varepsilon)$. When $Z = \varepsilon Z_0$, we get for ϕ_0 the apparently difficult boundary condition

$$\mathrm{i}\Omega D_0\big(\nabla_\perp \phi_0 \cdot \boldsymbol{n}_\perp\big) = \frac{\mathrm{i}\Omega D_0 p_0}{Z_0} + \widetilde{\mathcal{M}}\Big(\frac{D_0 \boldsymbol{V}_0 p_0}{Z_0}\Big),$$

which is, analogous to the no-flow case, likely to be an eigenvalue problem with discrete eigenvalues Z_0 (apart from the trivial solutions $p_0 = 0$, $\phi_0 = \phi_0(X) \propto \exp(-\mathrm{i}\Omega \int^X U_0(\xi)^{-1}\mathrm{d}\xi)$, i.e., hydrodynamically convected pressureless perturbations). If this conjecture is true, the possible eigenvalues vary with the geometry, and no other than the trivial solution is possible in a varying duct.

10. Conclusions. Generalizations of Webster's classic horn equation for nonuniform media, lined walls, and mean flow have been derived systematically, as an asymptotic perturbation problem for low Helmholtz number and slowly varying duct diameter. The conditions on frequency, acoustic medium, and duct geometry are explicitly indicated in terms of small parameter ε, the ratio between a typical length of duct variation and the duct diameter. The error and higher order corrections are also explicitly stated.

The presence of lining in a varying duct is shown to allow in general only trivial or merely hydrodynamic solutions. A curved duct is shown to produce the same equation if the radius of curvature is not smaller than the typical wavelength or duct length scale.

The approximation is nonuniform near a source or entrance. The prevailing boundary layer solution for an arbitrary duct cross section is given, together with the $\mathcal{O}(1)$ and $\mathcal{O}(\varepsilon)$ matching conditions to the outer ("Webster") region. From these expressions conditions are derived for which the $\mathcal{O}(\varepsilon)$-outer field is absent.

Appendix. On the spectrum of the Dirichlet-to-Neumann operator Ξ on smooth bounded domains in $\mathbb{R}^2$. We will show that the Dirichlet-to-Neumann operator Ξ, introduced in section 7 (see (80)), has a discrete spectrum of finite multiplicity. The basic idea is to relate the problem for the general simply connected open domain $\Omega \subset \mathbb{R}^2$ (which has apparently no explicit solution), via conformal mapping, to the corresponding problem for the unit disk D, which does have a simple explicit solution.

Note that the related result for an annular domain is entirely analogous.

Step 1. Consider the open unit-disk $D \subset \mathbb{R}^2$. Its boundary ∂D, the unit circle, is parametrized by the angle θ, with $0 \le \theta < 2\pi$. The set of functions

$$\mathrm{e}_n : \quad \theta \mapsto \mathrm{e}_n(\theta) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{\mathrm{i}n\theta}, \quad n \in \mathbb{Z}, \tag{A.1}$$

establishes an orthonormal basis in $L_2(\partial D; \mathrm{d}\theta)$.[4] We introduce for real a the linear operator $\mathcal{N}_a$ in $L_2(\partial D, \mathrm{d}\theta)$, defined via the way it acts on the basis $\{\mathrm{e}_n\}$,

$$\mathcal{N}_a : \quad \mathrm{e}_n \mapsto \mathcal{N}_a \mathrm{e}_n, \quad \text{with} \quad \mathcal{N}_a \mathrm{e}_n(\theta) = (|n| + a)\mathrm{e}_n(\theta), \tag{A.2}$$

followed by linear extension and closure.

[4] $L_2(U; w(x)\mathrm{d}x)$ denotes the space of square integrable functions, defined on U, with inner product $\int_U \overline{f(x)} g(x) w(x)\,\mathrm{d}x$.

Let $u : \partial D \to \mathbb{C}$ be a sufficiently smooth function. Let $u_\mathcal{H}$, the harmonic extension of u, denote the (unique) solution of the Dirichlet problem

$$\nabla^2 u_\mathcal{H}(\boldsymbol{x}) = 0 \quad \text{for} \quad \boldsymbol{x} \in D, \quad \text{while} \quad u_\mathcal{H}(\boldsymbol{x}) = u \quad \text{for} \quad \boldsymbol{x} \in \partial D. \tag{A.3}$$

The normal derivative at the boundary ∂D produces a function

$$\frac{\partial}{\partial n} u_\mathcal{H} : \quad \partial D \to \mathbb{C}. \tag{A.4}$$

Altogether this defines the linear mapping $u \mapsto \frac{\partial}{\partial n} u_\mathcal{H}$, which is called the *Dirichlet-to-Neumann operator* in $L_2(\partial D; \mathrm{d}\theta)$. By noting that $\mathrm{e}_{n\mathcal{H}}(\boldsymbol{x}) = (x \pm \mathrm{i}y)^{|n|} = r^{|n|}\,\mathrm{e}^{\pm \mathrm{i}|n|\theta}$, and hence $\frac{\partial}{\partial n}\mathrm{e}_{n\mathcal{H}} = |n|\mathrm{e}_n$ at ∂D, it is easily verified that this operator is just equal to $\mathcal{N}_0$.

Step 2. Consider the bounded open domain $\Omega \subset \mathbb{R}^2$ with piecewise smooth boundary $\partial\Omega$. Let $v : \partial\Omega \to \mathbb{C}$ be a sufficiently smooth function. As in the previous section (just replace D by Ω), we introduce

$$\Xi : \quad v \mapsto \Xi v = \frac{\partial}{\partial n} v_\mathcal{H}, \tag{A.5}$$

the Dirichlet-to-Neumann operator in $L_2(\partial\Omega; \mathrm{d}\theta)$. Thus $\Xi = \mathcal{N}_0$ if $\Omega = D$. We want to show that Ξ is nonnegative self-adjoint with a pure point spectrum of finite multiplicity. In the previous paragraph we showed this to be true in $L_2(\partial D; \mathrm{d}\theta)$.

The self-adjointness and nonnegativity follows, formally, from Green's first and second identities (see section 7). In order to achieve some spectral results, we invoke the Riemann mapping theorem and consider a conformal mapping $\beta : D \to \Omega$. The supposed smoothness of $\partial\Omega$ implies that the parametrization $\theta \mapsto \beta(\mathrm{e}^{\mathrm{i}\theta})$ for $\partial\Omega$ is such that both $|\beta'(\mathrm{e}^{\mathrm{i}\theta})|$ and its reciprocal are bounded.

Standard results from conformal mapping theory and harmonic functions on $\mathbb{R}^2$ lead to

$$\Xi v\big(\beta(\mathrm{e}^{\mathrm{i}\theta})\big) = \left(\frac{\partial}{\partial n} v_\mathcal{H}\right)\big(\beta(\mathrm{e}^{\mathrm{i}\theta})\big) = \big|\beta'(\mathrm{e}^{\mathrm{i}\theta})\big|^{-1} \frac{\partial}{\partial n}(v \circ \beta)_\mathcal{H}(\mathrm{e}^{\mathrm{i}\theta}). \tag{A.6}$$

This means that, instead of the original problem, we could study the eigenvalue problem

$$\mathcal{B}\mathcal{N}_0 u = \lambda u \tag{A.7}$$

in $L_2(\partial D, \mathrm{d}\theta)$, with $\mathcal{B}$ the multiplication operator defined by

$$(\mathcal{B}w)(\theta) = B(\theta)w(\theta) = \big|\beta'(\mathrm{e}^{\mathrm{i}\theta})\big|^{-1} w(\theta). \tag{A.8}$$

(Although the inverse $\mathcal{B}^{-1}$ involves no more than division by the function $B(\theta)$, we retain for clarity the operator symbolism.)

Step 3. In order to turn the operator $\mathcal{B}\mathcal{N}_0$ into a self-adjoint one, we consider the eigenvalue problem in $L_2(\partial D; B^{-1}(\theta)\mathrm{d}\theta)$, which is topologically equivalent to $L_2(\partial D; \mathrm{d}\theta)$. Note that

$$\big\{\theta \mapsto u(\theta)\big\} \mapsto \big\{\theta \mapsto B^{-\frac{1}{2}}(\theta) u(\theta)\big\} \tag{A.9}$$

furnishes a unitary transformation from $L_2(\partial D; B^{-1}(\theta)\mathrm{d}\theta)$ to $L_2(\partial D; \mathrm{d}\theta)$, because

$$\int_0^{2\pi} \overline{u(\theta)} v(\theta) B^{-1}(\theta)\,\mathrm{d}\theta = \int_0^{2\pi} \overline{\left(B^{-\frac{1}{2}}(\theta)u(\theta)\right)}\left(B^{-\frac{1}{2}}(\theta)v(\theta)\right)\mathrm{d}\theta. \tag{A.10}$$

At the same time this implies that the eigenvalue problem (A.7) is unitary equivalent to the eigenvalue problem

$$\mathcal{B}^{\frac{1}{2}}\mathcal{N}_0\mathcal{B}^{\frac{1}{2}}\varphi = \lambda\varphi, \tag{A.11}$$

with $\varphi = \mathcal{B}^{-\frac{1}{2}}u$.

Step 4. If we can show that $(\mathcal{I}+\mathcal{B}^{\frac{1}{2}}\mathcal{N}_0\mathcal{B}^{\frac{1}{2}})^{-1}$ (where $\mathcal{I}$ is the identity) is a compact self-adjoint operator, we are ready. In that case it has a discrete spectrum with finite multiplicity [31], and the same holds, a fortiori, for $\mathcal{B}^{\frac{1}{2}}\mathcal{N}_0\mathcal{B}^{\frac{1}{2}}$.

Take a positive and sufficiently small such that $\theta \mapsto B^{-1}(\theta) - a$ is still positive and uniformly bounded away from zero. By noting that $\mathcal{N}_a = \mathcal{N}_0 + a\mathcal{I}$, we can rewrite

$$\mathcal{I} + \mathcal{B}^{\frac{1}{2}}\mathcal{N}_0\mathcal{B}^{\frac{1}{2}} = \mathcal{B}^{\frac{1}{2}}\mathcal{N}_a^{\frac{1}{2}}\left\{\mathcal{N}_a^{-\frac{1}{2}}(\mathcal{B}^{-1} - a\mathcal{I})\mathcal{N}_a^{-\frac{1}{2}} + \mathcal{I}\right\}\mathcal{N}_a^{\frac{1}{2}}\mathcal{B}^{\frac{1}{2}}. \tag{A.12}$$

The operator between brackets, { }, is bounded, positive, and self-adjoint and has an inverse with the same properties. We thus find

$$\left(\mathcal{I} + \mathcal{B}^{\frac{1}{2}}\mathcal{N}_0\mathcal{B}^{\frac{1}{2}}\right)^{-1} = \mathcal{B}^{-\frac{1}{2}}\mathcal{N}_a^{-\frac{1}{2}}\left\{\mathcal{N}_a^{-\frac{1}{2}}(\mathcal{B}^{-1} - a\mathcal{I})\mathcal{N}_a^{-\frac{1}{2}} + \mathcal{I}\right\}^{-1}\mathcal{N}_a^{-\frac{1}{2}}\mathcal{B}^{-\frac{1}{2}}, \tag{A.13}$$

which is a composition of operators. Since the factor $\mathcal{N}_a^{-\frac{1}{2}}$ is compact, also $(\mathcal{I} + \mathcal{B}^{\frac{1}{2}}\mathcal{N}_0\mathcal{B}^{\frac{1}{2}})^{-1}$ is a compact operator.

Acknowledgement. This work was carried out in the context of the "TurboNoiseCFD" project of the European Union's 5th Framework "Growth" Programme.

We gratefully acknowledge the contribution on the discreteness of the spectrum of Ξ (see the appendix) by J. de Graaf, and the cooperation with T. D. Chandra on the analysis of the entrance boundary layer. We thank N. C. Ovenden for helpful discussions and critical reading of the text.

REFERENCES

[1] A. G. Webster, *Acoustical impedance, and the theory of horns and of the phonograph*, Proc. Natl. Acad. Sci. USA, 5 (1919), pp. 275–282; reprinted in J. Audio Engineering Soc., 25 (1977), pp. 24–28.

[2] P. M. Morse, *Vibration and Sound*, 2nd ed., McGraw-Hill, New York, 1948.

[3] A. D. Pierce, *Acoustics: An Introduction to Its Physical Principles and Applications*, McGraw–Hill, New York, 1981.

[4] M. B. Lesser and D. G. Crighton, *Physical acoustics and the method of matched asymptotic expansions*, in Physical Acoustics 11, W. P. Mason and R. H. N. Thurston, eds., Academic Press, New York, 1975, pp. 69–149.

[5] M. B. Lesser and J. A. Lewis, *Applications of matched asymptotic expansion methods to acoustics.* I. *The Webster equation and the stepped duct*, J. Acoust. Soc. Amer., 51 (1972), pp. 1664–1669.

[6] M. B. Lesser and J. A. Lewis, *Applications of matched asymptotic expansion methods to Acoustics.* II. *The open ended duct*, J. Acoust. Soc. Amer., 52 (1972), pp. 1406–1410.

[7] M. Van Dyke, *Slow variations in continuum mechanics*, in Advances in Applied Mechanics, Vol. 25, Academic Press, Orlando, FL, 1987, pp. 1–45.

[8] N. A. Eisenberg and T. W. Kao, *Propagation of sound through a variable-area duct with a steady compressible flow*, J. Acoust. Soc. Amer., 49 (1971), pp. 169–175.

[9] P. Huerre and K. Karamcheti, *Propagation of sound through a fluid moving in a duct of varying area*, in Proceedings of the Interagency Symposium on University Research in Transportation Noise, Stanford, CA, 1973, Stanford University Press, Vol. II, pp. 397–413.
[10] E. Lumsdaine and S. Ragab, *Effect of flow on quasi-one-dimensional acoustic wave propagation in a variable duct of finite length*, J. Sound and Vibration, 53 (1977), pp. 47–51.
[11] C. Thompson and R. Sen, *Acoustic Wave Propagation in a Variable Area Duct Carrying a Mean Flow*, in Proceedings of the AIAA/NASA 9th Aeroacoustics Conference, Williamsburg, VA, 1984, paper AIAA-84-2336.
[12] L. M. B. C. Campos, *On linear and non-linear wave equations for the acoustics of high-speed potential flows*, J. Sound and Vibration, 110 (1986), pp. 41–57.
[13] A. H. Nayfeh and D. P. Telionis, *Acoustic propagation in ducts with varying cross sections*, J. Acoust. Soc. Amer., 54 (1973), pp. 1654–1661.
[14] S. W. Rienstra, *Sound transmission in slowly varying circular and annular ducts with flow*, J. Fluid Mech., 380 (1999), pp. 279–296.
[15] S. W. Rienstra and W. Eversman, *A numerical comparison between multiple-scales and FEM solution for sound propagation in lined flow ducts*, J. Fluid Mech., 437 (2001), pp. 367–384.
[16] S. W. Rienstra, *Sound propagation in slowly varying lined flow ducts of arbitrary cross section*, J. Fluid Mech., 495 (2003), pp. 157–173.
[17] T. D. Chandra, *Perturbation and Operator Methods for Solving Stokes Flow and Heat Flow Problems*, Ph.D. thesis, Technische Universiteit Eindhoven, Eindhoven, The Netherlands, 2002.
[18] P. A. Lagerstrom, *Matched Asymptotic Expansions: Ideas and Techniques*, Springer-Verlag, New York, 1988.
[19] W. Eckhaus, *Asymptotic Analysis of Singular Perturbations*, North–Holland, Amsterdam, 1979.
[20] A. H. Nayfeh, *Perturbation Methods*, John Wiley & Sons, New York, 1973.
[21] V. Salmon, *Generalized plane wave horn theory*, J. Acoust. Soc. Amer., 17 (1946), pp. 199–218.
[22] V. Salmon, *A new family of horns*, J. Acoust. Soc. Amer., 17 (1946), pp. 212–218.
[23] E. Eisner, *Complete solutions of the Webster horn equation*, J. Acoust. Soc. Amer., 41 (1967), pp. 1126–1146.
[24] A. H. Benade and E. V. Jansson, *On plane and spherical waves in horns with nonuniform flare.* I. *Theory of radiation, resonance frequencies, and mode conversion*, Acustica, 31 (1974), pp. 79–98.
[25] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 5th ed., Alan Jeffrey, ed., Academic Press, London, 1994.
[26] J. Agullo, A. Barjau, and D. H. Keefe, *Acoustic propagation in flaring, axisymmetric horns*: I *A new family of unidirectional solutions*, Acustica—Acta Acustica, 85 (1999), pp. 278–284.
[27] S. W. Rienstra, *A classification of duct modes based on surface waves*, Wave Motion, 37 (2003), pp. 119–135.
[28] M. K. Myers, *On the acoustic boundary condition in the presence of flow,* J. Sound and Vibration, 71 (1980), pp. 429–434.
[29] W. Möhring, *Energy conservation, time-reversal invariance, and reciprocity in ducts with flow,* J. Fluid Mech., 431 (2001), pp. 223–237.
[30] W. Eversman, *The boundary condition at an impedance wall in a non-uniform duct with potential mean flow*, J. Sound and Vibration, 246 (2001), pp. 63–69.
[31] J. Weidmann, *Linear Operators in Hilbert Spaces*, Springer-Verlag, New York, 1980.

SIAM J. APPL. MATH.
Vol. 65, No. 6, pp. 2005–2032

REVEALING PAIRWISE COUPLING IN LINEAR-NONLINEAR NETWORKS*

DUANE Q. NYKAMP†

Abstract. Through an asymptotic analysis of a simple network, we derive an estimate of the coupling between a pair of units when all other units are unobservable. The analysis is based on a model where the response of each unit is a linear-nonlinear function of a white noise stimulus. The results accurately determine the coupling when all unmeasured units respond to the stimulus differently than the measured pair. To account for the possibility of unmeasured units similar to the measured pair, we cast our results in the framework of "subpopulations," which are defined as a group of units who respond to the stimulus similarly. We demonstrate that we can determine when correlations between two units are caused by a connection between their subpopulations, although the precise identity of the units involved in the connection may remain ambiguous. The result is rigorously valid only when the coupling is sufficiently weak to justify a second-order approximation in the coupling strength. We demonstrate through simulations that the results are still valid even with stronger coupling and in the presence of some deviations from the linear-nonlinear model. The analysis is presented in terms of neuronal networks, although the general framework is more widely applicable.

Key words. neural networks, correlations, Weiner analysis, white noise

AMS subject classification. 92C20

DOI. 10.1137/S0036139903437072

1. Introduction.

This analysis of coupling within networks is motivated by neuroscience, and we use the vocabulary of neuroscience throughout. The measured response properties of a neuron arise from the structure of the neural network in which the neuron is embedded. To understand the relationship between these response properties and the neural network structure, one would like to simultaneously measure the response of neurons and estimate their connectivity. However, it has proven difficult to estimate the connectivity from measurements of neural activity because only a small subset of neurons can be monitored simultaneously.

In particular, a direct connection between two measured neurons is difficult to distinguish from a connection onto both neurons that originates from a third, unmeasured neuron. We refer to the latter configuration as the common input configuration. We address the case where one simultaneously measures two neurons in a network and attempts to distinguish the direct connection configuration from the common input configuration.

This distinction is especially difficult because when studying a network, one typically does not directly measure the internal state of neurons, but records only their discrete output events, called spikes. From simultaneous recordings of two neurons' spike times, one can analyze the joint statistics of the two spike trains in hopes of detecting a direct connection. Two widely used tools are the joint peristimulus time histogram (JPSTH) and its integral, the shuffle-corrected correlogram [14, 1, 13]. Unfortunately, inferences from the JPSTH or correlogram about the connections between

*Received by the editors November 3, 2003; accepted for publication (in revised form) February 7, 2005; published electronically August 9, 2005. This research was supported in part by an NSF Mathematical Sciences Postdoctoral Research Fellowship and by NSF DMS-0415409.
http://www.siam.org/journals/siap/65-6/43707.html
†School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (nykamp@math.umn.edu).

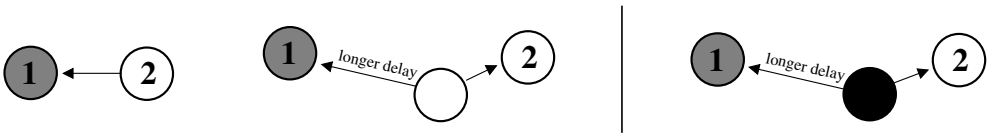


FIG. 1. *To determine "subpopulation connectivity," one needs to distinguish a direct connection from only certain kinds of common input. Three sample network configuration are shown, where neurons one and two are measured and the unlabeled neuron is not measured. The subpopulation of each neuron, which is defined within the context of a model, is indicated by the shading (white, gray, or black.) To determine subpopulation connectivity, as we have defined it, one must be able to distinguish the right configuration from the left two configurations. In both of the left configurations (but not in the right configuration), there is a connection from a neuron within neuron two's subpopulation (white) onto a neuron within neuron one's subpopulation (gray). Hence, we do not need to distinguish the left two configurations from each other in order to determine subpopulation connectivity.*

the two measured neurons are ambiguous because these measures cannot distinguish a direct connection from common input.

The joint statistics of the two spike trains alone may be insufficient to distinguish a direct connection from common input. If one could measure the neurons inducing the common input effects, then the joint statistics of all the measured spike trains would be sufficient, and one could use analysis tools such as partial coherence [15] to distinguish a direct connection from common input. However, when one cannot measure all possible sources of common input, one cannot rule out common input through partial coherence.

Our approach is to analyze the joint statistics, not just of the measured spike trains, but also of an experimentally controlled stimulus. The idea motivating this approach is that the joint stimulus-spike statistics may be sufficient to distinguish the direct connection configuration from the common input configuration even if the neurons inducing the common input are unmeasured.

It turns out that we cannot distinguish a direct connection from all possible cases of common input. Instead, we can characterize connectivity only in terms of certain *subpopulations* of neurons, defined so that each neuron in a subpopulation responds to the stimulus in a similar manner (the definition of responding "similarly" is made in the context of a model). The concept of subpopulation connectivity is illustrated in Figure 1. Imagine that the spikes of neuron one are correlated with a delayed version of the spikes of neuron two, consistent with a direct connection from neuron two onto neuron one. Our central result is that we can distinguish between (A) a direct connection from neuron two onto neuron one and (B) common input that does not originate from neuron two's subpopulation. On the other hand, if the common input does originate from neuron two's subpopulation, the common input may not be distinguishable from a direct connection. However, in this latter case, the common input does contain a connection from neuron two's subpopulation onto neuron one. Consequently, the identification of a direct connection from neuron two onto neuron one must be interpreted as the identification of a connection from neuron two's subpopulation onto neuron one. The precise identity of the neuron originating the connection remains ambiguous (it could be neuron two or another neuron in neuron two's subpopulation). To summarize this ambiguity, we say we can determine connectivity only at the level of subpopulations (and not at the level of individual neurons).

Our analysis is fundamentally model-driven. The structure imposed by an explicit

model gives the framework necessary for making the subtle distinction between a direct connection and most cases of common input. In this paper, we analyze a network modeled as interacting linear-nonlinear systems responding to a white noise stimulus. Clearly, this choice limits the applicability of this implementation to networks that can be approximated by this simple model. Our motivation for using this model is the ability to compute analytic expressions for necessary stimulus-spike statistics. We mention possible generalizations in the Discussion.

In section 2, we describe the model network and the assumptions required for the analysis. In section 3, we derive analytic expressions for measurable stimulus-spike statistics and solve the resulting system of equations for the coupling strength. We test our findings via simulations in section 4, and discuss the results in section 5.

2. The model.

2.1. The model network. We base our analysis on a model network of linear-nonlinear neurons that builds on the models we have presented previously [12, 10, 11, 9]. Let n be the (presumably unknown) number of neurons in the network. Let the random vector[1] $\mathbf{X}$ denote the stimulus. The components of $\mathbf{X}$ represent the spatio-temporal sequence of stimulus values, such as the pixel values for each refresh of a computer monitor.

The response of neuron $q = 1, 2, \ldots, n$ will depend on the convolution of the stimulus with a spatio-temporal kernel[2] $\bar{\mathbf{h}}_q$, normalized so that $\|\bar{\mathbf{h}}_q\| = 1$. To make later notation simpler, we view the kernel $\mathbf{h}_q$ as sliding along the stimulus with time, and denote by $\bar{\mathbf{h}}_q^i$ the kernel shifted for the discrete time point i. We implicitly view the temporal index of the stimulus as going backward in time, and write the convolution of the kernel with the stimulus as the dot product $\bar{\mathbf{h}}_q^i \cdot \mathbf{X}$.

Let the binary vector $\mathbf{R}$ represent the spike times of neurons in the network. A component $R_q^i = 1$ indicates that neuron q spiked at time i; otherwise, $R_q^i = 0$. When neuron p spikes, the probability that neuron q spikes j time steps later is modified by the connectivity factor $\bar{W}_{pq}^j$. The quantity $\bar{W}_{pq}^j$ is simply added to the convolution $\bar{\mathbf{h}}_q^i \cdot \mathbf{X}$.

The only nonlinear part of the linear-nonlinear model is that the above linear sum is composed with a static monotonically increasing nonlinearity $\bar{g}_q(\cdot)$. This output nonlinearity represents, for example, the neuron's spike generating mechanism and ensures that spiking probabilities stay between zero and one. The resulting linear-nonlinear network model is the following expression for the probability of a spike of neuron q at time i, conditioned on the stimulus and previous spikes (denoted $\mathbf{R}^{<i}$):

$$\Pr\left(R_q^i = 1 \middle| \mathbf{X} = \mathbf{x}, \mathbf{R}^{<i} = \mathbf{r}^{<i}\right) = \bar{g}_q\left(\bar{\mathbf{h}}_q^i \cdot \mathbf{x} + \sum_{p=1}^{n} \sum_{j>0} \bar{W}_{pq}^j r_p^{i-j}\right). \tag{2.1}$$

We let the recent stimulus $\mathbf{X}$ be a discrete approximation to temporal or spatio-temporal Gaussian white noise. For the analysis, we need to estimate stimulus-spike statistics conditioned on the stimulus. The estimation of these statistics implicitly assumes that we repeat each realization of the white noise stimulus multiple times.

[1] With the exceptions of $\bar{W}$, we will use capital variables to denote random quantities.

[2] We use overbars (e.g., $\bar{\mathbf{h}}$) to indicate original model parameters, and will remove the bars (e.g., $\mathbf{h}$) to indicate their estimates from data. In addition, we use subscripts to denote neuron index, and superscripts to denote temporal indices.

2.2. The weak coupling assumption. To facilitate our analysis, we make a weak coupling assumption, which asserts that the coupling $\bar{W}_{pq}^j$ is sufficiently small to justify a second-order approximation in $\bar{W}$. This assumption is really an assumption on how $\bar{W}$ scales with the number of neurons n. As one expands equations such as (2.1) in powers of $\bar{W}$, one obtains terms that are kth-order in $\bar{W}$ summed over the population k times. Hence, one obtains terms of the magnitude $(n\langle\bar{W}\rangle)^k$, where $\langle\bar{W}\rangle$ is an average of n values of $\bar{W}$. To truncate this series at finite k, one at minimum needs $n\langle\bar{W}\rangle < 1$. For a densely coupled large network, the coupling strength must, on average, scale at most like $1/n$. (Individual connections could be stronger, as long as the average scales like $1/n$.) We compute an approximation of order $k = 2$, and we ignore all terms that are third-order or higher in $\bar{W}$.

In our analysis, we go one step further. We ignore all second-order terms that are not summed over the population. Since, in this case, we are not summing over the population, it is no longer a scaling argument. This approximation simply asserts than any one connection cannot be too large. We will use $\approx$ to indicate equality within this modified second-order approximation in $\bar{W}$.

We use this approximation out of necessity, not because we believe it is justified by the biology. However, we demonstrate with simulations that the results often still hold even for larger coupling than needed for the analytic results.

2.3. Effective uncoupled neuron model. Our first step is to fit the spikes of each neuron separately to an uncoupled linear-nonlinear model of the form[3]

$$\Pr\big(R_q^i = 1\big|\mathbf{X} = \mathbf{x}\big) = g_q\big(\mathbf{h}_q^i \cdot \mathbf{x}\big), \tag{2.2}$$

where $\|\mathbf{h}_q^i\| = 1$. For the purpose of subpopulation definitions, below, we imagine we can do this for all neurons. In practice, of course, we can fit uncoupled models only to the two measured neurons. Fitting the uncoupled model (2.2) when the spikes were actually generated by the network model (2.1) defines the effective nonlinearities $g_q(\cdot)$ and kernels $\mathbf{h}_q^i$.

We derive expressions for the effective parameters in terms of the original parameters plus coupling effects. We simply need to calculate $\Pr(R_q^i = 1\big|\mathbf{X} = \mathbf{x})$ from the network model (2.1). Because we assume a second-order approximation in coupling strength $\bar{W}$, it turns out that a first-order approximation in $\bar{W}$ is sufficient for the effective single-neuron parameters.[4]

From a trivial generalization of the calculation in Appendix A.1 of [11], we can average the network model (2.1) over all spikes before time i to conclude that the probability of a spike at time i is

$$\Pr\big(R_q^i = 1\big|\mathbf{X} = \mathbf{x}\big) = \bar{g}_q(\bar{\mathbf{h}}_1^i \cdot \mathbf{x}) + \sum_{p=1}^{n}\sum_{j>0} \bar{W}_{pq}^j \bar{g}_q'(\bar{\mathbf{h}}_1^i \cdot \mathbf{x})\bar{g}_p(\bar{\mathbf{h}}_p^{i-j} \cdot \mathbf{x}) + O(\bar{W}^2). \tag{2.3}$$

Combining this expression with the uncoupled model (2.2), we obtain the following relationship between the effective kernels $\mathbf{h}_q^i$ and nonlinearities $g_q(\cdot)$, on one hand,

[3]Note the absence of bars to indicate effective parameters that can be estimated from data (at least for the measured neurons).

[4]We will show that all terms for the spike-pair statistics will be first- or second-order in $\bar{W}$ (all zero-order terms cancel out), and thus approximating the single-neuron parameters to first order is sufficient to retain a second-order approximation for the spike-pair statistics.

and the original model kernels $\bar{\mathbf{h}}_q^i$ and nonlinearities $\bar{g}_q(\cdot)$, on the other hand:

$$g_q(\mathbf{h}_q^i \cdot \mathbf{x}) = \bar{g}_q(\bar{\mathbf{h}}_q^i \cdot \mathbf{x}) + \sum_{p=1}^{n} \sum_{j>0} \bar{W}_{pq}^j \bar{g}_q'(\bar{\mathbf{h}}_q^i \cdot \mathbf{x}) \bar{g}_p(\bar{\mathbf{h}}_p^{i-j} \cdot \mathbf{x}) + O(\bar{W}^2). \tag{2.4}$$

Since $g_q(\mathbf{h}_q^i \cdot \mathbf{x})$ and $\bar{g}_q(\bar{\mathbf{h}}_a^i \cdot \mathbf{x})$ differ by only a first-order correction, and we are computing only to first order, we can simply erase the bars from $\bar{g}$ and $\bar{\mathbf{h}}$ in the $\bar{W}_{pq}^j$ term (creating a second-order error) to obtain

$$\bar{g}_q(\bar{\mathbf{h}}_q^i \cdot \mathbf{x}) = g_q(\mathbf{h}_q^i \cdot \mathbf{x}) - \sum_{p=1}^{n} \sum_{j>0} \bar{W}_{pq}^j g_q'(\mathbf{h}_q^i \cdot \mathbf{x}) g_p(\mathbf{h}_p^{i-j} \cdot \mathbf{x}) + O(\bar{W}^2). \tag{2.5}$$

This effective parameter relationship will be used in the following analysis to express all equations in terms of the effective parameters.

2.4. Subpopulation definition. A subpopulation is a group of neurons that respond to the stimulus in a similar manner. The effective kernel $\mathbf{h}$ derived from fitting a neuron's spikes to the uncoupled model (2.2) describes the relationship of neuronal spikes to the stimulus. (In some contexts, this kernel would be referred to as the neuron's *receptive field*.) We base our subpopulation definitions on this effective kernel. We define the similarity between two neurons based on the correlation coefficient between the linear components from the uncoupled model (2.2):

$$cc_{pq}^k = \mathrm{cor}(\mathbf{h}_p^i \cdot \mathbf{X}, \mathbf{h}_q^{i-k} \cdot \mathbf{X}), \tag{2.6}$$

where

$$\mathrm{cor}(A, B) = \frac{\mathrm{cov}(A, B)}{\sqrt{\mathrm{var}(A)\,\mathrm{var}(B)}}.$$

Note that $-1 \le cc_{pq}^k \le 1$. In fact, since each component of $\mathbf{X}$ is a unit normal random variable, the correlation coefficient is simply the cosine of the angle between the kernels:

$$cc_{pq}^k = \frac{\mathbf{h}_p^i \cdot \mathbf{h}_q^{i-k}}{\|\mathbf{h}_p^i\| \|\mathbf{h}_q^{i-k}\|} = \mathbf{h}_p^i \cdot \mathbf{h}_q^{i-k}. \tag{2.7}$$

(The last equality results because the kernels are normalized to be unit vectors.)

Define the maximum correlation coefficient as

$$cc_{pq}^{\max} = \max_k cc_{pq}^k. \tag{2.8}$$

If $cc_{pq}^{\max}$ is large, then neurons p and q respond to the stimulus similarly, and we consider the neurons as part of the same subpopulation. On the other hand, if $cc_{pq}^{\max}$ is small, then we consider the neurons as parts of different subpopulations. For the analysis, when we assume that neurons p and q are from different subpopulations, we will effectively assume that each cc_{pq}^k is $O(\bar{W})$. We show via simulations that, in practice, we can relax this condition somewhat.

3. The analysis.

3.1. Overview of the analysis. We assume we have access to only the spikes of neuron one and two (R_1^i and R_2^i) as well as the discrete white noise stimulus $\mathbf{X}$. Given the stimulus, the probability of spikes from the network is specified by the network model (2.1). We initially assume that all unmeasured neurons (with index $p > 2$) are from different subpopulations than those of neurons one or two (in particular, that cc_{p1}^k and cc_{p2}^k for all k are $O(\bar{W})$). Using this assumption, we can solve for the direct connection ($\bar{W}_{21}^j$ and $\bar{W}_{12}^j$) in terms of the joint statistics of the random variables R_1^i, R_2^i, and $\mathbf{X}$.

When we allow unmeasured neurons from the same subpopulations as the measured neurons, we do not change the algorithm to determine a direct connection. We demonstrate that, with this algorithm, common input from neuron one's subpopulation may be identified as a direct connection from neuron one onto neuron two. Similarly, common input from neuron two's subpopulation may be identified as a direct connection from neuron two onto neuron one. Although this results in a misidentification at the level of individual neurons, it still accurately identifies connectivity at the level of subpopulations (since, for example, common input from neuron two's subpopulation does contain a connection from neuron two's subpopulation onto neuron one's subpopulation).

For this analysis, we assume that we have an infinite dataset, so we can estimate the expected values of functions of the random variables. Since in practice, we will have much smaller datasets, we must reduce the bias in estimations from finite datasets using a procedure such as that outlined in [12, 11, 8]. We do not address such bias reduction here.

We give a brief overview of the analysis here and give more details of each step in the following sections. In the first step, we analyze the spikes of neuron one and neuron two separately. From their stimulus-spike statistics, we fit uncoupled linear-nonlinear models (2.2) as if we were using standard white noise analysis methods such as those outlined in [12]. This calculation is based on the mean spike rates[5] ($E\{R_1^i\}$ and $E\{R_2^i\}$) and the stimulus-spike correlations ($E\{\mathbf{X}R_1^i\}$ and $E\{\mathbf{X}R_2^i\}$). Since the spike times are really given by the network model (2.1), the effective kernels ($\mathbf{h}_q^i$ and $\mathbf{h}_2^i$) and nonlinearities ($g_1(\cdot)$ and $g_2(\cdot)$) are functions of network model parameters (including coupling and parameters from other neurons), as given by (2.4).

Next, we calculate the spike rates conditioned on a particular realization of the stimulus[6] ($E\{R_1^i|\mathbf{X}\}$ and $E\{R_2^i|\mathbf{X}\}$). These statistics are equivalent to the peristimulus time histogram (PSTH) commonly used in the neuroscience literature.

We look at spike pairs, where neuron two spikes k units of time before neuron one (note that k could be positive or negative). We subtract off the product of the PSTHs from the rate of spike pairs conditioned on the stimulus, forming

$$E\{R_1^i R_2^{i-k}|\mathbf{X}\} - E\{R_1^i|\mathbf{X}\}E\{R_2^{i-k}|\mathbf{X}\}.$$

The result is the JPSTH cast into the notation of the model.

If we take the expected value of the JPSTH over all realizations of the stimulus,

[5] Note that, due to the stationarity of the stimulus and model, many stimulus-spike statistics do not depend on time, despite the notation. The mean rates $E\{R_q^i\}$, for example, do not depend on the time point i.

[6] Here we assume that each realization of the stimulus is repeated multiple times.

we obtain the shuffle-corrected correlogram or covariogram

$$C_{21}^k = E\{R_1^i R_2^{i-k}\} - E\big\{E\{R_1^i|\mathbf{X}\}E\{R_2^{i-k}|\mathbf{X}\}\big\}. \tag{3.1}$$

Here we have used the fact that $E\{E\{R_1^i R_2^{i-k}|\mathbf{X}\}\} = E\{R_1^i R_2^{i-k}\}$. For a given value of k, C_{21}^k is effectively a sum over the diagonal of the JPSTH corresponding to the delay k. From analysis of the network model (2.1), we derive an equation for C_{21}^k in terms of model parameters.

As we argued in the introduction, the covariogram alone is insufficient to distinguish common input from a direct connection. In terms of the model parameters, there are too many unknowns to solve for $\bar{W}_{21}^k$ (or $\bar{W}_{12}^{-k}$ if $k < 0$). To obtain more equations, we combine white noise analysis methods with the JPSTH.

The key of the approach is to calculate the correlation[7] of the JPSTH with the stimulus:

$$D_{21}^{ki} = E\{\mathbf{X}R_1^i R_2^{i-k}\} - E\big\{\mathbf{X}E\{R_1^i|\mathbf{X}\}E\{R_2^{i-k}|\mathbf{X}\}\big\}. \tag{3.2}$$

Note that the stimulus-spike correlations (e.g., $E\{\mathbf{X}R_1^i\}$) were calculated by correlating the stimulus with a binary vector (e.g., the R_1^i). The above statistic D_{21}^{ki} is a correlation of the stimulus not with a spike vector, but with the vector composed of values from the diagonal of the JPSTH corresponding to delay k. This vector is, of course, not binary, but the correlation can be computed nearly identically. For a fixed k, the result D_{21}^{ki} will be a vector of the same dimension as the correlations $E\{\mathbf{X}R_1^i\}$ and $E\{\mathbf{X}R_2^{i-k}\}$ and hence the same dimension as the kernels $\mathbf{h}_1^i$ and $\mathbf{h}_2^{i-k}$.

Consequently, for a given k, we can decompose D_{21}^{ki} into components parallel to the kernels $\mathbf{h}_1^i$ and $\mathbf{h}_2^{i-k}$, calculating the coefficients A_1^k and A_2^k for which

$$D_{21}^{ki} = A_1^k \mathbf{h}_1^i + A_2^k \mathbf{h}_2^{i-k} + \mathbf{O}^{ki}, \tag{3.3}$$

where $\mathbf{O}^{ki}$ is perpendicular to $\mathbf{h}_1^i$ and $\mathbf{h}_2^{i-k}$. By analyzing the network model (2.1), we calculate expressions for A_1^k and A_2^k in terms of model parameters. From C_{21}^k, A_1^k, and A_2^k, we have three equations for each delay k.

If one compares the number of unknown parameters with the number of equations, the situation still looks hopeless. Assume that we calculate the statistics for the delays $k = \pm 1, \pm 2, \ldots, \pm M$, so that we have $2M \times 3 = 6M$ equations.[8] Assume also that the coupling is zero for delays longer than M time units. Then the coupling parameters are $\bar{W}_{pq}^j$ for $j \in \{1, 2, \ldots, M\}$ and $p, q \in \{1, 2, \ldots, n\}$, where n is the (presumably unknown) number of neurons, for a total of Mn^2 parameters.

If the number of neurons is more than two, the system appears vastly underdetermined. This limitation make sense. If we were sufficiently audacious as to claim that we could reconstruct the coupling of the entire network based on measures of just two neurons, our absurdity would be exposed by this reality check. Our goal is simply to estimate the direct connection $\bar{W}_{21}^j$ and $\bar{W}_{12}^j$ between the two measured neurons.

As demonstrated below, if we assume that unmeasured neurons are from different subpopulations than the measured neurons, all the coupling terms involving unmeasured neurons appear in the same combination[9] in all three sets of equations. We denote this combination by $\hat{U}^k$, and refer to $\hat{U}^k$ as the common input contribution[10]

[7] We use the term "correlation" loosely.

[8] We ignore the single-neuron statistics and single-neuron parameters for this rough calculation.

[9] For this overview, we ignore the presence of indirect connections (see section 3.3).

[10] It turns out that $\hat{U}^k$ contains only combinations of coupling terms that correspond to common input.

to delay k. This notation makes it clear there are really only two unknowns per delay k ($\hat{U}^k$ and either $\bar{W}_{21}^k$ or $\bar{W}_{12}^{-k}$, depending on whether k is positive or negative, respectively). We have a total of $2M \times 2 = 4M$ parameters for $6M$ equations.[11] The system is actually overdetermined. Moreover, due to the weak coupling assumption of section 2.2, the system is linear in $\hat{U}^k$, $\bar{W}_{21}^k$, and $\bar{W}_{12}^{-k}$. We can easily solve it via least squares and estimate the direct connection between neurons one and two as well as the effective common input.

This estimate is, of course, valid only when unmeasured neurons are from different subpopulations than the measured neurons. We address the case of unmeasured neurons from the same subpopulations as the measured neurons in section 3.4. There we argue that our estimate of $\bar{W}_{21}^k$ or $\bar{W}_{12}^{-k}$ accurately reconstructs connectivity between the subpopulations of neuron one and neuron two.

3.2. Single-neuron statistics. In the following sections, we present more details of the analysis outlined above. As many of the calculations are long, we present only the key details, referring where possible to similar calculations from previous papers.

For each of the measured neurons $q = 1, 2$, we analyze its spikes R_q^i and the stimulus $\mathbf{X}$ as though the uncoupled model (2.2) held. We view the parameters from the uncoupled model (2.2) as effective parameters that can be estimated from the stimulus-spike statistics. We have already calculated the effective parameter relationship (2.5) that relates effective parameters to the original model parameters.

In terms of the effective parameters, the stimulus-spike correlation is

$$\begin{aligned} E\{\mathbf{X}R_q^i\} &= E\{\mathbf{X}\Pr(R_q^i = 1 \big| \mathbf{X})\} \\ &= E\{\mathbf{X}g_q(\mathbf{h}_q^i \cdot \mathbf{X})\} \\ &= E\{g_q'(\mathbf{h}_q^i \cdot \mathbf{X})\}\mathbf{h}_q^i, \end{aligned} \tag{3.4}$$

and the mean rate is

$$\begin{aligned} E\{R_q^i\} &= E\{\Pr(R_q^i = 1 \big| \mathbf{X})\} \\ &= E\{g_q(\mathbf{h}_q^i \cdot \mathbf{X})\}, \end{aligned} \tag{3.5}$$

where we used the integration-by-parts formula (A.3) to obtain the final expression for $E\{\mathbf{X}R_q^i\}$.

Given (3.4) and the normalization $\|\mathbf{h}_q^i\| = 1$, the effective kernel can be calculated from the stimulus-spike correlation as

$$\mathbf{h}_q^i = \frac{E\{\mathbf{X}R_q^i\}}{\|E\{\mathbf{X}R_q^i\}\|}. \tag{3.6}$$

If we assume a two-parameter family of nonlinear functions for $g_q(\cdot)$, we can calculate those parameters from $E\{R_q^i\}$ and $\|E\{\mathbf{X}R_q^i\}\|$ (see [12]).

3.3. Neuron pair statistics. We repeat the calculations of Appendices A and B of [11], computing terms only up to the modified second-order approximation in $\bar{W}$, described above in section 2.2. To simplify the notation, we define $\bar{W}_{pq}^k = 0$ for $k \leq 0$.

[11] Although we could, in principle, look for higher-order corrections by retaining higher-order terms in $\bar{W}$, the system would not collapse to $4M$ parameters, and we would have to look for more equations.

After long, tedious calculations and use of the effective parameter relationship (2.5), most of the terms cancel out, and we are left with

$$
\begin{aligned}
&\Pr\bigl(R_1^i = 1 \,\&\, R_2^{i-k} = 1 \bigm| \mathbf{X} = \mathbf{x}\bigr) - \Pr(R_1^i = 1 | \mathbf{X} = \mathbf{x}) \Pr(R_2^{i-k} = 1 | \mathbf{X} = \mathbf{x}) \\
&\quad \approx \left[\bar{W}_{21}^k + \sum_{p=3}^{n} \sum_{j>0} \bar{W}_{2p}^{k-j} \bar{W}_{p1}^j g_p'(\mathbf{h}_p^{i-j} \cdot \mathbf{x}) \right] g_1'(\mathbf{h}_1^i \cdot \mathbf{x}) g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{x}) [1 - g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{x})] \\
&\qquad + \left[\bar{W}_{12}^{-k} + \sum_{p=3}^{n} \sum_{j>0} \bar{W}_{1p}^j \bar{W}_{p2}^{-k-j} g_p'(\mathbf{h}_p^{i+j} \cdot \mathbf{x}) \right] g_1(\mathbf{h}_1^i \cdot \mathbf{x}) [1 - g_1(\mathbf{h}_1^i \cdot \mathbf{x})] g_2'(\mathbf{h}_2^{i-k} \cdot \mathbf{x}) \\
&\qquad + \sum_{p=3}^{n} \sum_{j > \max(0,k)} \bar{W}_{p1}^j \bar{W}_{p2}^{j-k} g_1'(\mathbf{h}_1^i \cdot \mathbf{x}) g_2'(\mathbf{h}_2^{i-k} \cdot \mathbf{x}) g_p(\mathbf{h}_p^{i-j} \cdot \mathbf{x}) [1 - g_p(\mathbf{h}_p^{i-j} \cdot \mathbf{x})],
\end{aligned}
\tag{3.7}
$$

where $\approx$ indicates equality within our modified second-order approximation in $\bar{W}$. Note that if $k \leq 0$, then $\bar{W}_{21}^k = 0$ and $\bar{W}_{2p}^{k-j} = 0$, and the first expression in square brackets is zero. On the other hand, if $k \geq 0$, then $\bar{W}_{12}^{-k} = 0$ and $\bar{W}_{p2}^{-k-j} = 0$, and the second expression in square brackets is zero. Consequently, either the first or the second term is zero for any given k.

This expression is the expected value of the JPSTH, given that the stimulus $\mathbf{X} = \mathbf{x}$. Note that the first term is a direct connection with delay k from neuron two to neuron one, combined with an indirect connection through neuron p of total delay k. The second term is a direct connection, combined with an indirect connection, from neuron one to neuron two (of total delay $-k$, which is positive when this term is nonzero). The last term is due to common input from neuron p onto both neuron one and neuron two. (The expression $\bar{W}_{p1}^j \bar{W}_{p2}^{j-k}$ is nonzero only if neuron p is connected to both neuron one and neuron two.)

The covariogram (3.1) is the expected value of the JPSTH (3.7), and the statistic D (3.2) is the expected value of the JPSTH (3.7) times the stimulus $\mathbf{X}$. Without further assumptions on the unmeasured neurons, we cannot dissociate the contribution of unmeasured neurons from the contribution of measured neurons. In order to solve the equations, we assume that we can factor each expected value into (A) the expected value of an expression involving unmeasured neuron parameters multiplied by (B) the expected value of an expression involving measured neuron parameters. Note that unmeasured neuron parameters appear only in those terms that are second-order in $\bar{W}$. Given our second-order approximation in $\bar{W}$, this step assumes that, to zeroth order in $\bar{W}$, the $g_p(\mathbf{h}_p^{i-j} \cdot \mathbf{X})$ are independent of $g_1(\mathbf{h}_1^i \cdot \mathbf{X})$ and $g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{X})$ (i.e., the effective uncoupled models (2.2) for unmeasured neurons are independent of the effective uncoupled models for measured neurons). In particular, we are assuming that cc_{p1}^j and cc_{p2}^{j-k} are $O(\bar{W})$, which means that the unmeasured neurons are from different subpopulations than the measured neurons (as defined in section 2.4).

Under this assumption the covariogram (3.1) (i.e., the expected value of the JPSTH (3.7)) becomes[12]

$$
\begin{aligned}
C_{21}^k \approx\ & \hat{W}_{21}^k E\{g_1'(\mathbf{h}_1^i \cdot \mathbf{X}) g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{X}) [1 - g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{X})]\} \\
& + \hat{W}_{12}^{-k} E\{g_1(\mathbf{h}_1^i \cdot \mathbf{X}) [1 - g_1(\mathbf{h}_1^i \cdot \mathbf{X})] g_2'(\mathbf{h}_2^{i-k} \cdot \mathbf{X})\} \\
& + \hat{U}_{21}^k E\{g_1'(\mathbf{h}_1^i \cdot \mathbf{X}) g_2'(\mathbf{h}_2^{i-k} \cdot \mathbf{X})\},
\end{aligned}
\tag{3.8}
$$

[12]Recall that stationarity of the stimulus and model imply that the statistics in the equation for C_{21}^k (as well as A_1^k and A_2^k) do not depend on time point i, despite the notation.

where $\approx$ indicates equality within our modified second-order approximation in $\bar{W}$ and

$$\begin{aligned}
\hat{W}_{21}^{k} &= \bar{W}_{21}^{k} + \sum_{p=3}^{n} \sum_{j>0} \bar{W}_{2p}^{k-j} \bar{W}_{p1}^{j} E\{g_p'(\mathbf{h}_p^{i-j} \cdot \mathbf{X})\}, \\
\hat{W}_{12}^{-k} &= \bar{W}_{12}^{-k} + \sum_{p=3}^{n} \sum_{j>0} \bar{W}_{1p}^{j} \bar{W}_{p2}^{-k-j} E\{g_p'(\mathbf{h}_p^{i+j} \cdot \mathbf{X})\}, \\
\hat{U}_{21}^{k} &= \sum_{p=3}^{n} \sum_{j>\max(0,k)} \bar{W}_{p1}^{j} \bar{W}_{p2}^{j-k} E\{g_p(\mathbf{h}_p^{i-j} \cdot \mathbf{X})[1 - g_p(\mathbf{h}_p^{i-j} \cdot \mathbf{X})]\}. \qquad (3.9)
\end{aligned}$$

The new parameters $\hat{W}_{21}^{k}$ and $\hat{W}_{12}^{-k}$ are the effective direct connections between neurons one and two. (By definition, $\hat{W}_{21}^{k} = 0$ for $k \leq 0$ and $\hat{W}_{12}^{-k} = 0$ for $k \geq 0$.) Note that this effective direct connection factor $\hat{W}$ is a combination of both the direct connections and the indirect connections through any unmeasured neuron p. This fact indicates that we cannot distinguish between direct connections and indirect connections through unmeasured neurons. Our goal is to distinguish these effective direct connections $\hat{W}$ from the effective common input $\hat{U}$, which is the sum total effect from all unmeasured neurons p that project to both neuron one and neuron two.

We multiply the JPSTH (3.7) by $\mathbf{X}$, take the expected value, and use the integration-by-parts formula (A.3) to obtain an expression for D_{21}^{ki} (given by (3.2))

$$D_{21}^{ki} \approx A_1^k \mathbf{h}_1^i + A_2^k \mathbf{h}_2^{i-k} + \mathbf{O}^{ki}, \qquad (3.10)$$

where $\mathbf{O}^{ki}$ is an expression that is $O(\bar{W}^2)$ and involves the kernels of the unmeasured neurons. Since we are assuming that the unmeasured neurons are from different subpopulations than the measured neurons, $\mathbf{O}^{ki}$ can be viewed as orthogonal to[13] $\mathbf{h}_1^i$ and $\mathbf{h}_2^{i-k}$. The components of D_{21}^{ki} parallel to $\mathbf{h}_1^i$ and $\mathbf{h}_2^{i-k}$ are

$$\begin{aligned}
A_1^k &= \hat{W}_{21}^{k} E\{g_1''(\mathbf{h}_1^i \cdot \mathbf{X}) g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{X})[1 - g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{X})]\} \\
&\quad + \hat{W}_{12}^{-k} E\{g_1'(\mathbf{h}_1^i \cdot \mathbf{X})[1 - 2g_1(\mathbf{h}_1^i \cdot \mathbf{X})] g_2'(\mathbf{h}_2^{i-k} \cdot \mathbf{X})\} \\
&\quad + \hat{U}_{21}^{k} E\{g_1''(\mathbf{h}_1^i \cdot \mathbf{X}) g_2'(\mathbf{h}_2^{i-k} \cdot \mathbf{X})\}, \\
A_2^k &= \hat{W}_{21}^{k} E\{g_1'(\mathbf{h}_1^i \cdot \mathbf{X}) g_2'(\mathbf{h}_2^{i-k} \cdot \mathbf{X})[1 - 2g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{X})]\} \\
&\quad + \hat{W}_{12}^{-k} E\{g_1(\mathbf{h}_1^i \cdot \mathbf{X})[1 - g_1(\mathbf{h}_1^i \cdot \mathbf{X})] g_2''(\mathbf{h}_2^{i-k} \cdot \mathbf{X})\} \\
&\quad + \hat{U}_{21}^{k} E\{g_1'(\mathbf{h}_1^i \cdot \mathbf{X}) g_2''(\mathbf{h}_2^{i-k} \cdot \mathbf{X})\}. \qquad (3.11)
\end{aligned}$$

From measuring the spikes of neuron one and two (R_1^i and R_2^i) in response to the stimulus $\mathbf{X}$, we can calculate the effective uncoupled model parameters ($g_1(\cdot)$, $g_2(\cdot)$, $\mathbf{h}_1^i$, and $\mathbf{h}_2^{i-k}$), the covariogram C_{21}^k (via (3.1)), the statistic D_{21}^{ki} (via (3.2)), and consequently its components A_1^k and A_2^k. The nine expected values in (3.8) and (3.11) are simply Gaussian integrals of known functions that can be calculated. The only unknown quantities are the $\hat{W}_{12}^{-k}$, $\hat{W}_{21}^{k}$, and $\hat{U}_{21}^{k}$.

To emphasize that the direct connection is simply one variable per delay, we define

[13] Since, for any $p > 2$ and any j, the correlation coefficients cc_{p1}^{j} and cc_{p2}^{j-k} are assumed to be $O(\bar{W})$, the component of $\mathbf{h}_p^{i-j}$ parallel to $\mathbf{h}_1^i$ or $\mathbf{h}_2^{i-k}$ is $O(\bar{W})$. Hence, the component of $\mathbf{O}^{ki}$ parallel to these kernels must be $O(\bar{W}^3)$, which we can ignore.

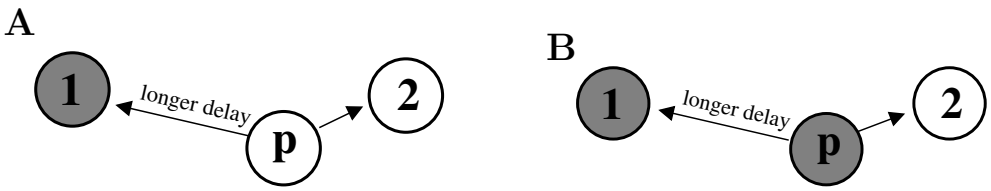


FIG. 2. *Schematic of common input from unmeasured neuron p onto measured neurons one and two. The gray shading indicates which neurons are from the same subpopulation. The connection from neuron p to neuron one is delayed, so the common input introduces a correlation between the spikes of neuron one and two that is similar to the correlation induced by a connection from neuron two to neuron one.* (A) *Neuron p is within neuron two's subpopulation. In this case, the common input may appear as a direct connection from neuron two onto neuron one.* (B) *Neuron p is within neuron one's subpopulation. In this case, the common input will not be confused with a direct connection.*

a new direct connection variable

$$\hat{W}^k = \begin{cases} \hat{W}_{12}^{-k} & \text{for } k < 0, \\ 0 & \text{for } k = 0, \\ \hat{W}_{21}^{k} & \text{for } k > 0. \end{cases} \tag{3.12}$$

Note that the equations for different delays k are uncoupled. For each $k \neq 0$, equations (3.8) and (3.11) are three linear equations for the two unknowns $\hat{U}_{21}^k$ and $\hat{W}^k$. This system is easily solved for the two unknowns using linear least squares. For $k = 0$, the only unknown is $\hat{U}_{21}^0$, for which we solve using (3.8).

3.4. Common input from the measured neurons' subpopulations. To complete the above analysis, we assumed that the unmeasured neurons were from different subpopulations than the measured neurons. In particular, we assumed that the correlation coefficients between the measured neurons and the unmeasured neurons (cc_{p1}^k and cc_{p2}^k for $p > 2$ and all k) were small.

In the brain, one typically finds groups of neurons that respond to a stimulus in a similar way and hence would be from the same subpopulation, as we defined them in section 2.4. Consequently, one would anticipate the presence of unmeasured neurons from the subpopulations of both neuron one and neuron two. Since such measured neurons could be the source of common input, we must address the case of common input from the neurons within the measured neurons' subpopulations. (With the exception of common input and indirect connection effects, all effects of unmeasured neurons had already been canceled in the analysis leading to the JPSTH (3.7), before we made any assumptions about subpopulations.)

We examine networks with a common input configuration where an unmeasured neuron p has a connection onto neuron two, and, with a longer delay, a connection onto neuron one, as schematized in Figure 2. To implement this, we let the connection onto neuron one have a delay of j time steps ($\bar{W}_{p1}^j \neq 0$) and the connection onto neuron two have a delay of $j - k$ time steps ($\bar{W}_{p2}^{j-k} \neq 0$), with $j > k > 0$. With this set of delays, the common input will introduce correlations between the measured neurons that mimic a direct connection from neuron two onto neuron one with a delay of k time steps.

We first show how common input from neuron two's subpopulation (Figure 2(A)) can be misidentified as a direct connection from neuron two onto neuron one. (Note

that this results in a correct identification of a connection from neuron two's subpopulation onto neuron one's subpopulation.) Since the connection from neuron p onto neuron two is due to the term $\bar{W}_{p2}^{j-k} \neq 0$, the correlation coefficient cc_{p2}^{j-k} will determine whether the neuron p acts as a member of neuron two's subpopulation (this inference comes from inspection of the last term in the JPSTH (3.7)). If cc_{p2}^{j-k} is large, neuron p will act as a member of neuron two's subpopulation.

We examine the extreme case where neuron p responds to nearly the same stimulus features as neuron two does $j-k$ time steps later; i.e., $\mathbf{h}_p^{i-j} = \mathbf{h}_2^{i-k} + O(\bar{W})$. Consequently, $cc_{p2}^{j-k} = 1+O(\bar{W})$. Then, the contribution of this neuron to the JPSTH (3.7) is

$$\bar{W}_{p1}^{j}\bar{W}_{p2}^{j-k} g_1'(\mathbf{h}_1^i \cdot \mathbf{x}) g_2'(\mathbf{h}_2^{i-k} \cdot \mathbf{x}) g_p(\mathbf{h}_2^{i-k} \cdot \mathbf{x})[1 - g_p(\mathbf{h}_2^{i-k} \cdot \mathbf{x})].$$

The contribution of neuron p to the covariogram is

$$C_{21}^k = \bar{W}_{p1}^{j}\bar{W}_{p2}^{j-k} E\{g_1'(\mathbf{h}_1^i \cdot \mathbf{X}) g_2'(\mathbf{h}_2^{i-k} \cdot \mathbf{X}) g_p(\mathbf{h}_2^{i-k} \cdot \mathbf{X})[1 - g_p(\mathbf{h}_2^{i-k} \cdot \mathbf{X})]\}. \tag{3.13}$$

Since the kernel $\mathbf{h}_p^{i-j}$ is identical to $\mathbf{h}_2^{i-k}$, an additional term will appear in A_2^k after the integration by parts, so that the contribution of neuron p to A_1^k and A_2^k is

$$\begin{aligned} A_1^k &= \bar{W}_{p1}^{j}\bar{W}_{p2}^{j-k} E\{g_1''(\mathbf{h}_1^i \cdot \mathbf{X}) g_2'(\mathbf{h}_2^{i-k} \cdot \mathbf{X}) g_p(\mathbf{h}_2^{i-k} \cdot \mathbf{X})[1 - g_p(\mathbf{h}_2^{i-k} \cdot \mathbf{X})]\}, \\ A_2^k &= \bar{W}_{p1}^{j}\bar{W}_{p2}^{j-k} E\{g_1'(\mathbf{h}_1^i \cdot \mathbf{X}) g_2''(\mathbf{h}_2^{i-k} \cdot \mathbf{X}) g_p(\mathbf{h}_2^{i-k} \cdot \mathbf{X})[1 - g_p(\mathbf{h}_2^{i-k} \cdot \mathbf{X})]\} \\ &\quad + \bar{W}_{p1}^{j}\bar{W}_{p2}^{j-k} E\{g_1'(\mathbf{h}_1^i \cdot \mathbf{X}) g_2'(\mathbf{h}_2^{i-k} \cdot \mathbf{X}) g_p'(\mathbf{h}_2^{i-k} \cdot \mathbf{X})[1 - 2g_p(\mathbf{h}_2^{i-k} \cdot \mathbf{X})]\}. \end{aligned} \tag{3.14}$$

Compare this contribution to the effect of a direct connection from neuron two to neuron one at a delay of k units of time (the $\bar{W}_{21}^k$ terms from (3.8) and (3.11)):

$$\begin{aligned} C_{21}^k &= \hat{W}_{21}^k E\{g_1'(\mathbf{h}_1^i \cdot \mathbf{X}) g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{X})[1 - g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{X})]\}, \\ A_1^k &= \hat{W}_{21}^k E\{g_1''(\mathbf{h}_1^i \cdot \mathbf{X}) g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{X})[1 - g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{X})]\}, \\ A_2^k &= \hat{W}_{21}^k E\{g_1'(\mathbf{h}_1^i \cdot \mathbf{X}) g_2'(\mathbf{h}_2^{i-k} \cdot \mathbf{X})[1 - 2g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{X})]\}. \end{aligned} \tag{3.15}$$

Ignoring the first term of A_2^k in (3.14), we observe that the relationship among C_{21}^k, A_1^k, and A_2^k in (3.13) and (3.14) is nearly identical to their relationship in (3.15). For the case $g_p = g_2$, the only difference is the additional common factor of $g_2'(\mathbf{h}_2^{i-k} \cdot \mathbf{X})$ in the expected value.

If the second term of A_2^k in (3.14) does dominate the first term, then the common input from neuron two's subpopulation leads to a relationship among the statistics C_{21}^k, A_1^k, and A_2^k that mimics a direct connection from neuron two to neuron one. Consequently, we would expect that applying the results of section 3.3 would indicate the presence of a direct connection. Simulations confirm that the second term of A_2^k in (3.14) does indeed dominate, as network configurations such as Figure 2(A) are categorized as direct connection configurations.

We next consider the case where neuron p is from neuron one's subpopulation (Figure 2(B)). If neuron p responds to the stimulus almost exactly as neuron one does, j time steps later, the analysis does not give a clear answer. If we assume $\mathbf{h}_p^{i-j} = \mathbf{h}_1^i + O(\bar{W})$ so that $cc_{p1}^j = 1 + O(\bar{W})$, we do not obtain a relationship among C_{21}^k, A_1^k, and A_2^k that mimics their relationship in (3.15). In this case, simulations indicate that this configuration appears as common input. (Of course, if the connection from

neuron p to neuron two had the longer delay, the network would be equivalent to Figure 2(A) with the roles of neuron one and two reversed. In this case, the network would mimic a direct connection from neuron one to neuron two.)

We conclude that we cannot distinguish between a direct connection from neuron two onto neuron one and common input from neuron two's subpopulation. (An equivalent statement holds with the roles of neuron one and neuron two reversed.) Since common input from neuron two's subpopulation does contain a connection from neuron two's subpopulation onto neuron one, we conclude that the connectivity is correctly identified at the level of subpopulations. In applications where the distinction among particular neurons within a subpopulation is unimportant, the ambiguity in the precise identification of the connection is not problematic. See the Discussion for more details.

4. Tests via simulation.

4.1. Simulations of small linear-nonlinear networks. We tested our analytic results with simulations of networks of linear-nonlinear neurons given by (2.1). We used kernels $\bar{\mathbf{h}}_q^i$ that capture some features of the responses of neurons in visual cortex [6]. For spatial grid point $\mathbf{j} = (j_1, j_2)$ and time t, the kernels were of the form

$$\bar{h}_q(\mathbf{j}, t) = (t - b_q) \exp\left(-\frac{t - b_q}{\tau_h} - \frac{|\mathbf{j}|^2}{10}\right) \sin((j_1 \cos\psi_q + j_2 \sin\psi_q) f_q + \phi_q) \tag{4.1}$$

for $t > b_q$ and $\bar{h}_q(\mathbf{j}, t) = 0$ otherwise. We sampled $\bar{h}_q(\mathbf{j}, t)$ on a $10 \times 10 \times 10$ grid and normalized it to the unit vector $\bar{\mathbf{h}}_q^i$. For the analysis, the only important geometry is the inner product between the kernels, $\bar{\mathbf{h}}_q^i \cdot \bar{\mathbf{h}}_p^{i-j}$, which is the correlation between normal random variables $\bar{\mathbf{h}}_p^{i-j} \cdot \mathbf{X}$ and $\bar{\mathbf{h}}_q^i \cdot \mathbf{X}$.

For each example, we simulated the network response to 500,000 units of time, adjusting the nonlinearities $\bar{g}_q(\cdot)$ to obtain between 10,000 and 15,000 spikes from each neuron. Each simulation was composed of 100 trials, each lasting 5,000 units of time. For ten trials, the stimulus was independent realizations of the Gaussian white noise. We repeated each realization ten times to form the 100 trials. The repetitions allowed estimation of the spiking probabilities $\Pr(R_1^i | \mathbf{X} = \mathbf{x})$ and $\Pr(R_2^{i-k} | \mathbf{X} = \mathbf{x})$ needed for the JPSTH (3.7) by averaging over the ten repetitions (equivalent to a shuffle correction).

The analysis was based on expected values of stimulus-spike statistics. Naive estimates of products of these statistics, including the shuffle correction, from finite datasets can be highly biased. We reduced these biases using techniques described elsewhere [12, 11, 8]. From the independent trials, we estimated confidence intervals as described in Appendix C.

To compute the Gaussian integrals in (3.8) and (3.11), we needed to choose a form for the nonlinear functions $g_q(y)$. To allow us to compute the integrals analytically, we assume that the nonlinear functions are error functions of the form

$$\bar{g}_q(y) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{y - \bar{y}_q}{\bar{\epsilon}_q \sqrt{2}}\right)\right], \tag{4.2}$$

where $\bar{y}_q$ is the threshold, $\bar{\epsilon}_q$ defines the steepness of the nonlinearity, and the error function is $\operatorname{erf}(y) = \frac{2}{\sqrt{\pi}} \int_0^y e^{-t^2} dt$. Note that $\lim_{y \to \infty} g_q(y) = 1$ and $\lim_{y \to -\infty} g_q(y) = 0$. The expressions for C_{21}^k, A_1^i, and A_2^{i-k} for the case of an error function nonlinearity are given in Appendix B. We demonstrate below that the results are not sensitive to

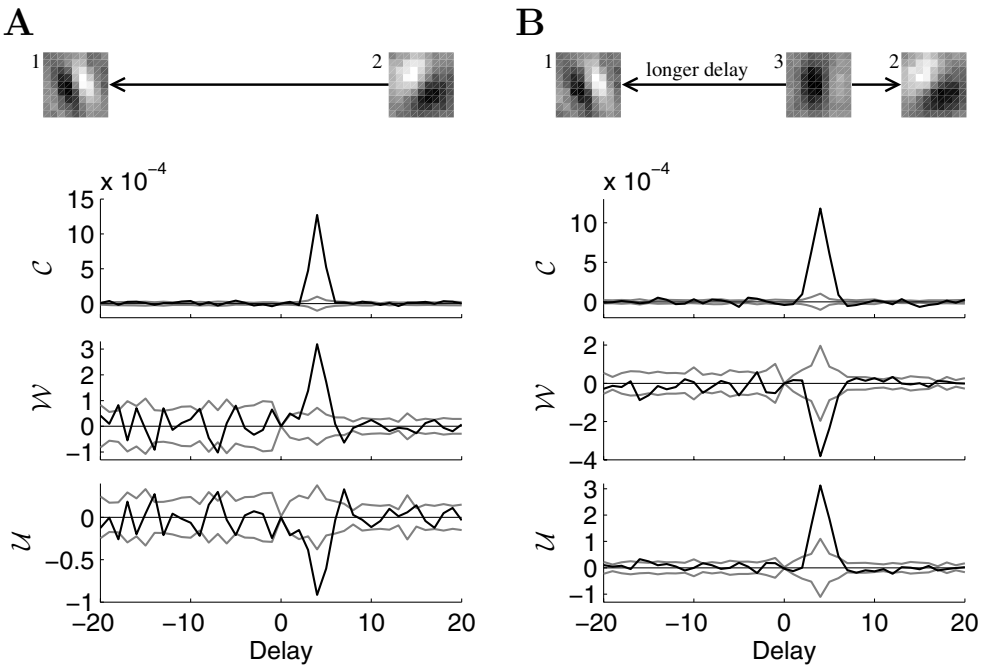


FIG. 3. *Results from sample simulations showing successful distinction between* (A) *the direct connection configuration and* (B) *the common input configuration, using only the spike times from neurons one and two. The top panels are schematics of the network architecture, with grayscale plots of the time slice in which each kernel (*$\mathbf{h}_1$*,* $\mathbf{h}_2$*, or* $\mathbf{h}_3$*, labeled by neuron number) reached its maximal value. (The spikes of neuron three were used to calculate* $\mathbf{h}_3$ *for this illustration. The remaining analysis used only the spikes of neuron one and two.) The bottom three panels plot with black lines the covariogram* $\mathcal{C}$*, direct connection measure* $\mathcal{W}$*, and common input measure* $\mathcal{U}$*. The gray lines estimate confidence intervals of one standard error. Delay is the spike time of neuron one minus the spike time of neuron two.* (A) *The direct connection from neuron two onto neuron one creates a positive covariogram* $\mathcal{C}$ *around a delay of four units of time. The significantly positive direct connection measure* $\mathcal{W}$ *at that delay indicates that the correlation was due to a direct connection. The negative common input measure* $\mathcal{U}$*, though indicating departure from the weak coupling assumption, does not confuse the direct connection interpretation. The direct connection was given by* $\bar{W}^4_{21} = 0.8$*,* $\bar{W}^3_{21} = \bar{W}^5_{21} = 0.4$*. (All other* $\bar{W}$ *were zero.) Parameters used:* $\tau_h = 2$*,* $\psi_1 = \pi/8$*,* $\psi_2 = -\pi/4$*,* $\phi_1 = 0$*,* $\phi_2 = \pi$*,* $f_1 = 1.0$*,* $f_2 = 0.3$*,* $b_1 = b_2 = 0$*,* $\bar{T}_1 = 2.3$*,* $\bar{T}_2 = 2.8$*,* $\bar{\epsilon}_1 = 0.5$*,* $\bar{\epsilon}_2 = 1.0$*.* (B) *For the network with common input from unmeasured neuron three, the covariogram* $\mathcal{C}$ *is nearly identical to the direct connection case from panel* A*. The fact that the correlation was due to common input is revealed by the positive* $\mathcal{U}$ *(and negative* $\mathcal{W}$*). The common input to neuron one was delayed four more units of time compared with that to neuron two:* $\bar{W}^6_{31} = \bar{W}^2_{32} = 1.8$*,* $\bar{W}^5_{31} = \bar{W}^7_{31} = \bar{W}^1_{32} = \bar{W}^3_{32} = 0.8$*. Parameters as in* (A) *except* $\psi_3 = 0$*,* $\phi_3 = -\pi/3$*,* $f_3 = 0.6$*,* $b_3 = 0$*,* $\bar{T}_1 = 2.6$*,* $\bar{T}_2 = 3.0$*,* $\bar{T}_3 = 2.4$*,* $\bar{\epsilon}_3 = 0.7$*.*

this particular choice of nonlinear function. One could perform a similar analysis for other nonlinear functions, although then one would presumably need to compute the integrals numerically.

We denote by $\mathcal{C}^k$ the covariogram C^k_{21} (3.1) estimated from a dataset. Similarly, we denote by $\mathcal{W}^k$ and $\mathcal{U}^k$ estimates of the direct connection $\hat{W}^k$ (3.12) and common input $\hat{U}^k_{21}$ (3.9), respectively.

To illustrate the method, we looked at minimal networks containing two or three neurons. First, we simulated a pair of neurons, where neuron two has a direct connection onto neuron one. The results are shown in Figure 3(A). The covariogram $\mathcal{C}$ shows

a peak at the delay corresponding to the connection. However, the covariogram does not indicate whether this spike correlation is due to a direct connection or common input from an unmeasured neuron.

This ambiguity is resolved by the measures $\mathcal{W}$ and $\mathcal{U}$. The direct connection measure $\mathcal{W}$ is significantly positive at the delay corresponding to the connection. On the other hand, the common input measure $\mathcal{U}$ is negative at that delay. Hence, $\mathcal{W}$ and $\mathcal{U}$ indicate that the spike correlation was caused by a direct connection rather than by a common input.

Note that the noise in $\mathcal{W}$ and $\mathcal{U}$ is dramatically greater than in the covariogram $\mathcal{C}$. This increase is presumably due to the subtlety of the distinction we are attempting to make. For this reason, we required long simulations with up to 15,000 spikes to obtain good results.

The reciprocal behavior observed between $\mathcal{W}$ and $\mathcal{U}$ is not predicted by the analysis. According to the analysis, $\mathcal{U}$ should be flat in the presence of a direct connection. The fact that $\mathcal{U}$ is negative is surprising. Simulations indicate that this behavior is a result of a breakdown in the weak coupling assumption. For a weaker direct connection (and much longer simulation to compensate for noise), the reflection in $\mathcal{U}$ disappears (not shown). The combination of a positive $\mathcal{W}$ and a negative $\mathcal{U}$ could theoretically be caused by either a positive direct connection or a negative common input. The ambiguity is removed by the positive $\mathcal{C}$, indicating that we indeed have a positive direct connection.

We next simulated three neurons, where the unmeasured neuron three was a source of common input to neurons one and two. In this example, neuron three was from a different subpopulation than neuron one or two, as defined in section 2.4. Figure 3(B) shows the results obtained from analyzing the spikes of neurons one and two. The covariogram $\mathcal{C}$ is essentially identical to the direct connection case in Figure 3(A). The covariogram cannot be used to distinguish this common input configuration from the direct connection configuration. This distinction can be made from the measures $\mathcal{W}$ and $\mathcal{U}$. In this case, the common input measure $\mathcal{U}$ is significantly positive, while the direct connection measure $\mathcal{W}$ is negative. Since $\mathcal{C}$ is positive, this combination correctly indicates the common input configuration.

We demonstrate in Figure 4 two simulations to confirm our analysis, with common input from neurons within the subpopulation of neuron one or neuron two (section 3.4). We retain the same common input configuration of Figure 3(B), but change the kernel of the unmeasured neuron three to match the kernel of either neuron two or neuron one.

For the case when the unmeasured neuron three was in neuron two's subpopulation (Figure 4(A)), the common input appears as a direct connection from neuron two onto neuron one, because $\mathcal{W}$ is significantly positive. The results fail to correctly identify that neuron two does not have a direct connection onto neuron one. If one cares about the distinction between neuron two and neuron three, then this result is unacceptable. If, on the other hand, the precise of identity of neurons within a subpopulation is unimportant, the results are adequate, as they correctly indicate that a neuron from neuron two's subpopulation has a direct connection onto neuron one.

For the case when the unmeasured neuron three was in neuron one's subpopulation (Figure 4(B)), the results correctly indicate the common input configuration. In this network, there was no direct connection from neuron two's subpopulation onto neuron one's subpopulation. Fortunately, the similarity between neuron three and neuron one does not affect the results.

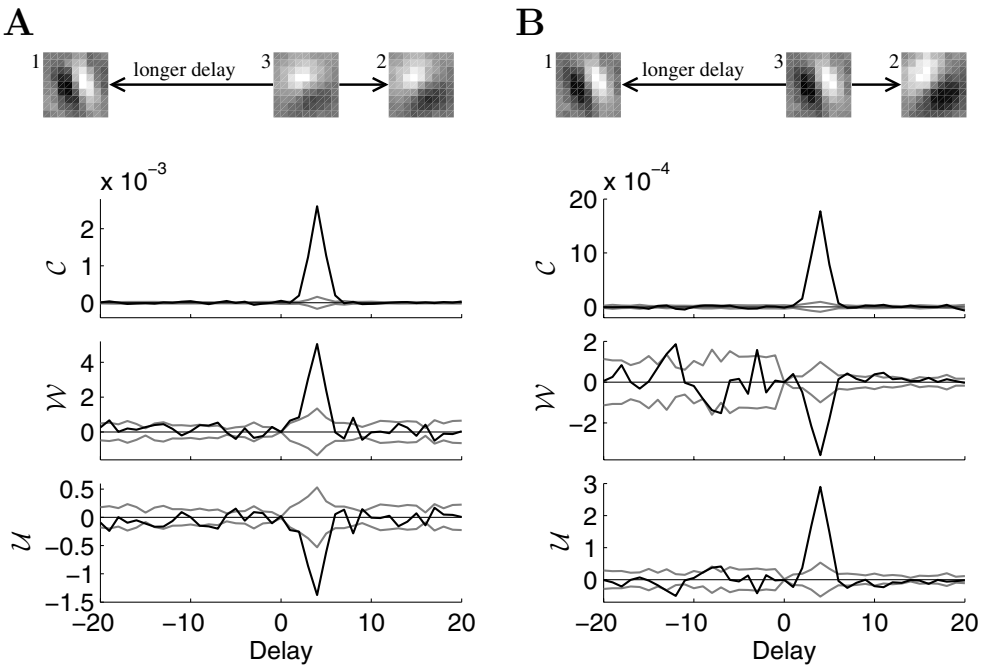


FIG. 4. *Tests of the effect of common input from an unmeasured neuron within the subpopulation of neuron one or neuron two. Panels as in Figure* 3. (A) *The network configuration is identical to the common input of Figure* 3(B) *except that the kernel of neuron three is similar to that of neuron two* ($cc_{32}^{\max} > 0.9$*; see* (2.8)*), so that neuron three is in neuron two's subpopulation. In this case, the common input is misidentified as a direct connection* ($\mathcal{W}$ *is positive).* $\mathcal{W}$ *can be interpreted as indicating a direct connection from a neuron within neuron two's subpopulation onto neuron one. The connectivity* $\bar{W}$ *and the parameters are identical to those of Figure* 3(B) *except* $b_2 = 2$, $\psi_3 = -3\pi/8$, $\phi_3 = 7\pi/8$, $f_3 = 0.4$, $\bar{T}_2 = 3.4$. (B) *When the kernel of neuron three is similar to that of neuron one* ($cc_{31}^{\max} > 0.9$*) so that neuron three is in neuron one's subpopulation, the common input is correctly identified* ($\mathcal{U}$ *is positive). In this case, there is no connection from neuron two's subpopulation to neuron one. The connectivity* $\bar{W}$ *and the parameters are identical to those of Figure* 3(B) *except* $b_1 = 6$, $\psi_3 = \pi/8$, $\phi_3 = 0.0$, $f_3 = 0.8$, $\bar{T}_1 = 3.0$.

As indicated by the analysis, our approach cannot distinguish a direct connection from an indirect connection via a third intermediate neuron. An example of an indirect connection is shown in Figure 5(A). Since the direct connection measure $\mathcal{W}$ is positive, the indirect connection is classified as a direct connection.

Although the analysis was based on an error function nonlinearity (4.2), the results are not sensitive to small changes in nonlinearity shape. In Figure 5, we demonstrate a simulation with a (truncated) power law nonlinearity: $\bar{g}_q(y) = \min\{A_q y^{\beta_q}, 1\}$ for $y > 0$, and $\bar{g}_q(y) = 0$ otherwise. This example includes both a direct connection from neuron one onto neuron two (we use the sign convention where this is a negative delay) and common input from an unmeasured neuron three (with a positive net delay, so that it mimics a connection from neuron two onto neuron one). We analyzed the spike responses from neurons one and two (R_1^i and R_2^i) and the stimulus $\mathbf{X}$ just as we did in the previous examples; i.e., we used the results of Appendix B, in which the nonlinearities $\bar{g}_q(\cdot)$ are assumed to be error functions.

The covariogram contains two similar peaks at a positive and negative delay and therefore cannot distinguish the two type of connections. The measures $\mathcal{W}$ and $\mathcal{U}$

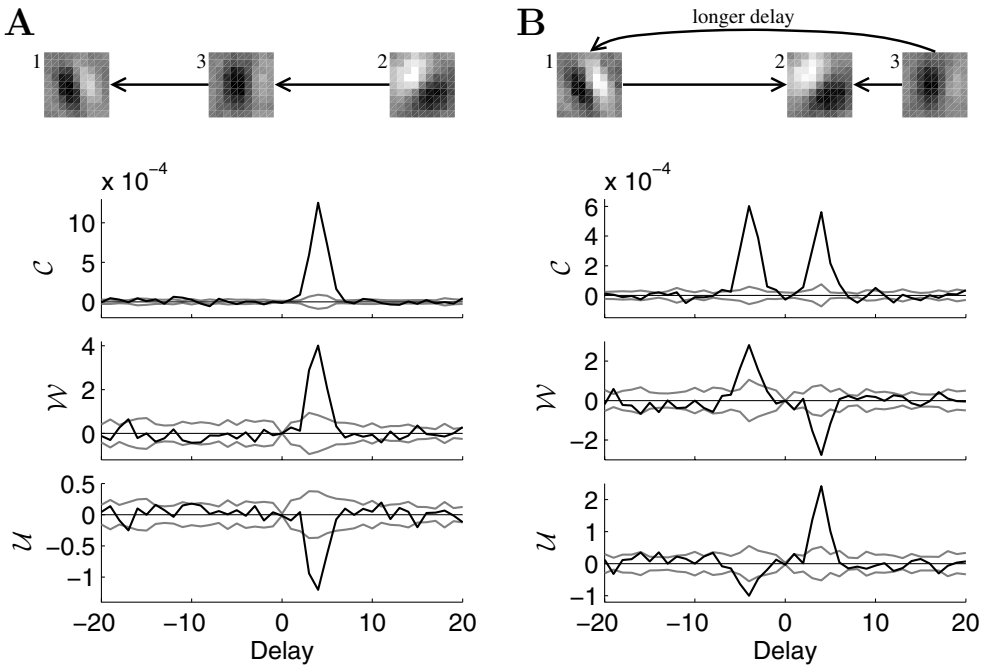


FIG. 5. *Further demonstrations of the approach. Panels as in Figure* 3. (A) *An indirect connection from neuron two onto neuron one through an unmeasured neuron three. This connection appears as a direct connection ($\mathcal{W}$ is positive). As shown by the analysis, we cannot distinguish such an indirect connection from a direct connection. The indirect connection is given by* $\bar{W}^2_{23} = \bar{W}^2_{31} = 1.6$, $\bar{W}^1_{23} = \bar{W}^3_{23} = \bar{W}^1_{31} = \bar{W}^3_{31} = 0.8$. *Parameters are identical to Figure* 3(B) *except* $\bar{T}_1 = 2.5$, $\bar{T}_2 = 2.8$, $\bar{T}_3 = 2.6$. (B) *A simulation with nonlinear functions* $\bar{g}_q(\cdot)$ *given by power laws. In this case, the network contains a direct connection from neuron one onto neuron two (corresponding to a negative delay) and common input from unmeasured neuron three onto neurons one and two (with a longer delay to neuron one to give a positive delay). The spikes of neuron one and two were analyzed as though the nonlinearities were error functions. Although the covariogram contains two virtually identical peaks, the measures $\mathcal{W}$ and $\mathcal{U}$ successfully identify the direct connection at negative delay and the common input at positive delay. The connections were given by* $\bar{W}^4_{12} = 0.6$, $\bar{W}^3_{12} = \bar{W}^5_{12} = 0.4$, $\bar{W}^6_{31} = \bar{W}^2_{32} = 1.8$, $\bar{W}^5_{31} = \bar{W}^7_{31} = \bar{W}^1_{32} = \bar{W}^3_{32} = 0.8$. *Kernel parameters are as in Figure* 3(B). *Power law parameters:* $A_1 = 0.02$, $A_2 = 0.035$, $A_3 = 0.05$, $\beta_1 = 2.6$, $\beta_2 = 2.0$, $\beta_3 = 2.3$.

differentiate between the origins of these peaks. Since $\mathcal{W}$ is positive at a negative delay, it indicates a direct connection from neuron one onto neuron two. On the other hand, since $\mathcal{U}$ is positive at a positive delay, it indicates a common input from a third neuron rather than any direct connection from neuron two onto neuron one. The method correctly identifies the circuitry of the model network even with a power law nonlinearity.

4.2. Simulation of integrate-and-fire networks. To test the robustness of the method to deviations from the linear-nonlinear model, we simulated a system of integrate-and-fire neurons. In this case, we viewed each time step as corresponding to a millisecond. The evolution of the voltage of neuron q in response to input $G_q(t)$ is given by $\tau_m \frac{\mathrm{d}V_q}{\mathrm{d}t} + V_q + G_q(t)(V_q - \mathcal{E}_s) = 0$. The spike times T_q^j of neuron q are those times when $V_q(T_q^j)$ reaches 1. After each spike, the voltage was reset to 0 and held there for an absolute refractory period of length τ_{ref}. Each neuron was driven by the

input conductance $G_q(t)$, which we specified by

$$G_q(t) = 0.05 \sum_{j>0} f(t - T_q^{\text{ext},j}) + \sum_{p=1}^{n} \sum_{j>0} W_{pq} f(t - T_p^j - d_{pq}),$$

where the first term is the response to external input events at times $T_p^{\text{ext},j}$ and the second term is due to internal coupling. The function $f(t) = \frac{e^2}{4}\left(\frac{t}{\tau_s}\right)^2 e^{-t/\tau_s}$ for $t > 0$, and $f(t) = 0$ otherwise. Here, W_{pq} specifies the strength of coupling from neuron p onto neuron q, and d_{pq} is the delay of that connection.

We set the external input to be a linear-nonlinear function of the stimulus. Accordingly, the $T_q^{\text{ext},j}$ were drawn from a modulated Poisson process with rate given by $\alpha_q\left[\mathbf{h}_q^i \cdot \mathbf{X}\right]_+$, where $[x]_+ = x$ if $x > 0$ and is zero otherwise.

We first simulated a network of three neurons that contained both a direct connection from neuron one onto neuron two and common input from neuron three onto neurons one and two (just as in Figure 5(B)). We used the same linear kernels (4.1) as before, sampling them on an $80 \times 10 \times 10$ grid in time and space. For realism, we sampled the white noise stimulus every ten units of time (i.e., every 10 ms). We simulated the network to 5,000 simulated seconds (nearly 1.4 simulated hours), recording 30,000 to 40,000 spikes per neuron. We needed such long simulations to obtain good results.

Figure 6(A) demonstrates that our analysis can distinguish common input from a direct connection even with integrate-and-fire neurons. The results are equivalent to Figure 5(B). The covariogram $\mathcal{C}$ show peaks corresponding to the direct connection and the common input. The source of these correlations is distinguished by the measures $\mathcal{W}$ and $\mathcal{U}$. The correlation at a negative delay is identified as a direct connection from neuron one onto neuron two; the correlation at a positive delay is identified as common input from a third neuron.

Since the integrate-and-fire neurons are driven by the stimulus in a fairly linear fashion, the basic relationship of neural response to the stimulus is similar to that assumed in the linear-nonlinear model (2.1). However, unlike model (2.1), the probability of a spike does depend strongly on previous spike times. The presence of a refractory period prevents the neuron from firing a spike immediately after spiking. Even after the refractory period, the voltage must integrate up to threshold, further increasing dependence among spike times. Figure 6(A) demonstrates that our approach can still work in the presence of these deviations from model assumptions.

As a final test of our approach, we simulated a slightly larger network of 20 integrate-and-fire neurons. The network included a direct connection from neuron one onto neuron two. In addition, four of the unmeasured neurons (neurons 3–20) were randomly selected to give common input onto both neurons one and two, where the connection onto neuron one had a delay that was 30 ms longer than the delay to neuron two. In this case, the measured spike trains had a correlation at a negative delay due to the direct connection and a correlation at a positive delay due to the common input, just as in the previous example. We randomly added additional connections to the network so that any given neuron had a 10% chance of connecting onto any given unmeasured neuron.

We simulated this network to 5,000 simulated seconds (nearly 1.4 simulated hours), measuring approximately 10,000–40,000 spikes per neuron. We discarded the spikes of all neurons except neurons one and two. The results from analyzing just these spikes are shown in Figure 6(B). In this case, the direct connection and

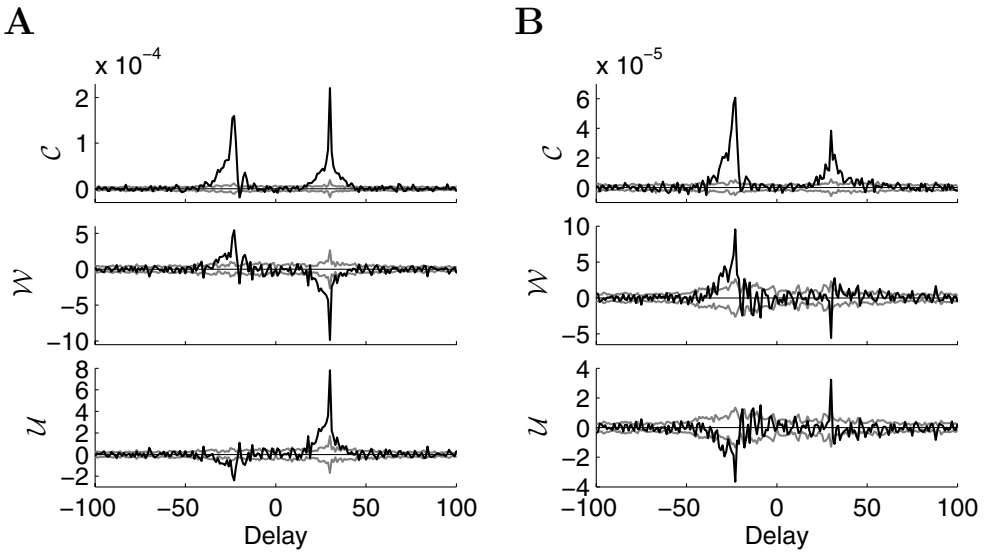


FIG. 6. *Demonstration of the results applied to networks of integrate-and fire neurons. Panels as in the bottom panels of Figure* 3. (A) *Results from a simulation of three neurons with network architecture identical to that pictured at the top of Figure* 5(B). *(The network contained both a direct connection at negative delay and common input at positive delay.) The direct connection measure* $\mathcal{W}$ *correctly identifies the direct connection from neuron one onto neuron two (appearing with negative delay). The common input measure* $\mathcal{U}$ *correctly identifies the common input at positive delay. Parameters:* $W_{12} = 0.1$, $W_{31} = W_{32} = 0.15$ *(all other* $W_{pq} = 0$*),* $d_{12} = 20$ *ms,* $d_{31} = 40$ *ms,* $d_{32} = 10$ *ms,* $\alpha_1 = \alpha_2 = 0.25$ ms^{-1}, $\alpha_3 = 0.3$ ms^{-1}, $\tau_m = 5$ *ms,* $\mathcal{E}_s = 6.5$, $\tau_s = 2$ *ms,* $\tau_{ref} = 2$ *ms. Parameters for* $\bar{\mathbf{h}}$ *are the same as those in Figure* 3(B) *except that* $\tau_h = 20$ *ms.* (B) *Results from a simulation of a random network of twenty neurons. The measure* $\mathcal{W}$ *correctly identified the direct connection from neuron one onto neuron two at negative delay (established by setting* $W_{12} = 0.12$, $d_{12} = 20$ *ms). The measure* $\mathcal{U}$ *correctly identified the common input at positive delay. Four neurons with index* $p > 2$ *were randomly selected to give this common input. For these p, the connection strength was randomly selected from* $W_{p1} \in (0.05, 0.15)$*, and* $W_{p2} = W_{p1}$*. For these four neurons, the delays were coordinated so that the delay to neuron one was* 30 *ms longer:* $d_{p2} = 2$ *ms,* $d_{p1} = 32$ *ms. The remaining connections were randomly generated as follows. For any* $p > 0$ *and* $q > 2$, $W_{pq} = 0$ *with* 90% *probability; otherwise the parameters* W_{pq} *and* d_{pq} *were randomly generated with* $W_{pq} \in (0.05, 0.15)$ *and* $d_{pq} \in (1, 40)$ *ms. Parameters for* $\bar{\mathbf{h}}_1$ *and* $\bar{\mathbf{h}}_2$ *are as in Figure* 3(A)*, except that* $\tau_h = 20$ *ms and* $b_2 = 2$ *ms. The remaining kernels with* $p > 2$ *were randomly generated with* $\psi_p \in (0, 2\pi)$, $\phi_p \in (0, 2\pi)$*, and* $f_p \in (0.2, 1.0)$*. We set* $\alpha_1 = \alpha_2 = 0.25$ ms^{-1} *and, for* $p > 2$*, randomly generated* $\alpha_p \in (0.15, 0.3)$ ms^{-1}*. We set* $\tau_m = 5$ *ms,* $\mathcal{E}_s = 6.5$, $\tau_s = 2$ *ms, and* $\tau_{ref} = 2$ *ms.*

common input are correctly identified by the measures $\mathcal{W}$ and $\mathcal{U}$, respectively. We did not constrain the unmeasured neurons to be from different subpopulations than the measured neurons. For the four common input neurons p, the maximal correlation coefficient $cc_{p2}^{\max}$ (see (2.8)) between neuron p and neuron two ranged from 0.0 to 0.7. Since the common input correlations mimicked a connection from neuron two to neuron one, these $cc_{p2}^{\max}$ were the critical measures for determining whether the common input would be identified as a direct connection. The simulation indicates that the common input neurons (at least on average) were considered to be from subpopulations different from that of neuron two.

5. Discussion. The results demonstrate that we can correctly identify subpopulation connectivity when neural response can be captured by the linear-nonlinear model (2.1), the coupling is not too strong, and we have a lot of data. Before we focus

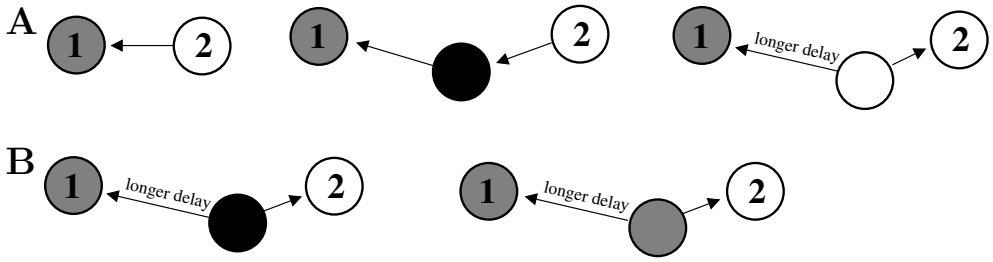


FIG. 7. *Schematic summary of the determination of subpopulation connectivity that we are able to achieve. Each pictured network configuration leads to a correlation between the spikes of neuron one and a delayed version of the spikes of neuron two. If one analyzed the joint statistics of the spikes of neurons one and two (e.g., with a covariogram), each example would appear to involve a direct connection from neuron two onto neuron one. (The unlabeled neuron is not measured.) Our result is that we can distinguish* (A) *network configurations in the top row from* (B) *network configurations in the bottom row using our analysis of the joint statistics of the stimulus and the spikes of neurons one and two. The subpopulation of each neuron is indicated by the shading (white, gray, or black).* (A) *We consider the network configurations in the top row to belong to the "direct connection class," as each configuration will be identified as a direct connection by our analysis. We cannot distinguish among the configurations in the direct connection class. Nonetheless, since each configuration contains a causal connection from neuron two's subpopulation (white) onto neuron one's subpopulation (gray), our analysis can still accurately determine connectivity at the level of subpopulations. From left to right, the configurations are a direct connection from neuron two onto neuron one, an indirect connection through an unmeasured neuron, and common input from neuron two's subpopulation.* (B) *We consider network configurations in the bottom row to belong to the "common input class," as each configuration will be identified as common input by our analysis. Since these configurations have no causal connection from neuron two's subpopulation (white) onto neuron one's subpopulation (gray), it is important that our analysis can distinguish them from the direct connection configurations* (A)*. From left to right, the configurations are common input from a different subpopulation (black) and common input from neuron one's subpopulation (gray).*

on the limitations caused by these conditions, we discuss the significance of subpopulation connectivity and the relationship between this approach and other works.

5.1. Identification of subpopulation connectivity. Recall that two neurons are in the same subpopulations if their effective kernels $\mathbf{h}_q$ (defined by fitting the uncoupled model (2.2)) are similar. In some contexts, neuroscientists would refer to these kernels as the neurons' *receptive fields*; in this case, a subpopulation would be a group of neurons with similar receptive fields.

We have shown that common input originating from within one neuron's subpopulation could appear like a direct connection from that neuron onto the other measured neuron. Hence, when we identify a direct connection between neurons, we can only conclude that there is a connection between those neurons' subpopulations.

We summarize our conclusions in Figure 7. The direct connection measure $\mathcal{W}$ and the common input measure $\mathcal{U}$ effectively divide network configurations into two classes; we will call these a direct connection class and a common input class. The top row (A) shows network configurations that would be classified as having a direct connection from neuron two onto neuron one. Besides the actual direct connection, this direct connection class contains an indirect connection through an unmeasured neuron and common input from neuron two's subpopulation. All three network configurations contain a causal connection from neuron two's subpopulation (white) onto neuron one's subpopulation (gray).

The bottom row (B) shows network configurations that would be classified as having common input. When the delay onto neuron one is longer (so that the correlations

mimic a direct connection from neuron two onto neuron one), the common input class contains networks with common input from different subpopulations and networks with common input from neuron one's subpopulation. In neither of these cases is there a connection from neuron two's subpopulation onto neuron one. Consequently, in order to accurately identify subpopulation connectivity, these configurations must be distinguished from the direct connection class of the first row. We have shown that, subject to the limitations mentioned above, we can make this distinction.

We argue that, in some experimental contexts, determining subpopulation connectivity is as informative as determining the actual connectivity between two measured neurons. In many experiments, electrodes are "blindly" inserted into the brain, and the precise identity of measured neurons remains unknown. In this situation, neurons are simply characterized by their response properties (e.g., their receptive fields), such as those captured by the effective kernels $\mathbf{h}_q^i$.

Since the precise identify of measured neurons is unknown, the best conclusion one can make about connectivity is that a neuron with response properties "A" is connected to a neuron with response properties "B." In other words, the best one can say is that a neuron from the subpopulation characterized by response properties "A" is connected to a neuron from the subpopulation characterized by response properties "B." This is the best possible conclusion even if we didn't have to worry about ambiguity introduced by connections from unmeasured neurons. Our central result is that we have developed an approach to achieve this best possible conclusion even in the presence of common input from unmeasured neurons.

5.2. Precise identity of subpopulations. The above discussion assumes the presence of discrete subpopulations. If this were the case, the statement that two neurons are from the same subpopulation would be unambiguous. Of course, in general, this is not the case. The response properties of neurons across a large population may be better modeled as a continuum, where the correlation coefficients $cc_{pq}^{\max}$ of (2.8) could be any value between 0 and 1. (Since cc_{pq}^{k} tends to zero for large $|k|$, the maximum is always nonnegative.) In order to make our subpopulation definition precise, we would like to have some cutoff value of $cc_{pq}^{\max}$, above which we could say that neurons p and q are from the same population and below which we could say they are from different subpopulations.

To explore this issue, we simulated the common input network of Figures 3(B) and 4 and the integrate-and-fire network of Figure 6(A), varying the model parameters to change[14] $cc_{32}^{\max}$. Although there was no clean cutoff, the cutoff value was around $cc_{32}^{\max} = 0.6$. If $cc_{32}^{\max}$ was near 0.6, the results were mixed, and the subpopulation of neuron three seemed to depend on model parameters. But for larger $cc_{32}^{\max}$, neuron three acted like a member of neuron two's subpopulation because the common input appeared as a direct connection in measure $\mathcal{W}$ (as in Figure 4(A)). Similarly, for $cc_{32}^{\max}$ much smaller than 0.6, neuron three acted like a member of a different subpopulation because the common input was correctly identified as common input (as in Figures 3(B) and 4(B)). Hence, at least for these coupling strengths and roughly equivalent firing rates, neurons p and q were effectively in the same subpopulation when $cc_{pq}^{\max}$ was well above 0.6. (See section 5.5 for examples of how strong coupling and disparity in firing rates can further complicate the picture.)

[14]To keep the discussion as simple as possible, we ensured that the maximum of cc_{32}^{k} occurred at the delay $k = 2$, since $\bar{W}_{32}^{j}$ was maximal at $j = 2$. Section 3.4 shows why cc_{32}^{2} is the important correlation coefficient for this case.

5.3. Heuristic explanation for results. To provide some intuition into how our approach successfully determines subpopulation connectivity, we give a heuristic explanation of why one should be able to distinguish subpopulation connectivity by analyzing the joint statistics of the measured spikes and the stimulus. We claim that one should expect that the relationship between the measured neurons' spikes and the stimulus will differ between the direct connection class of Figure 7(A) and the common input class of Figure 7(B).

For example, when the stimulus sequence happens to be optimal for neuron two and subsequently optimal for neuron one, the effectiveness of a connection from neuron two's subpopulation onto neuron one's subpopulation will be enhanced. (In this case, a spike from neuron two's subpopulation is likely to reach the neuron from neuron one's subpopulation when it is ready to fire.) Since in each network configuration in the direct connection class (Figure 7(A)) the correlation between neuron one and neuron two depends on a connection from neuron two's subpopulation onto neuron one, we expect the correlation to be especially strong for this particular stimulus sequence.

On the other hand, we would not expect the correlations in the common input class (Figure 7(B)) to be especially strong when the stimulus happens to be optimal for neuron two and subsequently optimal for neuron one. None of the connections leading to the correlation will be enhanced for this stimulus sequence, since no connection exists from neuron two's subpopulation onto neuron one's subpopulation. This example of an extreme stimulus sequence illustrates one case where the joint stimulus spike statistics will differ depending on subpopulation connectively. One might expect the differences to be evident even with other stimuli. Our results show that, at least for a simple model, one can exploit this difference to determine subpopulation connectivity.

5.4. Comparison to other approaches. Our approach succeeds in reconstructing subpopulation connectivity by combining spike correlation analysis [14, 1, 13] with white noise analysis [7, 5, 4]. It builds on previous work [10, 11] that did not address the presence of unmeasured neurons. We have previously reported [9] on our early attempts to address the unmeasured neurons where, since we did not require stimulus repeats, we had to assume that all unmeasured neurons had dissimilar kernels (effectively, that every unmeasured neuron was in its own subpopulation).

Our approach differs from the partial coherence of Rosenberg et al. [15] because it does not require measurement of the neuron producing the common input. In cases where one monitors multiple neurons simultaneously, partial coherence can rule out common input from the other measured neurons without appealing to the model assumptions underlying our analysis. Although there is a large literature in which researchers have developed methods to reconstruct the connectivity among measured neurons, we are unaware of others that explicitly account for unmeasured neurons. Without accounting for unmeasured neurons, common input from unmeasured neurons would be erroneously identified as a direct connection.

5.5. The weak coupling assumption. The analysis underlying the measures $\mathcal{W}$ and $\mathcal{U}$ relied on the assumption that the coupling $\bar{W}$ was small. The simulations demonstrate that one can obtain correct results even when the coupling is not weak. We used values of $\bar{W}^j_{pq}$ as large as 1.8 and values of $\sum_j \bar{W}^j_{pq}$ as large as 3.4. For this parameter range, the weak coupling assumption is not justified, yet the results successfully determined subpopulation connectivity.

At this point, we lack an analysis of the effects of strong coupling. We have discovered through simulations that violations of the weak coupling assumption can cause invalid results when the firing rates of the two measured neurons are greatly different. For example, if $E\{R_1^i\} \gg E\{R_2^i\}$, strong coupling can cause a direct connection from neuron one to neuron two to appear as common input. The same situation can also cause common input to appear as a direct connection from neuron two to neuron one. In other words, there is a bias for a faster neuron appearing to have a connection from a slower neuron and a bias against a slower neuron appearing to have a connection from a faster neuron. The strength of coupling at which the misidentification occurs depends on the degree of inequality between the firing rates.

For example, we analyzed a sequence of simulations of the direct connection of Figure 3(A) where we increased the disparity between the firing rate of neurons one and two. By the time neuron two fired ten times faster than neuron one, the direct connection was misidentified as common input, and we failed to reconstruct the subpopulation connectivity. On the other hand, when we halved the strength of the direct connection (and ran very long simulations), a direct connection was still accurately identified even when neuron two fired more than fifty times faster than neuron one.

We also analyzed a similar sequence of simulations of the common input configuration of Figure 3(B). The common input appeared as a direct connection from neuron two onto neuron one when neuron one fired over 20 times faster than neuron two. Because neuron three was not in neuron two's subpopulation ($c_{32}^{\max} < 0.2$), this misidentification is a failure in reconstructing subpopulation connectivity. When we increased the connection strengths by 50% (adjusting kernel parameters to keep $c_{32}^{\max} < 0.2$), the misidentification began when neuron one fired only eight times faster than neuron two. (As one might infer from the observations of section 5.2, when we changed the kernels to increase $cc_{32}^{\max}$, the common input was identified as a direct connection with lower disparities in firing rate.)

5.6. Improving statistical efficiency. Our reconstruction is based on an analysis of just a few stimulus-spike moments. We employed this moment-based approach because our intuition on such moments' behavior could guide development of this initial implementation of our subpopulation connectivity approach. One important demerit of this choice was made clear in our simulations, where we needed long simulations to obtain good results. To apply this approach to realistic neuroscience data, we will presumably need more statistically efficient techniques, such as maximum likelihood estimators, which will yield reliable estimates of subpopulation connectivity with less data.

5.7. Validations. Clearly, the assumptions of the analysis are idealizations that will never be satisfied by biological neuronal networks. The approach is viable only because accurate results can be obtained outside the strict assumptions (as demonstrated throughout section 4). However, section 5.5 demonstrated some violations of the assumptions that do lead to inaccurate results. Another possible source of inaccuracies is covariation in latency or excitability, as discussed by Brody [2]. Since such covariation is not addressed by the network model (2.1), this covariation could invalidate our results. To address possible sources of error, we must develop validation methods that can identify critical violations of assumptions that may skew the results. Such validations will allow one to trust that the results are accurate.

Ideally, one would like to test the accuracy of these results with *in vitro* experiments, where the actual connectivity can be determined by other means (such as with electrodes that enter neurons). Unfortunately, our approach depends on having

an experimentally controlled stimulus $\mathbf{X}$, where the relationship between the firing probabilities is given by (2.1). Given that *in vitro* preparations are typically severed from sensory receptors, this requirement may be difficult to achieve. A more promising testbed may be a lower organism, where the connectivity is known and neurons can be driven by a stimulus.

5.8. Extensions to other models. The model (2.1) was made as simple as possible to facilitate the analysis. It assumes, for example, that the response is an approximately linear function of the stimulus, that the network is in an asynchronous state, and that the internal dynamics of the neuron can be neglected. The results are valid only when the network is stimulated by white noise. An extension to more general elliptically symmetric stimuli should be possible. Since in this case, a linear-nonlinear model can be reconstructed (see, for example, [3]), the results should be attainable if one replaces the integration-by-parts formula (A.3) with a more general version.

We view the implementation presented in this paper simply as an example of a new framework of network analysis. The principle of analyzing joint input-output statistics may be generalized to reveal pairwise coupling in other network models. The current version should have only limited applicability to neuroscience experiments because the relationship of neural response to a stimulus will in most cases be more fundamentally nonlinear than the linear-nonlinear model (2.1). Extension of the results to more complicated models and stimuli will increase the range of applicability, allowing the approach to evolve into a useful tool for analyzing neuronal networks and other stimulus-driven networks.

Appendix A. Integration-by-parts formula. In our notation, we do not explicitly distinguish spatial versus temporal components of the stimulus, but rather let time be represented only by the temporal index of the kernels $\bar{\mathbf{h}}_q^i$ and the spikes R_q^i. We let each of the m components of $\mathbf{X}$ be independent standard normal variables, so that the probability density function of $\mathbf{X}$ is

$$\rho_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{m/2}} e^{-\frac{\|\mathbf{x}\|^2}{2}}. \tag{A.1}$$

To assist the reader, we derive an integration-by-parts formula (A.3), although such a formula is not new. Let $\mathbf{h}_k$ for $k = 1, 2, \ldots, K$ be linearly independent unit vectors corresponding to K kernels. We wish to compute

$$E\{\mathbf{X}F(\mathbf{h}_1 \cdot \mathbf{X}, \ldots, \mathbf{h}_K \cdot \mathbf{X})\},$$

where F is some smooth function with K arguments. Given the probability density function (A.1) for $\mathbf{X}$, this expected value is

$$\frac{1}{(2\pi)^{m/2}} \int \mathbf{x} F(\mathbf{h}_1 \cdot \mathbf{x}, \ldots, \mathbf{h}_K \cdot \mathbf{x}) e^{-\frac{\|\mathbf{x}\|^2}{2}} d\mathbf{x}.$$

Denote the standard unit vectors by $\mathbf{e}_j$, so that we can write the kernel $\mathbf{h}_k$ and the dot product $\mathbf{h}_k \cdot \mathbf{x}$ in component form as

$$\mathbf{h}_k = \sum_j h_{kj} \mathbf{e}_j \quad \text{and} \quad \mathbf{h}_k \cdot \mathbf{x} = \sum_j h_{kj} x_j,$$

where h_{kj} is the jth component of $\mathbf{h}_k$.

We calculate the components of $E\{\mathbf{X}F(\mathbf{h}_1\cdot\mathbf{X},\dots,\mathbf{h}_K\cdot\mathbf{X})\}$. Through integration by parts with respect to x_j, the jth component is

$$\begin{aligned}
&E\{X_jF(\mathbf{h}_1\cdot\mathbf{X},\dots,\mathbf{h}_K\cdot\mathbf{X})\}\\
&\quad=\frac{1}{(2\pi)^{m/2}}\int x_jF\left(\sum_k h_{1k}x_k,\dots,\sum_k h_{Kk}x_k\right)e^{-\frac{\|\mathbf{x}\|^2}{2}}d\mathbf{x}\\
&\quad=\frac{1}{(2\pi)^{m/2}}\sum_i h_{ij}\int F_i\left(\sum_k h_{1k}x_k,\dots,\sum_k h_{Kk}x_k\right)e^{-\frac{\|\mathbf{x}\|^2}{2}}d\mathbf{x}\\
&\quad=\sum_i h_{ij}E\{F_i(\mathbf{h}_1\cdot\mathbf{X},\dots,\mathbf{h}_K\cdot\mathbf{X})\},
\end{aligned}$$

where F_i indicates the partial derivative of F with respect to the ith variable.

Putting the components together, we conclude that

$$\begin{aligned}
E\{\mathbf{X}F(\mathbf{h}_1\cdot\mathbf{X},\dots,\mathbf{h}_K\cdot\mathbf{X})\}&=\sum_j E\{x_jF(\mathbf{h}_1\cdot\mathbf{X},\dots,\mathbf{h}_K\cdot\mathbf{X})\}\mathbf{e}_j\\
&=\sum_i E\{F_i(\mathbf{h}_1\cdot\mathbf{X},\dots,\mathbf{h}_K\cdot\mathbf{X})\}\left(\sum_j h_{ij}\mathbf{e}_j\right)\\
&=\sum_i E\{F_i(\mathbf{h}_1\cdot\mathbf{X},\dots,\mathbf{h}_K\cdot\mathbf{X})\}\mathbf{h}_i. \qquad \text{(A.2)}
\end{aligned}$$

The special case we need for our derivation is

$$\begin{aligned}
&E\{\mathbf{X}g_p(\mathbf{h}_p^i\cdot\mathbf{X})g_q(\mathbf{h}_q^{i-j}\cdot\mathbf{X})g_r(\mathbf{h}_r^{i-k}\cdot\mathbf{X})\}\\
&\quad=E\{g_p'(\mathbf{h}_p^i\cdot\mathbf{X})g_q(\mathbf{h}_q^{i-j}\cdot\mathbf{X})g_r(\mathbf{h}_r^{i-k}\cdot\mathbf{X})\}\mathbf{h}_p^i\\
&\qquad+E\{g_p(\mathbf{h}_p^i\cdot\mathbf{X})g_q'(\mathbf{h}_q^{i-j}\cdot\mathbf{X})g_r(\mathbf{h}_r^{i-k}\cdot\mathbf{X})\}\mathbf{h}_q^{i-j}\\
&\qquad+E\{g_p(\mathbf{h}_p^i\cdot\mathbf{X})g_q(\mathbf{h}_q^{i-j}\cdot\mathbf{X})g_r'(\mathbf{h}_r^{i-k}\cdot\mathbf{X})\}\mathbf{h}_r^{i-k} \qquad \text{(A.3)}
\end{aligned}$$

and the equivalent for fewer factors.

Appendix B. Equations for error function nonlinearity. The analysis for error function $\bar{g}_q(\cdot)$ (see (4.2)) mirrors the derivations outlined in [11]. In this appendix, we summarize the intermediate steps and then give the error function result for (3.8) and (3.11).

We define the effective error function parameters (ϵ_q, y_q) from the spikes of each neuron q by fitting to these spikes the uncoupled model (2.2) with nonlinearity,

$$g_q(y)=\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{y-y_q}{\epsilon_q\sqrt{2}}\right)\right].$$

Denote the inner product between kernels by $\cos\theta_{pq}^j=\mathbf{h}_p^i\cdot\mathbf{h}_q^{i-j}$. (Note that for Gaussian white noise $\cos\theta_{pq}^j=cc_{pq}^j$; see (2.7).) Define the following expressions as

functions of the parameters ϵ_q, y_q, and $\cos\theta_{pq}^j$:

$$\delta_q = \frac{1}{\sqrt{1+\epsilon_q^2}},$$

$$\lambda_{qp}^j = \frac{\delta_q y_q - \delta_p^2 \delta_q y_p \cos\theta_{qp}^j}{\sqrt{2(1-\delta_p^2\delta_q^2\cos^2\theta_{qp}^j)}},$$

$$\mu_{qp}^j = \frac{\delta_p\delta_q}{2\pi\sqrt{1-\delta_p^2\delta_q^2\cos^2\theta_{qp}^j}} \exp\left(-\frac{\delta_p^2 y_p^2 - 2\delta_p^2\delta_q^2 y_p y_q \cos\theta_{qp}^j + \delta_q^2 y_q^2}{2(1-\delta_p^2\delta_q^2\cos^2\theta_{qp}^j)}\right),$$

$$\xi_{qp}^j = \frac{\delta_q^2(1-\delta_p^2\cos^2\theta_{qp}^j)}{1-\delta_p^2\delta_q^2\cos^2\theta_{qp}^j}.$$

Define a double complementary error function

$$\operatorname{derfc}(a,b,c) = \frac{4}{\pi}\int_a^\infty dy\, e^{-y^2} \int_{\frac{b-cy}{\sqrt{1-c^2}}}^\infty dz\, e^{-z^2}. \tag{B.1}$$

The function derfc is a two-dimensional analogue of the complementary error function (see [10]).

Using the fact that $\mathbf{h}_p^i \cdot \mathbf{X}$ and $\mathbf{h}_q^{i-j} \cdot \mathbf{X}$ are joint unit normal random variables with correlation $\cos\theta_{qp}^j$, we compute the following expected values:

$$E\{g_p'(\mathbf{h}_p^i\cdot\mathbf{X})g_q(\mathbf{h}_q^{i-j}\cdot\mathbf{X})\} = \frac{\delta_p}{2\sqrt{2\pi}}\exp\left(-\frac{\delta_p^2y_p^2}{2}\right)\operatorname{erfc}(\lambda_{qp}^j),$$

$$E\{g_p'(\mathbf{h}_p^i\cdot\mathbf{X})(g_q(\mathbf{h}_q^{i-j}\cdot\mathbf{X}))^2\} = \frac{\delta_p}{4\sqrt{2\pi}}\exp\left(-\frac{\delta_p^2y_p^2}{2}\right)\operatorname{derfc}(\lambda_{qp}^j,\lambda_{qp}^j,\xi_{qp}^j),$$

$$E\{g_p'(\mathbf{h}_p^i\cdot\mathbf{X})g_q'(\mathbf{h}_q^{i-j}\cdot\mathbf{X})\} = \mu_{qp}^j,$$

$$E\{g_p'(\mathbf{h}_p^i\cdot\mathbf{X})g_q'(\mathbf{h}_q^{i-j}\cdot\mathbf{X})g_q(\mathbf{h}_q^{i-j}\cdot\mathbf{X})\} = \frac{\mu_{qp}^j}{2}\operatorname{erfc}\left(\frac{\lambda_{qp}^j(1-\xi_{qp}^j)}{\sqrt{1-(\xi_{qp}^j)^2}}\right),$$

$$E\{g_p''(\mathbf{h}_p^i\cdot\mathbf{X})g_q(\mathbf{h}_q^{i-j}\cdot\mathbf{X})\} = \frac{\delta_p^3y_p}{2\sqrt{2\pi}}\exp\left(-\frac{\delta_p^2y_p^2}{2}\right)\operatorname{erfc}(\lambda_{qp}^j) - \delta_p^2\cos\theta_{qp}^j\mu_{qp}^j,$$

$$E\{g_p''(\mathbf{h}_p^i\cdot\mathbf{X})(g_q(\mathbf{h}_q^{i-j}\cdot\mathbf{X}))^2\} = \frac{\delta_p^3y_p}{4\sqrt{2\pi}}\exp\left(-\frac{\delta_p^2y_p^2}{2}\right)\operatorname{derfc}(\lambda_{qp}^j,\lambda_{qp}^j,\xi_{qp}^j) - \delta_p^2\cos\theta_{qp}^j\mu_{qp}^j\operatorname{erfc}\left(\frac{\lambda_{qp}^j(1-\xi_{qp}^j)}{\sqrt{1-(\xi_{qp}^j)^2}}\right),$$

$$E\{g_p''(\mathbf{h}_p^i\cdot\mathbf{X})g_q'(\mathbf{h}_q^{i-j}\cdot\mathbf{X})\} = \frac{\delta_p^2[y_p-\delta_q^2y_q\cos\theta_{qp}^j]\mu_{qp}^j}{(1-\delta_p^2\delta_q^2\cos^2\theta_{qp}^j)}.$$

We rewrite (3.8) and (3.11) in terms of the above quantities:

$$\begin{aligned} C_{21}^k = {} & \hat{W}_{21}^k\frac{\delta_1}{2\sqrt{2\pi}}\exp\left(-\frac{\delta_1^2y_1^2}{2}\right)\left[\operatorname{erfc}(\lambda_{21}^k) - \frac{1}{2}\operatorname{derfc}(\lambda_{21}^k,\lambda_{21}^k,\xi_{21}^k)\right] \\ & + \hat{W}_{12}^{-k}\frac{\delta_2}{2\sqrt{2\pi}}\exp\left(-\frac{\delta_2^2y_2^2}{2}\right)\left[\operatorname{erfc}(\lambda_{12}^{-k}) - \frac{1}{2}\operatorname{derfc}(\lambda_{12}^{-k},\lambda_{12}^{-k},\xi_{12}^{-k})\right] \\ & + \hat{U}_{21}^k\mu_{21}^k, \end{aligned} \tag{B.2}$$

$$
\begin{aligned}
A_1^k = \hat{W}_{21}^k \Bigg\{ & \frac{\delta_1^3 y_1}{2\sqrt{2\pi}} \exp\left(-\frac{\delta_1^2 y_1^2}{2}\right) \left[\operatorname{erfc}(\lambda_{21}^k) - \frac{1}{2}\operatorname{derfc}(\lambda_{21}^k, \lambda_{21}^k, \xi_{21}^k)\right] \\
& - \delta_1^2 \cos\theta_{21}^k \mu_{21}^k \left[1 - \operatorname{erfc}\left(\frac{\lambda_{21}^k(1-\xi_{21}^k)}{\sqrt{1-(\xi_{21}^k)^2}}\right)\right] \Bigg\} \\
& + \hat{W}_{12}^{-k} \mu_{12}^{-k} \left[1 - \operatorname{erfc}\left(\frac{\lambda_{12}^{-k}(1-\xi_{12}^{-k})}{\sqrt{1-(\xi_{12}^{-k})^2}}\right)\right] \\
& + \hat{U}_{21}^k \frac{\delta_1^2[y_1 - \delta_2^2 y_2 \cos\theta_{21}^k]\mu_{21}^k}{(1-\delta_1^2\delta_2^2\cos^2\theta_{21}^k)},
\end{aligned} \tag{B.3}
$$

$$
\begin{aligned}
A_2^k = & \hat{W}_{21}^k \mu_{21}^k \left[1 - \operatorname{erfc}\left(\frac{\lambda_{21}^k(1-\xi_{21}^k)}{\sqrt{1-(\xi_{21}^k)^2}}\right)\right] \\
& + \hat{W}_{12}^{-k} \Bigg\{ \frac{\delta_2^3 y_2}{2\sqrt{2\pi}} \exp\left(-\frac{\delta_2^2 y_2^2}{2}\right) \left[\operatorname{erfc}(\lambda_{12}^{-k}) - \frac{1}{2}\operatorname{derfc}(\lambda_{12}^{-k}, \lambda_{12}^{-k}, \xi_{12}^{-k})\right] \\
& - \delta_2^2 \cos\theta_{12}^{-k} \mu_{12}^{-k} \left[1 - \operatorname{erfc}\left(\frac{\lambda_{12}^{-k}(1-\xi_{12}^{-k})}{\sqrt{1-(\xi_{12}^{-k})^2}}\right)\right] \Bigg\} \\
& + \hat{U}_{21}^k \frac{\delta_2^2[y_2 - \delta_1^2 y_1 \cos\theta_{12}^{-k}]\mu_{12}^{-k}}{(1-\delta_2^2\delta_1^2\cos^2\theta_{12}^{-k})}.
\end{aligned} \tag{B.4}
$$

The key point of these long formulas is that, with the exception of $\hat{W}_{21}^k$, $\hat{W}_{12}^{-k}$, and $\hat{U}_{21}^k$, all expressions are functions of the error function parameters of the measured neurons (ϵ_1, ϵ_2, y_1, and y_2) and $\cos\theta_{21}^k = \cos\theta_{12}^{-k}$. The kernels (and hence $\cos\theta_{21}^k$) are computed from (3.6). The parameters ϵ_q and y_q, for $q = 1, 2$, can be calculated from (3.4) and (3.5) with the use of the formulas

$$
\begin{aligned}
E\{g_q(\mathbf{h}_q^i \cdot \mathbf{X})\} &= \frac{1}{2}\operatorname{erfc}\left(\frac{\delta_q y_q}{\sqrt{2}}\right), \\
E\{g_q'(\mathbf{h}_q^i \cdot \mathbf{X})\} &= \frac{\delta_q}{\sqrt{2\pi}} \exp\left(-\frac{\delta_q^2 y_q^2}{2}\right).
\end{aligned}
$$

Appendix C. Estimating confidence intervals. We estimate the confidence interval of our measures using essentially the procedure outlined in Appendix B of [10]. Besides changing the base variables to those needed for the current analysis, we make the following two minor changes. First, we calculate the covariances of inner products accurately using the covariances among all the factors in the product. Second, since the statistics from different delays are uncoupled in our equations, we ignore covariances among statistics from different delays.

Acknowledgment. The author thanks Dario Ringach for numerous helpful discussions throughout the development of this research.

REFERENCES

[1] A. M. H. J. Aertsen, G. L. Gerstein, M. K. Habib, and G. Palm, *Dynamics of neuronal firing correlation: Modulation of "effective connectivity,"* J. Neurophysiol., 61 (1989), pp. 900–917.

[2] C. D. Brody, *Correlations without synchrony*, Neural Comput., 11 (1999), pp. 1537–1551.
[3] E. J. Chichilnisky, *A simple white noise analysis of neural light responses*, Network: Comput. Neural Syst., 12 (2001), pp. 199–213.
[4] Y. Dan, J.-M. Alonso, W. M. Usrey, and R. C. Reid, *Coding of visual information by precisely correlated spikes in the lateral geniculate nucleus*, Nature Neurosci., 1 (1998), pp. 501–507.
[5] E. DeBoer and P. Kuyper, *Triggered correlation*, IEEE Trans. Biomed. Eng., 15 (1968), pp. 169–179.
[6] S. Marcelja, *Mathematical description of the responses of simple cortical cells*, J. Opt. Soc. Am., 70 (1980), pp. 1297–1300.
[7] P. N. Marmarelis and V. Z. Marmarelis, *Analysis of Physiological Systems: The White Noise Approach*, Plenum Press, New York, 1978.
[8] D. Q. Nykamp, *Measuring linear and quadratic contributions to neuronal response*, Network: Comput. Neural Syst., 14 (2003), pp. 673–702.
[9] D. Q. Nykamp, *Reconstructing stimulus-driven neural networks from spike times*, in Advances in Neural Information Processing Systems 15, S. Becker, S. Thrun, and K. Obermayer, eds., MIT Press, Cambridge, MA, 2003, pp. 309–316.
[10] D. Q. Nykamp, *Spike correlation measures that eliminate stimulus effects in response to white noise*, J. Comp. Neurosci., 14 (2003), pp. 193–209.
[11] D. Q. Nykamp, *White noise analysis of coupled linear-nonlinear systems*, SIAM J. Appl. Math., 63 (2003), pp. 1208–1230.
[12] D. Q. Nykamp and D. L. Ringach, *Full identification of a linear-nonlinear system via cross-correlation analysis*, J. Vision, 2 (2002), pp. 1–11.
[13] G. Palm, A. M. H. J. Aertsen, and G. L. Gerstein, *On the significance of correlations among neuronal spike trains*, Biol. Cybern., 59 (1988), pp. 1–11.
[14] D. H. Perkel, G. L. Gerstein, and G. P. Moore, *Neuronal spike trains and stochastic point processes.* II. *Simultaneous spike trains*, Biophys. J., 7 (1967), pp. 419–440.
[15] J. R. Rosenberg, A. M. Amjad, P. Breeze, D. R. Brillinger, and D. M. Halliday, *The Fourier approach to the identification of functional coupling between neuronal spike trains*, Prog. Biophys. Mol. Biol., 53 (1989), pp. 1–31.

SIAM J. APPL. MATH.
Vol. 65, No. 6, pp. 2033–2048

MATHEMATICAL ANALYSIS OF THE GENERALIZED NATURAL MODES OF AN INHOMOGENEOUS OPTICAL FIBER*

E. M. KARTCHEVSKI†, A. I. NOSICH‡, AND G. W. HANSON§

Abstract. The eigenvalue problem for generalized natural modes of an inhomogeneous optical fiber without a sharp boundary is formulated as a problem for the set of time-harmonic Maxwell equations with the Reichardt condition at infinity in the cross-sectional plane. The generalized eigenvalues (including, as subsets, the well-known guided and leaky modes) of this problem are the complex propagation constants on a logarithmic Riemann surface. A theorem on spectrum localization is proved, and then the original problem is reduced to a nonlinear spectral problem with a compact integral operator. It is proved that the set of all eigenvalues of the original problem can only be a set of isolated points on the Riemann surface, and it is also proved that each eigenvalue depends continuously on the frequency and refraction index and can appear and disappear only at the boundary of the Riemann surface.

Key words. electromagnetic theory, optical fiber, waveguides, eigenvalue problem, guided modes

AMS subject classifications. 35P30, 45C05, 65R20, 78A50

DOI. 10.1137/040604376

1. Introduction.

Optical fibers are dielectric waveguides (DWs), i.e., regular dielectric rods, having various cross sectional shapes, and where generally the refractive index of the dielectric may vary in the waveguide's cross section. Although existing technologies often result in a refractive index that is anisotropic, frequently it is possible to assume that the fiber is isotropic, which is the case investigated in this work. The study of the source-free electromagnetic fields, called *natural modes*, that can propagate on DWs necessitates that longitudinally the rod extend to infinity. Since often DWs are not shielded, the medium surrounding the waveguide transversely forms an unbounded domain, typically taken to be free space. This fact plays an extremely important role in the mathematical analysis of natural waveguide modes, and brings into consideration a variety of possible formulations. Each different formulation can be cast as an eigenvalue problem for the set of time-harmonic Maxwell equations, but they differ in the form of the condition imposed at infinity in the cross-sectional plane, and hence in the functional class of the natural-mode field. As we discuss below, this also restricts the localization of the eigenvalues in the complex plane of the eigenparameter.

Historically, the first DWs to be studied were step-index waveguides having circular cross section; interior to the waveguide, the refractive index was either homo-

*Received by the editors February 23, 2004; accepted for publication (in revised form) February 14, 2005; published electronically August 9, 2005. This work was partly supported by the Russian Foundation for Basic Research, grant 03-01-96184, and by the United States National Research Council COBASE Program.

http://www.siam.org/journals/siap/65-6/60437.html

†Department of Applied Mathematics, Kazan State University, 18 Kremliovskaia Street, Kazan, Russia, 420008 (evgenii.karchevskii@ksu.ru).

‡Institute of Radio Physics and Electronics, National Academy of Sciences, Kharkov, Ukraine, 61085 (alex@emt.kharkov.ua).

§Department of Electrical Engineering, University of Wisconsin-Milwaukee, 3200 North Cramer Street, Milwaukee, WI 53211 (george@uwm.edu).

geneous or coaxial-layered. In these cases, by using separation of variables, modal eigenvalue problems are easily reduced to families of transcendental dispersion equations associated with the azimuthal indices (see, e.g., [1], [2]). All questions concerning discreteness and existence of the natural-mode spectrum are settled "automatically" due to general results from the theory of complex variables and the analytic properties of cylindrical functions with integer indices and complex arguments.

For these circular cross section DWs the first class of natural modes to be studied were *purely guided modes*, which have real-valued wavenumbers. The fields of the guided modes are confined near to the waveguide, decaying exponentially transversely away from the waveguide, so that they belong to the space L_2 in the whole cross-sectional plane. Corresponding eigenvalue problems are self-adjoint. Later it was discovered that the guided modes of a circular DW can turn into (i.e., be analytically continued as) so-called *leaky-wave modes*, existing on the "improper" sheet of a square-root Riemann surface, with the wavenumbers migrating off the real axis of the "proper" sheet onto the "improper" sheet as some parameters of the structure vary [3]. It was noticed that leaky modes can be studied as solutions of a more general eigenvalue problem, without cross-sectional field confinement, due to some relaxed, although never explicitly formulated, condition at infinity.

Although leaky waves exist on an "improper" Riemann sheet, they have considerable physical importance in wave excitation and fiber discontinuity problems. In particular, it is known that the electromagnetic fields existing on a dielectric waveguide can be represented as a discrete sum of bound modes (which are the mentioned guided modes generated by the eigenvalues of the propagation constant on the real axis of the "proper" sheet) and a continuous sum (i.e., integral) of so-called *radiation modes* (whose physical sense still causes discussions) [1], [2], [4]. It has been shown that, although leaky waves are not themselves a part of a "proper" spectral field representation, in many cases the continuum of the radiation modes may be approximated by a discrete sum of leaky modes [5], representing the near field of a source-excited fiber. Often the leaky-wave sum can be reduced to a single term, providing a concise analytical form for the near-zone radiation field. Furthermore, various features in the far-field radiation pattern of a real, finite-length, source-driven fiber can be interpreted in terms of leaky-wave excitation. In addition to source-driven waveguides, leaky modes on longitudinally invariant fibers are important in the analysis of radiation and mode-conversion effects associated with waveguide discontinuities such as fiber splices [6], radiation from waveguide bends [7], and radiation from anisotropic fibers [8], [9]. Some properties of leaky modes on dielectric waveguides, and, in particular, dielectric fibers, are presented in [2], [3], [5], [6], [7], [8], [9], [10], [11], [12].

In addition to leaky modes, it was discovered that on the "proper" sheet, but off its real axis, one can also find other generalized eigenvalues (modal wavenumbers) [13] known as *complex modes*. Analogous results were obtained numerically for gradient-index DWs of arbitrary cross section [14]. These complex modes are often important in near-field fiber discontinuity problems and mode-matching analysis. It is important to note that all of these known types of natural modes can transform (be continued) one into another, following variation of some geometrical or material parameter or frequency. Due to the presence of the two-dimensional unbounded domain and the resulting Green's functions represented as Hankel functions, it is easy to see that the dispersion equations contain logarithmic as well as square-root–type branch points.

If the cross section is not circular, the study of the natural modes encounters both methodological and numerical problems. In [15] an elliptic DW was studied by using expansions in terms of Mathieu functions. However, in that work as well as in other studies of waveguides having complicated cross sections, or of multirod waveguides, the modal problems are reduced not to transcendental equations but to infinite matrix equations or integral equations (IEs). Hence, it is necessary to base the analysis on the theory of operator-functions depending on parameters. Once again, by restricting the desired field behavior in the cross-sectional plane, one arrives at different formulations of the eigenvalue problem in terms of the transverse condition at infinity; eigenvalue localization and the function class of the natural mode field are tied up with this condition.

In recent years, research on the natural modes of arbitrarily shaped DWs has been focused on the development of efficient and reliable computational methods. For instance, in [16] the eigenvalue problem for the natural modes of arbitrary DWs was studied by splitting the differential operator into self-adjoint and perturbation parts and using a discretization in terms of the eigenfunctions of the self-adjoint operator. This enabled the authors to develop a very efficient numerical technique, although its convergence was not proven.

In the papers on numerical methods for DWs, the mathematical grounding of the methods was frequently neglected; however, useful insight into the encountered difficulties and modal behavior has been discussed (e.g., see [17], [18]). The most rigorous efforts were connected with IE formulations. Within this class the domain IE method has the attractive advantage of being applicable to cross-sectionally inhomogeneous (and, in fact, anisotropic) DWs [19], [20]. A problem with domain IEs is that they are strongly singular, which previously prevented their use in a mathematical study of the spectrum of the eigenvalues, with the exception of [21] for the purely guided modes of an inhomogeneous DW. For real-valued propagation constants it was proven in [21] that the operator of the domain IE is semi-Fredholm.

A rigorous mathematical study of an arbitrary-shaped DW was performed in [22] within the guided (proper) mode formulation. This enabled the authors to make extensive use of the theory of unbounded self-adjoint operators. For example, by using the min-max principle, they proved the existence of guided modes, the number of which is finite and depends on frequency. However, generalized natural modes having complex valued propagation constants cannot be studied by this approach.

The above considerations give a new thrust to the idea of elaborating a generalized formulation of the modal eigenproblem in order to bring together all the possible natural-mode solutions. All of the known natural-mode solutions (i.e., guided modes, leaky modes, complex modes) satisfy the Reichardt condition [23] at infinity. The wavenumbers may be generally considered on the appropriate logarithmic Riemann surface. The Reichardt condition in this problem is connected with the fact that the wavenumber may be complex. For real wavenumbers on the principal ("proper") sheet of this Riemann surface, one can reduce the Reichardt condition to either the Sommerfeld radiation condition or to the condition of exponential decay. The Reichardt condition may be considered as a generalization of the Sommerfeld radiation condition and can be applied for complex wavenumbers. This condition may also be considered as the continuation of the Sommerfeld radiation condition from a part of the real axis of the complex parameter (wavenumber) to the appropriate logarithmic Riemann surface.

During recent years the Reichardt condition has been widely used for statements of various wave propagation problems [24], [25], [26]. By using the Reichardt condition, the problems on generalized modes of microstrip and slot lines on a cylindrical substrate were investigated in [27], [28], [29]. Tensor Green's functions of generic open waveguides with compact cross sections were analyzed in [30] by using Fourier transforms and IE techniques in the transform domain. It was shown that the complex-valued poles of analytic continuations of the Green's functions satisfy a certain eigenvalue problem. Their residues can be interpreted as the generalized natural modes. In this case, the eigenvalue problem should be formulated with the Reichardt condition at infinity. Reducing Maxwell's equations to an IE and converting the latter to a Fredholm second kind equation enabled the proof of some important properties of the spectrum of the generalized modes. Furthermore, in [31], [32] a similar formulation was applied to study generalized guided modes in DWs, and a numerical algorithm was developed based on a Galerkin discretization in terms of a trigonometric basis.

In this paper we extend the approach of [30], [31], [32] to the analysis of generalized natural modes of arbitrary-cross-section DWs having inhomogeneous (although continuous) refractive index. Here, we use the model of DW without a sharp boundary, as was proposed in [33]. Such an approach enables one to reduce the original problem to a nonlinear spectral problem with a compact integral operator, and was originally introduced in [34] and used in [35]. We present a unified and rigorous theory of generalized natural modes in terms of the Reichardt condition at infinity.

The rest of this paper is organized as follows. Physical assumptions, basic equations, and notation are presented in section 2. In section 3 we formulate the modal eigenvalue problem as a problem for the set of time-harmonic Maxwell equations with the Reichardt condition at infinity in the cross-sectional plane. The eigenvalues of this problem are the complex propagation constants of the natural modes, and we introduce a classification of modal eigenvalues in terms of their location on the logarithmic Riemann surface. In section 4 we prove a theorem on localization of eigenvalues, where it is established that there exists a domain free of eigenvalues on this surface. In section 5 we investigate the eigenvalues as functions of frequency and refractive index, and we reduce the original problem to a nonlinear spectral problem with a compact integral operator. Using general results from the spectral theory of operator-valued functions [36], we prove that the set of all eigenvalues of the original problem can only be a set of isolated points on the logarithmic Riemann surface, and also we prove that each eigenvalue depends continuously on frequency and refractive index, and can appear and disappear only at the boundary of the logarithmic Riemann surface.

2. Basic relations. We consider the generalized natural modes of the regular DW shown in Figure 1. Let the three-dimensional space $\{(x_1, x_2, x_3) : -\infty < x_1, x_2, x_3 < \infty\}$ be occupied by an isotropic source-free medium, and let the refractive index be prescribed as a positive real-valued function $n = n(x_1, x_2)$ independent of the longitudinal coordinate x_3 and equal to a constant n_∞ outside a cylinder. The axis of the cylinder is parallel to the x_3-axis, and its cross section is a bounded domain Ω with a Lipschitz boundary on the plane $\mathbb{R}^2 = \{(x_1, x_2) : -\infty < x_1, x_2 < \infty\}$. Denote by Ω_∞ the unbounded domain $\Omega_\infty = \mathbb{R}^2 \setminus \overline{\Omega}$, and denote by n_+ the maximum of the function n in the domain Ω, where $n_+ > n_\infty$. Let the function n belong to the space of real-valued twice continuously differentiable functions in $\mathbb{R}^2$.

The modal problem can be formulated as a vector eigenvalue problem for the set of harmonic Maxwell equations, assuming that electric and magnetic field vectors

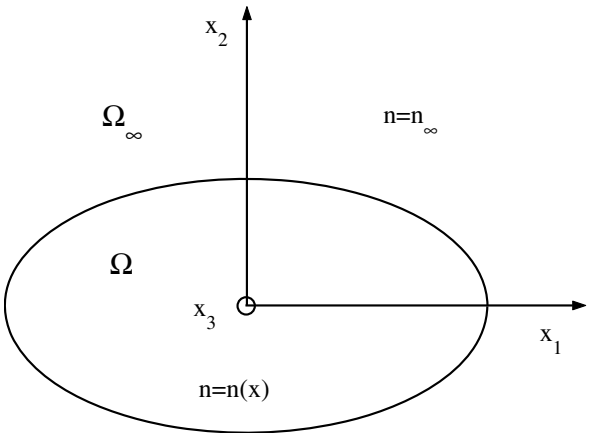


FIG. 1. *Geometry of a dielectric waveguide.*

have the form

$$\mathcal{E}(x, x_3, t) = \mathrm{Re}\left(\mathrm{E}(x)\exp\left(i\beta x_3 - i\omega t\right)\right), \tag{1}$$

$$\mathcal{H}(x, x_3, t) = \mathrm{Re}\left(\mathrm{H}(x)\exp\left(i\beta x_3 - i\omega t\right)\right). \tag{2}$$

Here $x = (x_1, x_2)$, $\omega > 0$ is the radian frequency, β is the complex-valued modal wavenumber (or propagation constant), and E and H are complex amplitudes of $\mathcal{E}$ and $\mathcal{H}$. For the sake of clarity, we note that, unlike in [22], we consider the propagation constant β as an unknown complex parameter and $\omega > 0$ as a given parameter. Such a choice seems to be commonly adopted in the fiber optics and microwave research communities due to the easy control of frequency.

For the fields of the form (1), (2), the set of Maxwell equations becomes

$$\mathrm{Rot}_\beta \mathrm{E} = i\omega\mu_0 \mathrm{H}, \quad x \in \mathbb{R}^2, \tag{3}$$

$$\mathrm{Rot}_\beta \mathrm{H} = -i\omega\varepsilon_0 n^2 \mathrm{E}, \quad x \in \mathbb{R}^2. \tag{4}$$

Here ε_0, μ_0 are the free-space dielectric and magnetic constants, respectively, and

$$\mathrm{Rot}_\beta \mathrm{E} = \begin{bmatrix} \partial E_3/\partial x_2 - i\beta E_2 \\ i\beta E_1 - \partial E_3/\partial x_1 \\ \partial E_2/\partial x_1 - \partial E_1/\partial x_2 \end{bmatrix}. \tag{5}$$

By $C^2(\mathbb{R}^2)$ denote the space of twice continuously differentiable in $\mathbb{R}^2$ complex-valued functions. We shall be seeking nonzero solutions $[\mathrm{E}, \mathrm{H}]$ of set (3), (4) in the space $(C^2(\mathbb{R}^2))^6$.

Let F be a three-dimensional vector-function,

$$\mathrm{F} = \begin{bmatrix} F_1 \\ F_2 \\ F_3 \end{bmatrix} \in \left(C^2(\mathbb{R}^2)\right)^3,$$

and let $u \in C^2(\mathbb{R}^2)$ be a scalar function. By definition, set

$$\mathrm{Div}_\beta \mathrm{F} = \frac{\partial F_1}{\partial x_1} + \frac{\partial F_2}{\partial x_2} + i\beta F_3, \tag{6}$$

$$\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}, \tag{7}$$

$$\mathrm{Grad}_\beta u = \begin{bmatrix} \partial u/\partial x_1 \\ \partial u/\partial x_2 \\ i\beta u \end{bmatrix}, \quad \mathrm{grad} u = \begin{bmatrix} \partial u/\partial x_1 \\ \partial u/\partial x_2 \\ 0 \end{bmatrix}, \tag{8}$$

$$\mathrm{grad}_2 u = \begin{bmatrix} \partial u/\partial x_1 \\ \partial u/\partial x_2 \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} F_1 \\ F_2 \end{bmatrix}. \tag{9}$$

By direct calculation it is easy to obtain the following equations:

$$\mathrm{Div}_\beta \left(\mathrm{Grad}_\beta u\right) = \Delta u - \beta^2 u, \tag{10}$$

$$\mathrm{Div}_\beta \left(\mathrm{Rot}_\beta \mathrm{F}\right) = 0, \tag{11}$$

$$\mathrm{Div}_\beta \left(u\mathrm{F}\right) = u\mathrm{Div}_\beta \mathrm{F} + \left(\mathrm{F}, \mathrm{grad} u\right), \tag{12}$$

$$\mathrm{Rot}_\beta \left(\mathrm{Grad}_\beta u\right) = 0, \tag{13}$$

$$\mathrm{Rot}_\beta \left(\mathrm{Rot}_\beta \mathrm{F}\right) = -\Delta \mathrm{F} + \beta^2 \mathrm{F} + \mathrm{Grad}_\beta \left(\mathrm{Div}_\beta \mathrm{F}\right), \tag{14}$$

where

$$(\mathrm{F}, \mathrm{L}) = \sum_{i=1}^{3} F_i L_i. \tag{15}$$

LEMMA 2.1. *If* $[\mathrm{E}, \mathrm{H}]$ *is a solution of the set* (2.3), (2.4), *then for* $x \in \mathbb{R}^2$

$$\mathrm{Rot}_\beta \left(\mathrm{Rot}_\beta \mathrm{E}\right) = k^2 n^2 \mathrm{E}, \tag{16}$$

$$\mathrm{Rot}_\beta \left(n^{-2} \mathrm{Rot}_\beta \mathrm{H}\right) = k^2 \mathrm{H}, \tag{17}$$

$$\mathrm{Div}_\beta \left(n^2 \mathrm{E}\right) = 0, \tag{18}$$

$$\mathrm{Div}_\beta \left(\mathrm{H}\right) = 0, \tag{19}$$

where $k^2 = \varepsilon_0 \mu_0 \omega^2$.

Proof. Applying the Rot_β operator to both sides of (3) and (4), we obtain (16), (17). Applying the Div_β operator to both sides of (3) and (4) and using (11), we obtain (18), (19). ∎

LEMMA 2.2. *If* $[\mathrm{E}, \mathrm{H}]$ *is a solution of the set* (2.3), (2.4), *then*

$$\mathrm{Div}_\beta \left(\left(n^2 - n_\infty^2\right) \mathrm{E}\right) = n_\infty^2 (\mathrm{E}, n^{-2} \mathrm{grad} n^2), \quad x \in \mathbb{R}^2. \tag{20}$$

Proof. Using (12) leads to

$$\mathrm{Div}_\beta \left(\left(n^2 - n_\infty^2\right) \mathrm{E}\right) = \left(n^2 - n_\infty^2\right) \mathrm{Div}_\beta \mathrm{E} + \left(\mathrm{E}, \mathrm{grad}\left(n^2 - n_\infty^2\right)\right), \quad x \in \mathbb{R}^2. \tag{21}$$

Taking into account (18) and (12), we arrive at

$$-\mathrm{Div}_\beta \mathrm{E} = (\mathrm{E}, n^{-2} \mathrm{grad} n^2), \quad x \in \mathbb{R}^2. \tag{22}$$

Combining (21) and (22), we obtain (20). ∎

LEMMA 2.3. *If* $[\mathrm{E},\mathrm{H}]$ *is a solution of the set* (2.3), (2.4), *then*

$$\left[\Delta + \left(k^2 n_\infty^2 - \beta^2\right)\right] \begin{bmatrix} \mathrm{E} \\ \mathrm{H} \end{bmatrix} = 0, \quad x \in \Omega_\infty. \tag{23}$$

Proof. The function n is equal to a constant n_∞ in the domain Ω_∞. Therefore we obtain (23) from (16)–(19) and (14). □

3. Reichardt condition. Because the domain Ω_∞ is unbounded, to have the problem formulation complete we have to specify the behavior of E and H at infinity. This can be done in various ways; for the problem under consideration the most general condition is the Reichardt condition [23], as discussed below. Denote by Ω_R a circle $\Omega_R = \left\{x \in \mathbb{R}^2 : |x| \le R\right\}$, and by Γ_R the boundary of Ω_R.

DEFINITION 3.1. *Let* R_0 *be a large positive constant such that* $\Omega \subset \Omega_{R_0}$. *We say that functions* E *and* H *satisfy the Reichardt condition if the functions* E *and* H *can be represented for all* $x \in \mathbb{R}^2 \setminus \Omega_{R_0}$ *as*

$$\begin{bmatrix} \mathrm{E} \\ \mathrm{H} \end{bmatrix} = \sum_{l=-\infty}^{\infty} \begin{bmatrix} \mathrm{A}_l \\ \mathrm{B}_l \end{bmatrix} H_l^{(1)}(\chi r) \exp(il\varphi), \tag{24}$$

where $H_l^{(1)}$ *is the Hankel function of the first kind and index* l *(see, e.g.,* [37]*),* (r,φ) *are the polar coordinates of the point* x, *and* $\chi = \sqrt{k^2 n_\infty^2 - \beta^2}$. *The series in* (3.1) *should converge uniformly and absolutely.*

DEFINITION 3.2. *By* Λ *denote the Riemann surface of the function* $\ln \chi(\beta)$. *A nonzero vector* $[\mathrm{E},\mathrm{H}] \in (C^2(\mathbb{R}^2))^6$ *is referred to as a* generalized eigenvector *(or eigenmode) of the problem* (2.3), (2.4), *and* (3.1) *corresponding to an eigenvalue* $\beta \in \Lambda$ *if the relations of problem* (2.3), (2.4), *and* (3.1) *are valid.*

In order to discuss the Reichardt condition in more detail, we need to analyze the Riemann surface Λ and consider the different types of modes that are possible.

3.1. Riemann surface Λ. The Hankel functions $H_l^{(1)}(\chi(\beta)r)$ are many-valued functions of the variable β. If we want to consider these functions as holomorphic functions, it is seen that β should be considered on the set Λ, which is the Riemann surface of the function $\ln \chi(\beta)$. This is due to the fact that Hankel functions can be represented as

$$H_l^{(1)}(\chi r) = c_l^{(1)}(\chi r) \ln(\chi r) + R_l^{(1)}(\chi r), \tag{25}$$

where $c_l^{(1)}(\chi r)$ and $R_l^{(1)}(\chi r)$ are holomorphic single-valued functions (see, e.g., [37]). The Riemann surface Λ is infinitely sheeted, with each sheet having two branch points, $\beta = \pm k n_\infty$. More precisely, due to the branching of $\chi(\beta)$ itself, we consider an infinite number of logarithmic branches Λ_m, $m = 0, \pm 1, \ldots$, each consisting of two square-root sheets of the complex variable β: $\Lambda_m^{(1)}$ and $\Lambda_m^{(2)}$. By $\Lambda_0^{(1)}$ denote the principal ("proper") sheet of Λ, which is specified by the conditions

$$-\pi/2 < \arg \chi(\beta) < \frac{3\pi}{2}, \quad \mathrm{Im}(\chi(\beta)) \ge 0, \quad \beta \in \Lambda_0^{(1)}. \tag{26}$$

The "improper" sheet $\Lambda_0^{(2)}$ is specified by the conditions

$$-\pi 2 < \arg \chi(\beta) < \frac{3\pi}{2}, \quad \mathrm{Im}(\chi(\beta)) < 0, \quad \beta \in \Lambda_0^{(2)}. \tag{27}$$

Denote also the whole real axis of $\Lambda_0^{(1)}$ as $R_0^{(1)}$, and that of $\Lambda_0^{(2)}$ as $R_0^{(2)}$. All the other pairs of sheets $\Lambda_{m\neq 0}^{(1),(2)}$ differ from $\Lambda_0^{(1),(2)}$ by a shift in $\arg\chi(\beta)$ equal to $2\pi m$, and satisfy the conditions

$$
\begin{aligned}
&-\pi/2+2\pi m<\arg\chi(\beta)<\frac{3\pi}{2}+2\pi m, \quad \operatorname{Im}(\chi(\beta))\geq 0, \quad \beta\in\Lambda_m^{(1)},\\
&-\pi/2+2\pi m<\arg\chi(\beta)<\frac{3\pi}{2}+2\pi m, \quad \operatorname{Im}(\chi(\beta))< 0, \quad \beta\in\Lambda_m^{(2)}.
\end{aligned}
\tag{28}
$$

Hence, on $\Lambda_0^{(1)}$ there is only a pair of branch-cuts dividing it from $\Lambda_0^{(2)}$; they run along the real axis at $|\beta|<kn_\infty$ and along the imaginary axis. On $\Lambda_0^{(2)}$, additionally, there is a pair of branch-cuts dividing it from $\Lambda_{\pm 1}^{(2)}$; they run along the real axis at $|\beta|>kn_\infty$.

3.2. Purely guided, complex, and leaky-wave modes. Denote a set of points on the real axis $R_0^{(1)}$ of the sheet $\Lambda_0^{(1)}$ by G, that is, the union of two intervals:

$$
G=\{\beta\in R_0^{(1)}: kn_\infty<|\beta|<kn_+\}. \tag{29}
$$

By $C_0^{(1)}$ denote the set

$$
C_0^{(1)}=\{\beta\in\Lambda_0^{(1)}:\operatorname{Re}\beta\neq 0\}\setminus R_0^{(1)}. \tag{30}
$$

Propagation constants β of purely guided modes, complex modes, and leaky-wave modes belong to sets $G\subset\Lambda_0^{(1)}$, $C_0^{(1)}\subset\Lambda_0^{(1)}$, and $\Lambda_0^{(2)}\setminus R_0^{(2)}$, respectively.

If $-\pi/2<\arg\chi<3\pi/2$, then the large-argument asymptotic forms of the Hankel functions of the first kind are known (see, e.g., [37]) to be

$$
H_l^{(1)}(\chi r)=\sqrt{\frac{2}{\pi\chi r}}\exp\left[i\left(\chi r-\frac{l\pi}{2}-\frac{\pi}{4}\right)\right]\left[1+O\left(\frac{1}{\chi r}\right)\right], \quad r\to\infty. \tag{31}
$$

Hence, if $-\pi/2<\arg\chi<3\pi/2$, $\operatorname{Im}(\chi)\neq 0$, and a function $[\mathrm{E},\mathrm{H}]$ satisfies the Reichardt condition, then this function satisfies the following condition at infinity:

$$
\begin{bmatrix}\mathrm{E}\\ \mathrm{H}\end{bmatrix}=\exp(i\chi r)\,O\left(\frac{1}{\sqrt{r}}\right), \quad r\to\infty. \tag{32}
$$

It is easy to see that for purely guided and complex modes, $\operatorname{Im}(\chi)>0$. Therefore corresponding eigenmodes $[\mathrm{E},\mathrm{H}]$ decay at infinity as $\exp(-\operatorname{Im}(\chi)r)r^{-1/2}$. Eigenvectors $[\mathrm{E},\mathrm{H}]$ of leaky-wave modes grow at infinity as $\exp(-\operatorname{Im}(\chi)r)r^{-1/2}$ because $\operatorname{Im}(\chi)<0$ for them.

3.3. Radiation modes. By D denote the set

$$
D=\{\beta\in\Lambda_0^{(1)}:\operatorname{Re}\beta=0\}\bigcup\{\beta\in R_0^{(1)}:|\beta|<kn_\infty\}. \tag{33}
$$

The continuous spectrum of radiation modes belongs to domain D, and each radiation mode can be expressed as (see [1])

$$
\begin{bmatrix}\mathrm{E}\\ \mathrm{H}\end{bmatrix}=\sum_{l=-\infty}^{\infty}\begin{bmatrix}\mathrm{A}_l\\ \mathrm{B}_l\end{bmatrix}H_l^{(1)}(\chi r)\exp(il\varphi)+\sum_{l=-\infty}^{\infty}\begin{bmatrix}\mathrm{C}_l\\ \mathrm{D}_l\end{bmatrix}H_l^{(2)}(\chi r)\exp(il\varphi), \tag{34}
$$

where $x \in \mathbb{R}^2 \setminus \Omega_{R_0}$ and $H_l^{(2)}$ is the Hankel function of the second kind and index l (see, e.g., [37]).

If $-\pi/2 < \arg\chi < 3\pi/2$, then the large-argument asymptotic forms of the Hankel functions of the second kind are known (see, e.g., [37]) to be

$$H_l^{(2)}(\chi r) = \sqrt{\frac{2}{\pi\chi r}} \exp\left[-i\left(\chi r - \frac{l\pi}{2} - \frac{\pi}{4}\right)\right]\left[1 + O\left(\frac{1}{\chi r}\right)\right], \quad r \to \infty. \tag{35}$$

It is easy to see that for radiation modes $\mathrm{Im}(\chi) = 0$, and that the radiation modes satisfy the following condition at infinity:

$$\begin{bmatrix} \mathrm{E} \\ \mathrm{H} \end{bmatrix} = O\left(\frac{1}{\sqrt{r}}\right), \quad r \to \infty. \tag{36}$$

The Reichardt condition (24) for all functions which satisfy (23) and all $\beta \in D$ is equivalent to the Sommerfeld condition

$$\left(\frac{\partial}{\partial r} - i\chi\right)\begin{bmatrix} \mathrm{E} \\ \mathrm{H} \end{bmatrix} = o\left(\frac{1}{\sqrt{r}}\right), \quad r \to \infty, \tag{37}$$

a fact which was proven in [38]. Therefore, radiation modes do not satisfy the Reichardt condition (24). In section 4 we will prove that the set D is free of the eigenvalues of problem (3), (4), and (24). In section 5, using the Reichardt condition (24), we will reduce problem (3), (4), and (24) to a problem with a purely point spectrum. Therefore, in this work we will not investigate the continuous spectrum of radiation modes.

3.4. Mode notation. The eigenvectors corresponding to the eigenvalues $\beta \in R_0^{(1)}$ such that $|\beta| < kn_\infty$ and satisfying the Sommerfeld condition (37) do not exist in a "passive" DW (i.e., when $\mathrm{Im}n^2 = 0$), which we investigate in this paper. However, if the waveguide is "active," i.e., if $\mathrm{Im}n^2 < 0$, then such modes, radiating to $r \to \infty$ (i.e., satisfying the Sommerfeld condition (37)) and propagating along x_3 without attenuation, may exist. In contrast, the eigenvectors corresponding to the eigenvalues $\beta \in G \subset R_0^{(1)}$ satisfy the condition of exponential decay at infinity. We suggest calling all natural modes generated by the real-axis eigenvalues *eigenmodes*, and, to distinguish between them, calling the first ones *radiating eigenmodes* and the second *guided-wave eigenmodes*. Note, however, that our radiating eigenmodes should not be confused with the "radiation modes" discussed in the previous section. Note that the condition (24) leads to a non–self-adjoint problem in general, which becomes self-adjoint if $\beta \in G$, i.e., for the guided-wave eigenmodes.

If $\beta \in \Lambda_0^{(1),(2)}$ but off $R_0^{(1)}$, then the corresponding modes will be called *quasi eigenmodes*: they consist of the exponentially decaying *"proper" complex quasi eigenmodes* if $\beta \in C_0^{(1)}$, the exponentially growing *leaky-wave quasi eigenmodes* if $\beta \in \Lambda_0^{(2)} \setminus R_0^{(2)}$, and the exponentially growing *"anti-guided" quasi eigenmodes* if $\beta \in R_0^{(2)}$ such that $|\beta| > kn_\infty$.

For all $m \neq 0$, $l = 0, \pm1, \pm2, \ldots$, and $\beta \in \bigcup_{m\neq0}(\Lambda_m^{(1)} \bigcup \Lambda_m^{(2)})$ we have

$$H_l^{(1)}(\chi \exp(i2\pi m)r) = \alpha_l^{(m)} H_l^{(1)}(\chi r) + \gamma_l^{(m)} H_l^{(2)}(\chi r), \qquad \alpha_l^{(m)}, \gamma_l^{(m)} \neq 0. \tag{38}$$

All of the modes whose wavenumbers are located on the higher-order pairs of sheets $\Lambda_{m\neq0}^{(1),(2)}$ will be collectively called *pseudoeigenmodes* because, according to (31), (35),

and (38), they are composed of a sum of incoming and outgoing cylindrical waves. Another justification of this terminology is that all of the possible eigenmodes in a "passive" DW are solutions of a self-adjoint problem, whereas quasi eigenmodes and pseudoeigenmodes satisfy non–self-adjoint problems.

The eigenvalues β on Λ possess a symmetry which is a consequence of equivalency between positive and negative directions along the longitudinal axis x_3 and time t (see [33]). Namely, if β is an eigenvalue and $[\mathrm{E},\mathrm{H}]$ is a corresponding generalized eigenvector, then $-\beta$ is also an eigenvalue, with the generalized eigenvector given by $[-\mathrm{E},\mathrm{H}]$. Further, because $\operatorname{Im}\omega = 0$ and $\operatorname{Im} n = 0$, the complex-conjugate numbers $\pm\overline{\beta}$ are eigenvalues as well, with the eigenvectors given by $\left[\mp\overline{\mathrm{E}}, -\overline{\mathrm{H}}\right]$. All these facts can be easily verified by direct substitution into (3), (4), and (24). We shall call the above-mentioned modes *forward*, *backward*, *conjugate*, and *backward-conjugate* modes, respectively.

4. Localization of the eigenvalues.

THEOREM 4.1. *The sets* $B = \{\beta \in R_0^{(1)} : |\beta| \geq kn_+\}$ *and* D *are free of the eigenvalues of problem* (2.3), (2.4), *and* (3.1).

Proof. Suppose that conditions (3), (4), and (24) are satisfied for some $[\mathrm{E},\mathrm{H}] \in (C^2(\mathbb{R}^2))^6$ and $\beta \in B$. Multiplying both sides of (17) by $\overline{\mathrm{H}}$, integrating over $\mathbb{R}^2$, and using (31), we obtain

$$
\begin{aligned}
k^2 \int_{\mathbb{R}^2} |\mathrm{H}|^2 dx &= \int_{\mathbb{R}^2} \left(\mathrm{Rot}_\beta \left(\frac{1}{n^2} \mathrm{Rot}_\beta \mathrm{H} \right), \overline{\mathrm{H}} \right) dx \\
&= \int_{\mathbb{R}^2} \left(\frac{1}{n^2} \mathrm{Rot}_\beta \mathrm{H}, \overline{\mathrm{Rot}_\beta \mathrm{H}} \right) dx \\
&\geq \frac{1}{n_+^2} \int_{\mathbb{R}^2} \left(\mathrm{Rot}_\beta \mathrm{H}, \overline{\mathrm{Rot}_\beta \mathrm{H}} \right) dx \\
&= \frac{1}{n_+^2} \int_{\mathbb{R}^2} \left(\mathrm{Rot}_\beta \left(\mathrm{Rot}_\beta \mathrm{H} \right), \overline{\mathrm{H}} \right) dx.
\end{aligned} \tag{39}
$$

Combining this with (19) and (14), we obtain

$$
\begin{aligned}
k^2 \int_{\mathbb{R}^2} |\mathrm{H}|^2 dx &\geq \frac{1}{n_+^2} \int_{\mathbb{R}^2} \left(-\Delta \mathrm{H} + \beta^2 \mathrm{H}, \overline{\mathrm{H}} \right) dx \\
&= \frac{1}{n_+^2} \int_{\mathbb{R}^2} |\operatorname{grad} \mathrm{H}|^2 dx + \frac{\beta^2}{n_+^2} \int_{\mathbb{R}^2} |\mathrm{H}|^2 dx.
\end{aligned} \tag{40}
$$

Therefore, we have

$$
\left(\beta^2 - k^2 n_+^2 \right) \int_{\mathbb{R}^2} |\mathrm{H}|^2 dx + \int_{\mathbb{R}^2} |\operatorname{grad} \mathrm{H}|^2 dx \leq 0. \tag{41}
$$

Hence, if $\beta \in B$ and $|\beta| > kn_+$, then $\mathrm{H} = 0$ for $x \in \mathbb{R}^2$, and

$$
\mathrm{E} = \frac{-1}{(i\omega\varepsilon_0 n^2)} \mathrm{Rot}_\beta \mathrm{H} = 0 \tag{42}
$$

for $x \in \mathbb{R}^2$. If $\beta \in B$ and $|\beta| = kn_+$, then the function H is equivalent to a constant in $\mathbb{R}^2$, but if H satisfies (24), then it must vanish at infinity for all $\beta \in B$. Therefore,

if $\beta \in B$ and $|\beta| = kn_+$, then $\mathrm{H} = 0$ for $x \in \mathbb{R}^2$, and $\mathrm{E} = 0$ for $x \in \mathbb{R}^2$. Therefore the vector $[\mathrm{E}, \mathrm{H}]$ is not an eigenvector of problem (3), (4), and (24) if $\beta \in B$.

Now suppose that conditions (3), (4), and (24) are satisfied for some $[\mathrm{E}, \mathrm{H}] \in (C^2(\mathbb{R}^2))^6$ and $\beta \in D$. Multiplying both sides of (16) by $\overline{\mathrm{E}}$, integrating over Ω_R where $R \geq R_0$, and using (14) and (18), we obtain

$$
\begin{aligned}
(43) \quad k^2 \int_{\Omega_R} n^2 \left|\mathrm{E}\right|^2 dx &= \int_{\Omega_R} \left(\mathrm{Rot}_\beta \left(\mathrm{Rot}_\beta \mathrm{E}\right), \overline{\mathrm{E}}\right) dx \\
&= \int_{\Omega_R} \left(-\Delta \mathrm{E} + \beta^2 \mathrm{E} + \mathrm{Grad}_\beta \left(\mathrm{Div}_\beta \mathrm{E}\right), \overline{\mathrm{E}}\right) dx \\
&= \int_{\Omega_R} \left|\mathrm{grad}\, \mathrm{E}\right|^2 dx - \int_{\Gamma_R} \left(\frac{\partial \mathrm{E}}{\partial \left|x\right|}, \overline{\mathrm{E}}\right) dx + \beta^2 \int_{\Omega_R} \left|\mathrm{E}\right|^2 dx \\
&\quad - \int_{\Omega_R} \left|\mathrm{Div}_\beta \mathrm{E}\right|^2 dx.
\end{aligned}
$$

For all $\beta \in D$ the number β^2 is real, and therefore we have

$$
(44) \qquad \mathrm{Im} \int_{\Gamma_R} \left(\frac{\partial \mathrm{E}}{\partial \left|x\right|}, \overline{\mathrm{E}}\right) dx = 0, \quad R \geq R_0.
$$

If we combine this with (24), we obtain

$$
(45) \qquad 2\pi\chi R \sum_{l=-\infty}^{\infty} \mathrm{Im} \left[H_l^{(2)} \left(\chi R\right) H_l^{(1)'} \left(\chi R\right)\right] \left|\mathrm{A}_l\right|^2 = 0, \quad R \geq R_0.
$$

We also have

$$
(46) \qquad \mathrm{Im} \left[H_l^{(2)} \left(\chi R\right) H_l^{(1)'} \left(\chi R\right)\right] = \frac{2}{\pi \chi R}, \quad l = 0, \pm 1, \pm 2, \ldots,
$$

which leads to $A_l = 0$ for all l and any $R \geq R_0$. Hence $\mathrm{E} = 0$ for $r \geq R_0$. Under the assumption of the smoothness of the function n, we have $\mathrm{E} = 0$ for $x \in \Omega_{R_0}$ (see [39, p. 190]) and

$$
(47) \qquad \mathrm{H} = \frac{1}{(i\omega\mu_0)} \mathrm{Rot}_\beta \mathrm{E} = 0
$$

for $x \in \mathbb{R}^2$. Therefore the vector $[\mathrm{E}, \mathrm{H}]$ is not an eigenvector of problem (3), (4), and (24) if $\beta \in D$. The proof of the theorem is complete. □

5. Discreteness and dependence of the eigenvalues on parameters. Now we shall prove that the set of all eigenvalues of problem (3), (4), and (24) can be only a set of isolated points on Λ. We shall also investigate the behavior of eigenvalues β of the problem (3), (4), and (24) as functions of parameters $n_\infty \in \mathbb{R}_+$ and $\omega \in \mathbb{R}_+$, where $\mathbb{R}_+$ is the set of all positive numbers, $\mathbb{R}_+ = \{x > 0\}$. We will use general results of the theory of operator-valued functions [36]. The results in [36] were obtained for operators of the form $I + \mathcal{B}(\beta)$, where I is the identity operator and the operator $\mathcal{B}(\beta)$ is compact for all β. Therefore we shall reduce the problem (3), (4), and (24) to a nonlinear spectral problem with a compact integral operator.

LEMMA 5.1. *Suppose that* $[\mathrm{E},\mathrm{H}]$ *is an eigenvector of the problem* (2.3), (2.4), *and* (3.1) *corresponding to an eigenvalue* $\beta \in \Lambda$. *Then*

$$\mathrm{E}(x) = \left(\mathcal{B}(\beta)\mathrm{E}\right)(x), \quad x \in \mathbb{R}^2, \tag{48}$$

where

$$\begin{aligned}\left(\mathcal{B}(\beta)\mathrm{E}\right)(x) = k^2 \int_\Omega \left(n^2(y) - n_\infty^2\right) \Phi(\beta; x, y)\mathrm{E}(y)dy \\ + \mathrm{Grad}_\beta \int_\Omega (\mathrm{E}, n^{-2}\mathrm{grad}n^2)(y)\Phi(\beta; x, y)dy,\end{aligned} \tag{49}$$

$$\Phi\left(\beta; x, y\right) = \frac{i}{4} H_0^{(1)}\left(\chi(\beta)\left|x - y\right|\right). \tag{50}$$

Proof. For all $\beta \in \Lambda$ and $x \in \mathbb{R}^2$ we have

$$\mathrm{E}(x) = \left(k^2 n_\infty^2 + \mathrm{Grad}_\beta \mathrm{Div}_\beta\right) \frac{1}{n_\infty^2} \int_\Omega \left(n^2(y) - n_\infty^2\right) \Phi(\beta; x, y)\mathrm{E}(y)dy. \tag{51}$$

This result is well known for $\beta \in G$ (see, e.g., [40]). The desired assertion for all $\beta \in \Lambda$ is obtained by applying the method of Green functions to the vector Helmholtz equation for the electric field with the use of the relation

$$\int_{\Gamma_R} \left(\frac{\partial \mathrm{E}(y)}{\partial |y|} \Phi(\beta; x, y) - \frac{\partial \Phi(\beta; x, y)}{\partial |y|} \mathrm{E}(y)dy \right) = 0, \quad R \geq R_0, \tag{52}$$

which is valid for any $\beta \in \Lambda$ and an arbitrary function E satisfying the Reichardt condition (24). The validity of relation (52) was proved in [38], [23].

By the supposition of the lemma, $\mathrm{E} \in (C^2(\mathbb{R}^2))^3$. The function n is twice continuously differentiable in $\mathbb{R}^2$ too. Therefore, the following divergence relation is valid:

$$\begin{aligned}\mathrm{Div}_\beta \int_\Omega \left(n^2(y) - n_\infty^2\right) \Phi(\beta; x, y)\mathrm{E}(y)dy \\ = \int_\Omega \mathrm{Div}_\beta \left[\left(n^2(y) - n_\infty^2\right) \mathrm{E}(y)\right] \Phi(\beta; x, y)dy, \quad x \in \mathbb{R}^2.\end{aligned} \tag{53}$$

Taking into account (53) and (20), we obtain the assertion of the lemma. □

For all $\beta \in \Lambda$ the operator $\mathcal{B}(\beta)$ determined by (49) will be considered as an operator in the space of complex-valued functions $[L_2(\Omega)]^3$. By definition, set

$$\mathcal{A}(\beta) = I - \mathcal{B}(\beta),$$

where I is the identity operator in $[L_2(\Omega)]^3$. The kernel of the integral operator $\mathcal{B}(\beta)$ is weakly singular for all $\beta \in \Lambda$, and the domain Ω has a Lipschitz boundary. Therefore, the operator $\mathcal{B}(\beta)$ is compact for all $\beta \in \Lambda$ (see, e.g., [41]).

DEFINITION 5.2. *A nonzero vector* $\mathrm{F} \in [L_2(\Omega)]^3$ *is called an eigenvector of an operator-valued function* $\mathcal{A}(\beta)$ *corresponding to an eigenvalue* $\beta \in \Lambda$ *if the relation*

$$\mathcal{A}(\beta)\mathrm{F} = 0 \tag{54}$$

is valid. The set of all $\beta \in \Lambda$ *for which the operator* $\mathcal{A}(\beta)$ *does not have the bounded inverse operator in* $[L_2(\Omega)]^3$ *is called the spectrum of problem* (5.7)

Next we shall prove a theorem on the spectral equivalence of the problem (3), (4), and (24) and the problem (54), but before doing this we consider the following.

DEFINITION 5.3. *A nonzero vector* $u \in C^2(\mathbb{R}^2)$ *is called a* generalized eigenvector *of the problem*

$$\left[\Delta + \left(k^2 n^2 - \beta^2\right)\right] u = 0, \quad x \in \mathbb{R}^2, \tag{55}$$

$$u = \sum_{l=-\infty}^{\infty} a_l H_l^{(1)}(\chi r) \exp(il\varphi) \quad \text{for all } r \geq R_0 \tag{56}$$

(where the series is supposed to converge uniformly and absolutely), corresponding to an eigenvalue $\beta \in \Lambda$ *if the relations* (5.8) *and* (5.9) *are valid.*

LEMMA 5.4. *The set of all eigenvalues of problem* (5.8) *and* (5.9) *can only be a set of isolated points on* Λ. *The sheet* $\Lambda_0^{(1)}$, *except for the set* G, *is free of the eigenvalues of the problem* (5.8) *and* (5.9).

The proof is found in [42]. Note that the solutions of problem (55) and (56) represent the solutions of the weak-guidance approximation of the original problem (3), (4), and (24).

THEOREM 5.5. *Suppose that* $[\mathrm{E}, \mathrm{H}] \in (C^2(\mathbb{R}^2))^6$ *is an eigenvector of the problem* (2.3), (2.4), *and* (3.1) *corresponding to an eigenvalue* $\beta_0 \in \Lambda$. *Then* $\mathrm{F} = \mathrm{E} \in [L_2(\Omega)]^3$ *is an eigenvector of the operator-valued function* $\mathcal{A}(\beta)$ *corresponding to the same eigenvalue* β_0. *Suppose that* $\mathrm{F} \in [L_2(\Omega)]^3$ *is an eigenvector of the operator-valued function* $\mathcal{A}(\beta)$ *corresponding to an eigenvalue* $\beta_0 \in \Lambda$, *and also suppose that the same number* β_0 *is not an eigenvalue of the problem* (5.8) *and* (5.9). *Let* $\mathrm{E} = \mathcal{B}(\beta_0)\mathrm{F}$ *and* $\mathrm{H} = (i\omega\mu_0)^{-1}\mathrm{Rot}_{\beta_0}\mathrm{E}$ *for* $x \in \mathbb{R}^2$. *Then* $[\mathrm{E}, \mathrm{H}] \in (C^2(\mathbb{R}^2))^6$, *and* $[\mathrm{E}, \mathrm{H}]$ *is an eigenvector of the problem* (2.3), (2.4), *and* (3.1) *corresponding to the same eigenvalue* β_0.

Proof. From Lemma 5.1 we obtain the first assertion of the theorem. Now we shall prove the second assertion of the theorem. Suppose that $\mathrm{F} \in [L_2(\Omega)]^3$ is an eigenvector of the operator-valued function $\mathcal{A}(\beta)$ corresponding to an eigenvalue $\beta \in \Lambda$. Assume $\mathrm{E} = \mathcal{B}(\beta)\mathrm{F}$ for $x \in \mathbb{R}^2$. The kernel of the integral operator $\mathcal{B}(\beta)$ is weakly singular for any $\beta \in \Lambda$. By virtue of the well-known property of the integral operator with weakly singular kernel on the domain with a Lipschitz boundary (see, e.g., [41]), we have $\mathrm{E} \in [\mathrm{C}(\overline{\Omega})]^3$. The function n belongs to the space of twice continuously differentiable functions in $\mathbb{R}^2$. By virtue of the well-known properties of the area potential (see, e.g., [41]), we have $\mathrm{E} \in [\mathrm{C}^2(\mathbb{R}^2)]^3$.

Applying the operator Div_β to both sides of (48), and using (10) and (53), we obtain

$$\begin{aligned} \mathrm{Div}_\beta \mathrm{E}(x) = k^2 \int_\Omega \mathrm{Div}_\beta \left[\left(n^2(y) - n_\infty^2\right) \mathrm{E}(y)\right] \Phi(\beta; x, y) dy \\ + \left(\Delta - \beta^2\right) \int_\Omega (\mathrm{E}, n^{-2}\mathrm{grad} n^2)(y) \Phi(\beta; x, y) dy \end{aligned} \tag{57}$$

for all $x \in \mathrm{R}^2$. If we combine this with Poisson's formula

$$\left(\Delta + k^2 n_\infty^2 - \beta^2\right) \int_\Omega \left(n^2(y) - n_\infty^2\right) \Phi(\beta; x, y) \mathrm{E}(y) dy = -\left(n^2(x) - n_\infty^2\right) \mathrm{E}(x), \tag{58}$$

we get

$$\mathrm{Div}_\beta \mathrm{E}(x) = k^2 \int_\Omega \mathrm{Div}_\beta \left[\left(n^2(y) - n_\infty^2\right) \mathrm{E}(y)\right] \Phi(\beta; x, y) dy - k^2 n_\infty^2 \int_\Omega (\mathrm{E}, n^{-2}\mathrm{grad} n^2)(y) \Phi(\beta; x, y) dy - (E, n^{-2}\mathrm{grad} n^2)(x) \tag{59}$$

for all $x \in \mathrm{R}^2$. Using (12), we have

$$\mathrm{Div}_\beta \left[\left(n^2 - n_\infty^2\right) \mathrm{E}\right] = \mathrm{Div}_\beta \left(n^2 \mathrm{E}\right) - n_\infty^2 \mathrm{Div}_\beta \mathrm{E}, \tag{60}$$

$$(\mathrm{E}, n^{-2}\mathrm{grad} n^2) = n^{-2}\mathrm{Div}_\beta \left(n^2 \mathrm{E}\right) - \mathrm{Div}_\beta \mathrm{E}. \tag{61}$$

If we combine this with (59), we see that the function $u = n^{-2}\mathrm{Div}_\beta \left(n^2 \mathrm{E}\right)$ satisfies

$$u = \int_\Omega k^2 \left(n^2(y) - n_\infty^2\right) \Phi(\beta; x, y) u(y) dy, \quad x \in \mathbb{R}^2.$$

If the number β is not an eigenvalue of the problem (55) and (56), then this equation has only the trivial solution (see [42]). Therefore, we have

$$\mathrm{Div}_\beta \left(n^2 \mathrm{E}\right) = 0, \quad x \in \mathbb{R}^2. \tag{62}$$

Using this, (48), and (61), for $x \in \mathbb{R}^2$, we obtain

$$\mathrm{E}(x) = k^2 \int_\Omega \left(n^2(y) - n_\infty^2\right) \Phi(\beta; x, y) \mathrm{E}(y) dy - \mathrm{Grad}_\beta \int_\Omega \Phi(\beta; x, y) \mathrm{Div}_\beta \mathrm{E}(y) dy. \tag{63}$$

Assume $\mathrm{H} = (i\omega\mu_0)^{-1}\mathrm{Rot}_\beta \mathrm{E}$, $x \in \mathbb{R}^2$; i.e., $[\mathrm{E}, \mathrm{H}]$ satisfies (3). Combining (63) and (13), we have

$$\mathrm{H}(x) = -i\omega\varepsilon_0 \mathrm{Rot}_\beta \int_\Omega \left(n^2(y) - n_\infty^2\right) \Phi(\beta; x, y) \mathrm{E}(y) dy, \quad x \in \mathbb{R}^2. \tag{64}$$

Therefore, if $\mathrm{E} \in [\mathrm{C}^2(\mathbb{R}^2)]^3$, then $\mathrm{H} \in [\mathrm{C}^2(\mathbb{R}^2)]^3$.

Now we shall prove that $[\mathrm{E}, \mathrm{H}]$ satisfies (4). Multiplying both sides of (63) by $i\omega\varepsilon_0 n_\infty^2$, applying the operator Rot_β to both sides of (64), and combining the results, we obtain

$$\mathrm{Rot}_\beta \mathrm{H} + i\omega\varepsilon_0 n_\infty^2 \mathrm{E} = -i\omega\varepsilon_0 \mathrm{Rot}_\beta \mathrm{Rot}_\beta \int_\Omega \left(n^2(y) - n_\infty^2\right) \Phi(\beta; x, y) \mathrm{E}(y) dy + i\omega\varepsilon_0 n_\infty^2 k^2 \int_\Omega \left(n^2(y) - n_\infty^2\right) \Phi(\beta; x, y) \mathrm{E}(y) dy - i\omega\varepsilon_0 n_\infty^2 \mathrm{Grad}_\beta \int_\Omega \Phi(\beta; x, y) \mathrm{Div}_\beta \mathrm{E}(y) dy \tag{65}$$

for all $x \in \mathbb{R}^2$. If we combine this with (14) and (53), we obtain

$$\mathrm{Rot}_\beta \mathrm{H} + i\omega\varepsilon_0 n_\infty^2 \mathrm{E} = i\omega\varepsilon_0 \left[\Delta + (k^2 n_\infty^2 - \beta^2)\right] \int_\Omega \left(n^2(y) - n_\infty^2\right) \Phi(\beta; x, y) \mathrm{E}(y) dy - i\omega\varepsilon_0 \mathrm{Grad}_\beta \int_\Omega \mathrm{Div}_\beta \left[\left(n^2(y) - n_\infty^2\right) \mathrm{E}(y)\right] \Phi(\beta; x, y) dy - i\omega\varepsilon_0 n_\infty^2 \mathrm{Grad}_\beta \int_\Omega \Phi(\beta; x, y) \mathrm{Div}_\beta \mathrm{E}(y) dy$$

for all $x \in \mathbb{R}^2$. Using this, (62), and (58), we have

$$\mathrm{Rot}_\beta \mathrm{H} + i\omega\varepsilon_0 n_\infty^2 \mathrm{E} = -i\omega\varepsilon_0 \left(n^2 - n_\infty^2\right) \mathrm{E}, \quad x \in \mathbb{R}^2. \tag{66}$$

Therefore $[\mathrm{E}, \mathrm{H}]$ satisfies (4).

Using the Bessel function addition theorem (see, e.g., [37]), we can readily prove that the number β and the vector $[\mathrm{E}, \mathrm{H}]$ satisfy condition (24). The proof of the theorem is complete. □

THEOREM 5.6. *The set of all eigenvalues of the problem* (2.3), (2.4), *and* (3.1) *can be only a set of isolated points on* Λ. *Each eigenvalue* β *of the problem* (2.3), (2.4), *and* (3.1) *depends continuously on* $(\omega, n_\infty) \in \mathbb{R}^2_+$ *and can appear and disappear only at the boundary of* Λ, *i.e., at* $\beta = \pm k n_\infty$ *and at infinity on* Λ.

Proof. For any $(x, y) \in \Omega^2$ and any $(\omega, n_\infty) \in \mathbb{R}^2_+$ the kernel of the operator $\mathcal{A}(\beta)$ is analytic in $\beta \in \Lambda$. Hence, the operator-valued function $\mathcal{A}(\beta)$ is holomorphic in $\beta \in \Lambda$ for any $(\omega, n_\infty) \in \mathbb{R}^2_+$. The operator-valued function $\mathcal{A}(\beta; \omega, n_\infty)$ is jointly continuous in $(\beta; \omega, n_\infty) \in \Lambda \times \mathbb{R}^2_+$. For all $(\beta; \omega, n_\infty) \in \Lambda \times \mathbb{R}^2_+$ the operator $\mathcal{B}(\beta; \omega, n_\infty)$ is compact. Therefore, using Theorems 4.1 and 5.5 and Lemma 5.4, we see that the operator $\mathcal{A}(\beta; \omega, n_\infty)$ has a bounded inverse operator in $[L_2(\Omega)]^3$ for all $\beta \in B \bigcup D$ and $(\omega, n_\infty) \in \mathbb{R}^2_+$. Hence, for each $(\omega, n_\infty) \in \mathbb{R}^2_+$ the spectrum of problem (54) can be only a set of isolated points on Λ, which are the eigenvalues of the operator-valued function $\mathcal{A}(\beta)$; each eigenvalue β of the operator-valued function $\mathcal{A}(\beta)$ depends continuously on $(\omega, n_\infty) \in \mathbb{R}^2_+$ and can appear and disappear only at the boundary of Λ, i.e., at $\beta = \pm k n_\infty$ and at infinity on Λ (see [36]). Using Theorem 5.5, we obtain the assertion of the current theorem, which is now complete. □

REFERENCES

[1] D. MARCUSE, *Theory of Dielectric Optical Waveguides*, Academic Press, New York, 1974.

[2] A. W. SNYDER AND J. D. LOVE, *Optical Waveguide Theory*, Chapman and Hall, London, 1983.

[3] B. Z. KATSENELENBAUM, *Symmetric and non-symmetric excitation of infinite dielectric cylinder*, Zhurnal Tekhnicheskoi Fiziki, 19 (1949), pp. 1168–1181 (in Russian).

[4] T. ROZZI AND M. MONGIARDO, *Open Electromagnetic Waveguides*, The Institution of Electrical Engineers, London, 1997.

[5] A. W. SNYDER, *Leaky-ray theory of optical waveguides of circular cross section*, Appl. Phys., 4 (1974), pp. 273–298.

[6] C. M. MILLER, *Optical Fiber Splices and Connectors: Theory and Methods*, Marcel Dekker, New York, 1986.

[7] F. WILCZEWSKI, *Bending loss of leaky modes in optical fibers with arbitrary index profiles*, Optics Letters, 19 (1994), pp. 1031–1033.

[8] A. W. SNYDER AND A. ANKIEWICZ, *Anisotropic fibers with nonaligned optical (stress) axes*, J. Opt. Soc. Amer. A, 3 (1986), pp. 856–863.

[9] M. LU AND M. M. FEJER, *Anisotropic dielectric waveguides*, J. Opt. Soc. Amer. A, 10 (1993), pp. 246–261.

[10] R. SAMMUT AND A. W. SNYDER, *Leaky modes on circular optical waveguides*, Applied Optics, 15 (1976), pp. 477–482.

[11] R. SAMMUT AND A. W. SNYDER, *Leaky modes on a dielectric waveguide: Orthogonality and excitation*, Applied Optics, 15 (1976), pp. 1040–1044.

[12] R. SAMMUT, *A comparison of leaky mode and leaky ray analysis of circular optical fibers*, J. Opt. Soc. Amer. A, 66 (1976), pp. 370–371.

[13] G. N. VESELOV AND S. B. RAEVSKIY, *On the spectrum of complex waves of the circular dielectric waveguide*, Radiotekhnika, 2 (1983), pp. 55–58 (in Russian).

[14] T. F. JABLONSKI, *Complex modes in open lossless dielectric waveguides*, J. Opt. Soc. Amer. A, 11 (1994), pp. 1272–1282.

[15] L. A. LYUBIMOV, G. I. VESELOV, AND N. A. BEI, *Dielectric waveguide of elliptic cross-section*, Radiotekhnika i Elektronika, 6 (1961), pp. 1871–1880 (English translation in Radio Engineering Electronic Physics).

[16] T. F. Jablonski and M. J. Sowinski, *Analysis of dielectric guiding structures by the iterative eigenfunction expansion method*, IEEE Trans. Microwave Theory Techniques, MTT-37 (1989), pp. 63–70.

[17] N. N. Voitovich, B. Z. Katsenelenbaum, A. N. Sivov, and A. D. Shatrov, Natural waves of dielectric waveguides of complicated cross-sections, Radiotekhnika i Elektronika, 24 (1979), pp. 1245–1263 (English translation in Radio Engineering Electronic Physics).

[18] A. N. Kleyev, A. B. Manenkov, and A. G. Rozhnev, *Numerical methods of calculating dielectric waveguides or fiber lightguides*, J. Commun. Technol. Electronics (English translation), 39 (1994), pp. 90–115.

[19] J. S. Bagby, D. P. Nyquist, and B. C. Drachman, *Integral formulation for analysis of integrated dielectric waveguides*, IEEE Trans. Microwave Theory Tech., MTT-29 (1985), pp. 906–915.

[20] J. M. van Splunter, H. Blok, N. H. G. Baken, and M. F. Dane, *Computational analysis of propagation properties of integrated-optical waveguides using a domain integral equation*, in Proceedings of the URSI International Symposium on EM Theory, Budapest, 1986, URSI, Ghent, Belgium, pp. 321–323.

[21] H. P. Urbach, *Analysis of the domain integral operator for anisotropic dielectric waveguides*, SIAM J. Math. Anal., 27 (1996), pp. 204–220.

[22] A. Bamberger and A. S. Bonnet, *Mathematical analysis of the guided modes of an optical fiber*, SIAM J. Math. Anal., 21 (1990), pp. 1487–1510.

[23] H. Reichardt, *Ausstrahlungsbedingungen fur die wellengleihung*, Abh. Math. Sem. Univ. Hamburg, 24 (1960), pp. 41–53.

[24] S. V. Sukhinin, *On the discreteness of natural frequencies of open acoustic resonators*, Dynamics of Continuous Media, 49 (1981), pp. 157–163 (in Russian).

[25] A. Y. Poyedinchuk, Y. A. Tuchkin, and V. P. Shestopalov, *On the regularization of spectral problems of the wave scattering by non-closed screens*, Soviet Phys. Dokl., 295 (1987), pp. 1358–1362 (in Russian).

[26] Yu. V. Shestopalov, Yu. G. Smirnov, and E. V. Chernokozhin, *Logarithmic Integral Equations in Electromagnetics*, VSP, Leiden, The Netherlands, 2000.

[27] A. I. Nosich, A. Y. Poedinchuk, and V. P. Shestopalov, *Discrete spectrum of the characteristic waves in open partially screened dielectric core*, Soviet Phys. Dokl. (English translation), 30 (1985), pp. 669–671.

[28] A. I. Nosich and A. Y. Svezhentsev, *Accurate computation of mode characteristics for open-layered circular cylindrical microstrip and slot lines*, Microwave and Optical Technology Lett., 4 (1991), pp. 274–277.

[29] A. I. Nosich and A. Y. Svezhentsev, *Principal and higher order modes of microstrip and slot lines on a cylindrical substrate*, Electromagnetics, 13 (1993), pp. 85–94.

[30] A. I. Nosich, *Radiation conditions, limiting absorption principle, and general relations in open waveguide scattering*, J. Electromag. Waves Applicat., 8 (1994), pp. 329–353.

[31] E. M. Karchevskii, *Analysis of the eigenmode spectra of dielectric waveguides*, J. Comput. Math. Math. Phys., 39 (1999), pp. 1493–1498.

[32] E. M. Karchevskii, *The fundamental wave problem for cylindrical dielectric waveguides*, Differential Equations, 36 (2000), pp. 1109–1111.

[33] A. I. Nosich, *On correct formulation and general properties of wave scattering by discontinuities in open waveguides*, in Proceedings of the International Conference on Mathematical Methods in Electromagnetic Theory (MMET 90), Gurzuf, 1990, Test-Radio Publishing, pp. 100–112.

[34] C. Muller, *Grundproblems der Mathematischen Theorie Elektromagnetischer Schwingungen*, Springer, Berlin, 1957.

[35] D. Colton and R. Kress, *Time harmonic electromagnetic waves in an inhomogeneous medium*, Proc. Roy. Soc. Edinburgh, 116A (1990), pp. 279–293.

[36] S. Steinberg, *Meromorphic families of compact operators*, Arch. Ration. Mech. Anal., 31 (1968), pp. 372–379.

[37] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1965.

[38] I. N. Vekua, *On metaharmonic functions*, Trudy Tbilisskogo Mat. Inst., 12 (1943), pp. 105–174 (in Russian).

[39] L. Hormander, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, 1976.

[40] R. F. Harrington, *Time-Harmonic Electromagnetic Fields*, McGraw–Hill, New York, 1961.

[41] V. S. Vladimirov, *Equations of Mathematical Physics*, Marcel Dekker, New York, 1971.

[42] E. M. Karchevskii and S. I. Solov'ev, *Investigation of a spectral problem for the Helmholtz operator on the plane*, Differential Equations, 36 (2000), pp. 631–634.

SIAM J. APPL. MATH.
Vol. 65, No. 6, pp. 2049–2066

INVERSE MEDIUM SCATTERING PROBLEMS FOR ELECTROMAGNETIC WAVES*

GANG BAO† AND PEIJUN LI†

Abstract. Consider a time-harmonic electromagnetic plane wave incident on a medium enclosed by a bounded domain in $\mathbb{R}^3$. In this paper, existence and uniqueness of the variational problem for forward scattering are established. An energy estimate for the scattered field with a uniform bound with respect to the wavenumber is obtained in the case of low frequency on which the Born approximation is based. A continuation method for the inverse medium scattering problem, which reconstructs the scatterer of an inhomogeneous medium from boundary measurements of the scattered wave, is developed. The algorithm requires multifrequency scattering data. Using an initial guess from the Born approximation, each update is obtained via recursive linearization on the wavenumber k by solving one forward problem and one adjoint problem of Maxwell's equations.

Key words. inverse medium scattering, Maxwell's equations, recursive linearization

AMS subject classifications. 65N21, 78A46

DOI. 10.1137/040607435

1. Introduction.

Consider the systems of time-harmonic Maxwell's equations in three dimensions

$$\nabla \times E^t = \mathrm{i}\omega\mu^* H^t, \tag{1.1}$$

$$\nabla \times H^t = -\mathrm{i}\omega\varepsilon^* E^t, \tag{1.2}$$

where E^t and H^t are the total electric field and magnetic field, respectively; $\omega > 0$ is the frequency; and ε^* and μ^* are the electric permittivity and the magnetic permeability, respectively. Denote by $\varepsilon_0 > 0$, $\mu_0 > 0$ the permittivity and permeability of the vacuum. The fields are further assumed to be nonmagnetic; i.e., $\mu^* = \mu_0$. Rewriting $\varepsilon^* = \varepsilon_0\varepsilon$, $\varepsilon = 1 + q(x)$ is the relative permittivity, where $q(x)$ is the scatterer, which is assumed to have a compact support, and $\Re(q(x)) > -1$.

Taking the curl of (1.1) and eliminating the magnetic field H^t, we obtain the uncoupled equation for the electric field E^t:

$$\nabla \times (\nabla \times E^t) - k^2\varepsilon E^t = 0, \tag{1.3}$$

where $k = \omega\sqrt{\varepsilon_0\mu_0}$ is called the wavenumber, satisfying $0 < k_{\min} \le k \le k_{\max} < \infty$. The total electric field E^t consists of the incident field E^i and the scattered field E:

$$E^t = E^i + E.$$

Assume that the incident field is a plane wave of the normalized form [5]

$$E^i = \mathrm{i}k\vec{p}\,e^{\mathrm{i}kx\cdot\vec{n}}, \tag{1.4}$$

*Received by the editors April 27, 2004; accepted for publication (in revised form) March 11, 2005; published electronically August 9, 2005. This research was supported in part by NSF grant DMS 01-04001 and ONR grant N000140210365. Preliminary progress of this research was announced in [3].

http://www.siam.org/journals/siap/65-6/60743.html

†Department of Mathematics, Michigan State University, East Lansing, MI 48824-1027 (bao@math.msu.edu, lipeijun@math.msu.edu).

where $\vec{n} \in \mathbb{S}^2$ is the propagation direction and $\vec{p} \in \mathbb{S}^2$ is the polarization satisfying $\vec{p} \cdot \vec{n} = 0$. Evidently, such an incident wave satisfies the homogeneous equation

$$\nabla \times (\nabla \times E^i) - k^2 E^i = 0. \tag{1.5}$$

It follows from (1.3) and (1.5) that the scattered field satisfies

$$\nabla \times (\nabla \times E) - k^2 \varepsilon E = k^2 q(x) E^i. \tag{1.6}$$

In addition, the scattered field is required to satisfy the following Silver–Müller radiation condition:

$$\lim_{r \to \infty} r \left[\nabla \times E \times \frac{x}{r} - \mathrm{i}kE \right] = 0,$$

where $r = |x|$. In practice, it is convenient to reduce the problem to a bounded domain by introducing an artificial surface. Let Ω be the compact support of the scatterer $q(x)$. Assume that $R > 0$ is a constant such that the support of the scatterer, Ω, is included in the ball $B = \{x \in \mathbb{R}^3 : |x| < R\}$. Let S be the sphere of the ball, i.e., $S = \{x \in \mathbb{R}^3 : |x| = R\}$. Denote by ν the outward unit normal to S. A suitable boundary condition then has to be imposed on S. For simplicity, we employ the first order absorbing boundary condition (impedance boundary condition) [12] as

$$\nu \times (\nabla \times E) + \mathrm{i}k\nu \times (\nu \times E) = 0 \quad \text{on } S. \tag{1.7}$$

Given the incident field E^i, the forward problem is to determine the scattered field E for the known scatterer $q(x)$, which is assumed further to be in $L^\infty(B)$. Based on the Helmholtz decomposition and a compact imbedding result, the forward problem is shown to have a unique solution for all but possibly a discrete set of wavenumbers. Furthermore, an energy estimate for the scattered field, with a uniform bound with respect to the wavenumber, is given in the low frequency case. The estimate provides a theoretical basis for our linearization algorithm. For numerical solution of the forward scattering problem in an open domain, the reader is referred to [14, 15, 16, 21] and references therein. The inverse medium scattering problem is to determine the scatterer $q(x)$ from the measurements of near field current densities, the tangential trace of the scattered field $\nu \times E|_S$, given the incident field. Although this is a classical problem in inverse scattering theory, little is known on reconstruction methods, especially in the three dimensional case, due to the nonlinearity, ill-posedness, and large scale computation associated with the inverse scattering problem. We refer the reader to [1, 6, 10, 11, 23] for related results on the inverse medium problem. See [5] for an account of recent progress on the general inverse scattering problem.

The goal of this work is to present a recursive linearization method that solves the inverse medium scattering problem of Maxwell's equations in three dimensions. The reader is referred to [2, 4] for recursive linearization approaches for solving the inverse medium scattering problems in two dimensions. Our algorithm requires multi-frequency scattering data, and the recursive linearization is obtained by a continuation method on the wavenumber. It first solves a linear equation (Born approximation) at the lowest wavenumber, which may be done by using the fast Fourier transform (FFT). Updates are subsequently obtained by using higher and higher wavenumbers. Following the idea of the Kaczmarz method [6, 18, 19], we use partial data to perform the nonlinear Landweber iteration at each wavenumber. For each iteration, one forward and one adjoint state of Maxwell's equations are solved, which may be implemented by using the symmetric second order edge (Nédélec) elements.

The plan of this paper is as follows. Analysis of the variational problem for forward scattering is presented in section 2. Based on the Helmholtz decomposition, a compact imbedding result, and the Lax–Milgram lemma, the well-posedness of the forward scattering is proved. An important energy estimate is given. Section 3 is devoted to the numerical study of inverse medium scattering. Using the initial guess of the reconstruction derived from the Born approximation, a regularized iterative linearization algorithm is proposed. Numerical examples are presented in section 4. The paper is concluded with some remarks and future directions in section 5.

2. Analysis of the variational problem for forward scattering. In this section, the variational formulation for the forward scattering problem is discussed. The analysis provides a criterion for weak scattering, which plays an important role in the inversion algorithm.

To state our boundary value problem, following [17], we first introduce the standard Sobolev spaces:

$$
\begin{aligned}
&L_t^2(S) = \{u \in (L^2(S))^3 : \nu \cdot u = 0 \ \text{ on } S\},\\
&H_0^1(B) = \{u \in H^1(B) : u = 0 \ \text{ on } S\},\\
&H(\mathrm{curl}, B) = \{u \in (L^2(B))^3 : \nabla \times u \in (L^2(B))^3\},\\
&H_{\mathrm{imp}}(\mathrm{curl}, B) = \{u \in H(\mathrm{curl}, B) : \nu \times u \in L_t^2(S)\},
\end{aligned}
$$

where $H_{\mathrm{imp}}(\mathrm{curl}, B)$ is an appropriate subspace of $H(\mathrm{curl}, B)$ for solving problems involving the impedance boundary condition. Correspondingly, these spaces are equipped with the norms

$$
\begin{aligned}
&\| u \|_{L_t^2(S)} = \| u \|_{(L^2(S))^3},\\
&\| u \|_{H^1(B)}^2 = \| u \|_{L^2(B)}^2 + \| \nabla u \|_{(L^2(B))^3}^2,\\
&\| u \|_{H(\mathrm{curl},B)}^2 = \| u \|_{(L^2(B))^3}^2 + \| \nabla \times u \|_{(L^2(B))^3}^2,\\
&\| u \|_{H_{\mathrm{imp}}(\mathrm{curl},B)}^2 = \| u \|_{H(\mathrm{curl},B)}^2 + \| \nu \times u \|_{L_t^2(S)}^2.
\end{aligned}
$$

For convenience, denote the $(L^2(B))^3$ and $(L^2(S))^3$ inner products by

$$
(u, v) = \int_B u \cdot \overline{v} dx \quad \text{and} \quad \langle u, v \rangle = \int_S u \cdot \overline{v} ds,
$$

respectively, where the overline denotes the complex conjugate. Introduce the bilinear form $a : H_{\mathrm{imp}}(\mathrm{curl}, B) \times H_{\mathrm{imp}}(\mathrm{curl}, B) \to \mathbb{C}$,

$$
a(E, \phi) = (\nabla \times E, \nabla \times \phi) - k^2(\varepsilon E, \phi) + \mathrm{i}k\langle \nu \times E, \nu \times \phi \rangle,
$$

and the linear functional on $H_{\mathrm{imp}}(\mathrm{curl}, B)$,

$$
b(\phi) = k^2(qE^i, \phi).
$$

Then we have the weak form of the boundary value problem (1.6) and (1.7): find $E \in H_{\mathrm{imp}}(\mathrm{curl}, B)$ such that

$$
a(E, \phi) = b(\phi) \quad \forall\, \phi \in H_{\mathrm{imp}}(\mathrm{curl}, B). \tag{2.1}
$$

Throughout the paper, C stands for a positive generic constant, whose value may change step by step but should always be clear from the context.

Before presenting the main result for the variational problem, we state several useful lemmas. The reader is referred to [17] for detailed discussions and proofs.

LEMMA 2.1 (Helmholtz decomposition). *The spaces X and Y are closed subspaces of $H_{\rm imp}(\mathrm{curl}, B)$, which is the direct sum of the spaces X and Y; i.e.,*

$$H_{\rm imp}(\mathrm{curl}, B) = X \oplus Y.$$

Here

$$X = \{u \in H_{\rm imp}(\mathrm{curl}, B) : \mathrm{div}(\varepsilon u) = 0 \quad \text{in } B\}$$

and

$$Y = \{\nabla \xi : \xi \in H_0^1(B)\}.$$

LEMMA 2.2 (compact imbedding). *The space X is compactly imbedded into the space $(L^2(B))^3$.*

LEMMA 2.3 (Friedrichs inequality). *There exists a positive constant C, independent of the wavenumber, such that for all $u \in X$*

$$\| u \|_{(L^2(B))^3} \leq C \left(\| \nabla \times u \|_{(L^2(B))^3} + \| \nu \times u \|_{(L^2(S))^3} \right).$$

Next we prove the well-posedness of the variational problem (2.1) and obtain an energy estimate for the scattered field with a uniform bound with respect to the wavenumber in the case of low frequency.

THEOREM 2.1. *If the wavenumber is sufficiently small, the variational problem* (2.1) *admits a unique weak solution in $H_{\mathrm{imp}(\mathrm{curl}, B)}$ given by $E = u + \nabla p$, while $u \in X, p \in H_0^1(B)$. Furthermore, we have the estimate*

$$\| E \|_{H_{\mathrm{imp}(\mathrm{curl}, B)}} \leq Ck|\Omega|^{1/2} \| q \|_{L^\infty(B)}, \tag{2.2}$$

where the constant C is independent of k and Ω is the compact support of the scatterer.

Proof. Using the Helmholtz decomposition, we take $E = u + \nabla p$ and $\phi = v + \nabla \xi$, for any $v \in X, \xi \in H_0^1(B)$. Observe that $a(u, \nabla \xi) = 0$, for any $\xi \in H_0^1(B)$, by the definition of X. Therefore, we decompose the variational equation (2.1) into the form

$$a(u, v) + a(\nabla p, v) + a(\nabla p, \nabla \xi) = b(v) + b(\nabla \xi) \quad \forall\, v \in X,\ \xi \in H_0^1(B). \tag{2.3}$$

First, we determine $p \in H_0^1(B)$ by the solution of

$$a(\nabla p, \nabla \xi) = b(\nabla \xi) \quad \forall\, \xi \in H_0^1(B),$$

which gives explicitly

$$-(\varepsilon \nabla p, \nabla \xi) = (qE^i, \nabla \xi) \quad \forall\, \xi \in H_0^1(B).$$

The existence and uniqueness of the solution p in $H_0^1(B)$ may be proved by a direct application of the Lax–Milgram lemma with the estimate

$$\| \nabla p \|_{(L^2(B))^3} \leq Ck|\Omega|^{1/2} \| q \|_{L^\infty(B)}. \tag{2.4}$$

Rewrite (2.3) as

$$a(u, v) = b(v) - a(\nabla p, v) \quad \forall\, v \in X, \tag{2.5}$$

and decompose the bilinear form a into $a = a_1 + k^2 a_2$, where

$$a_1(u,v) = (\nabla \times u, \nabla \times v) + \mathrm{i}k\langle \nu \times u, \nu \times v\rangle,$$
$$a_2(u,v) = -(\varepsilon u, v).$$

Using the inequality of arithmetic and geometric means, we conclude from Lemma 2.3 that a_1 is coercive:

$$|a_1(u,u)| \geq Ck(\| \nabla \times u \|^2_{(L^2(B))^3} + \| \nu \times u \|^2_{(L^2(S))^3}) \geq Ck \| u \|^2_{H_{\mathrm{imp}}(\mathrm{curl},B)} \quad \forall\, u \in X.$$

The continuity of the bilinear form a_1 follows from the Cauchy–Schwarz inequality.

Next we prove the compactness of a_2. Define an operator $\mathcal{A} : (L^2(B))^3 \to X$ by

$$a_1(\mathcal{A}u, v) = a_2(u,v) \quad \forall\, v \in X,$$

which gives

$$(\nabla \times \mathcal{A}u, \nabla \times v) + \mathrm{i}k\langle \nu \times \mathcal{A}u, \nu \times v\rangle = -(\varepsilon u, v) \quad \forall\, v \in X.$$

Using the Lax–Milgram lemma again, it follows that

$$\| \mathcal{A}u \|_{H_{\mathrm{imp}}(\mathrm{curl},B)} \leq \frac{C}{k} \| u \|_{(L^2(B))^3}, \tag{2.6}$$

where the constant C is independent of k. Thus $\mathcal{A}$ is bounded from $(L^2(B))^3$ to X, and X is compactly imbedded into $(L^2(B))^3$. Hence $\mathcal{A} : (L^2(B))^3 \to (L^2(B))^3$ is a compact operator.

Define a function $w \in (L^2(B))^3$ by requiring $w \in X$ and satisfying

$$a_1(w,v) = b(v) - a(\nabla p, v) \quad \forall\, v \in X.$$

More specifically, we have by using the Stokes formula that

$$a_1(w,v) = k^2(qE^i, v) + k^2(\varepsilon \nabla p, v) \quad \forall\, v \in X.$$

It follows from the Lax–Milgram lemma that

$$\| w \|_{H_{\mathrm{imp}}(\mathrm{curl},B)} \leq C(k^2|\Omega|^{1/2} \| q \|_{L^\infty(B)} + k \| \nabla p \|_{(L^2(B))^3}).$$

An application of (2.4) yields

$$\| w \|_{H_{\mathrm{imp}}(\mathrm{curl},B)} \leq Ck^2|\Omega|^{1/2} \| q \|_{L^\infty(B)}. \tag{2.7}$$

Using the operator $\mathcal{A}$, we can see that the problem (2.5) is equivalent to finding $u \in (L^2(B))^3$ such that

$$(\mathcal{I} + k^2\mathcal{A})u = w. \tag{2.8}$$

When the wavenumber k is small enough, the operator $\mathcal{I} + k^2\mathcal{A}$ has a uniformly bounded inverse. We then have the estimate

$$\| u \|_{(L^2(B))^3} \leq C \| w \|_{(L^2(B))^3}, \tag{2.9}$$

where the constant C is independent of k. However, rearranging (2.8), we have $u = w - k^2\mathcal{A}u$, so $u \in X$ and, by the estimate (2.6) for the operator $\mathcal{A}$, we have

$$\| u \|_{H_{\rm imp}({\rm curl},B)} \le \| w \|_{H_{\rm imp}({\rm curl},B)} + Ck \| u \|_{(L^2(B))^3}.$$

Combining the estimates (2.9) and (2.7) leads to

$$\| u \|_{H_{\rm imp}({\rm curl},B)} \le Ck^2|\Omega|^{1/2} \| q \|_{L^\infty(B)}. \tag{2.10}$$

Finally, it follows from the definition of the norm in $H_{\rm imp}({\rm curl}, B)$ that

$$\| E \|_{H_{\rm imp}({\rm curl},B)} \le \| u \|_{H_{\rm imp}({\rm curl},B)} + \| \nabla p \|_{(L^2(B))^3}.$$

The proof is complete by noting the estimates (2.10) and (2.4) for sufficiently small wavenumbers. □

Remark 2.1. The energy estimate of the scattered field (2.2) provides a criterion for weak scattering. From this estimate, it is easily seen that, fixing any two of the three quantities, i.e., the wavenumber, the compact support of the scatterer Ω, and the $L^\infty(B)$ norm of the scatterer, the scattering is weak when the third one is small. Especially for the given scatterer $q(x)$, i.e., the norm and the compact support are fixed, the scattering is weak when the wavenumber is small.

Remark 2.2. For a general wavenumber, from (2.8) the uniqueness and existence follow from the Fredholm alternative. If the scatterer $q(x)$ is more regular, say of $C_0^2(B)$ [8], unique continuation may be used to prove the uniqueness and thus the existence of the forward scattering problem (1.6), (1.7) for all $k > 0$. Otherwise, if k^2 is not the eigenvalue for Maxwell's equations in the domain B, then the operator $\mathcal{I} + k^2\mathcal{A}$ has a bounded inverse. However, the bound depends on the wavenumber. Therefore, the constant C in the estimate (2.2) depends on the wavenumber.

From the above discussion, we have the following theorem on the well-posedness of the variational problem (2.1).

THEOREM 2.2. *Given the scatterer $q \in L^\infty(B)$, for all but possibly a discrete set of wavenumbers, the variational problem* (2.1) *admits a unique weak solution in $H_{\rm imp}({\rm curl}, B)$, given by $E = u + \nabla p$, while $u \in X$, $p \in H_0^1(B)$.*

3. Inverse medium scattering. In this section, a regularized recursive linearization method for solving the inverse medium scattering problem of Maxwell's equations in three dimensions is proposed. The algorithm, obtained by a continuation method on the wavenumber, requires multifrequency scattering data. At each wavenumber, the algorithm determines a forward model which produces the prescribed scattering data. At a low wavenumber, the scattered field is weak. Consequently, the nonlinear equation becomes essentially linear, known as the Born approximation. The algorithm first solves this nearly linear equation at the lowest wavenumber to obtain low-frequency modes of the true scatterer. The approximation is then used to linearize the nonlinear equation at the next higher wavenumber to produce a better approximation which contains more modes of the true scatterer. This process is continued until a sufficiently high wavenumber, where the dominant modes of the scatterer are essentially recovered.

3.1. Low-frequency modes of the scatterer. Rewrite (1.6) as

$$\nabla \times (\nabla \times E) - k^2 E = k^2 q(x)(E^i + E), \tag{3.1}$$

where the incident wave is taken as $E^i = \mathrm{i}k\vec{p}_1 e^{\mathrm{i}kx\cdot\vec{n}_1}$. Consider a test function $F = \mathrm{i}k\vec{p}_2 e^{\mathrm{i}kx\cdot\vec{n}_2}$, where $\vec{p}_2, \vec{n}_2 \in \mathbb{S}^2$ satisfy $\vec{p}_2 \cdot \vec{n}_2 = 0$. Hence F satisfies (1.5).

Multiplying (3.1) by F and integrating over B on both sides, we have

$$\int_B F \cdot [\nabla \times (\nabla \times E)]dx - k^2 \int_B F \cdot E dx = k^2 \int_B q(x) F \cdot E^i dx + k^2 \int_B q(x) F \cdot E dx.$$

Integration by parts yields

$$\int_B E \cdot [\nabla \times (\nabla \times F)]dx + \int_S \big[E \times (\nabla \times F) - F \times (\nabla \times E)\big] \cdot \nu ds - k^2 \int_B F \cdot E dx = k^2 \int_B q(x) F \cdot E^i dx + k^2 \int_B q(x) F \cdot E dx.$$

We have, by noting (1.5),

$$\int_S \big[E \times (\nabla \times F) - F \times (\nabla \times E)\big] \cdot \nu ds = k^2 \int_B q(x) F \cdot E^i dx + k^2 \int_B q(x) F \cdot E dx.$$

Using the boundary condition (1.7) of the scattered field and the special form of the incident wave E^i and F, we get

$$-\int_S (\nu \times E) \cdot (\vec{n}_2 \times \vec{p}_2) e^{\mathrm{i}kx\cdot\vec{n}_2} ds + \int_S \big[\nu \times (\nu \times E)\big] \cdot \vec{p}_2 e^{\mathrm{i}kx\cdot\vec{n}_2} ds = \int_B q(x) F \cdot E^i dx + \int_B q(x) F \cdot E dx.$$

A simple calculation yields

$$\int_B q(x) e^{\mathrm{i}kx\cdot(\vec{n}_1+\vec{n}_2)} dx = \frac{1}{(\vec{p}_1 \cdot \vec{p}_2)k^2} \int_S (\nu \times E) \cdot (\vec{n}_2 \times \vec{p}_2 + \nu \times \vec{p}_2) e^{\mathrm{i}kx\cdot\vec{n}_2} ds + \frac{\mathrm{i}}{(\vec{p}_1 \cdot \vec{p}_2)k} \int_B q(x) \vec{p}_2 \cdot E e^{\mathrm{i}kx\cdot\vec{n}_2} dx. \tag{3.2}$$

From Theorem 2.1 and Remark 2.1, for a small wavenumber, the scattered field is weak and the inverse scattering problem becomes essentially linear. Dropping the nonlinear (second) term of (3.2), we obtain the linearized integral equation

$$\int_B q_0(x) e^{\mathrm{i}kx\cdot(\vec{n}_1+\vec{n}_2)} dx = \frac{1}{(\vec{p}_1 \cdot \vec{p}_2)k^2} \int_S (\nu \times E) \cdot (\vec{n}_2 \times \vec{p}_2 + \nu \times \vec{p}_2) e^{\mathrm{i}kx\cdot\vec{n}_2} ds, \tag{3.3}$$

which is the Born approximation. The function $q_0(x)$ will be used as the starting point for our recursive linearization algorithm.

Since the scatterer $q_0(x)$ has a compact support, we use the notation

$$\hat{q}_0(\xi) = \int_B q_0(x) e^{\mathrm{i}kx\cdot(\vec{n}_1+\vec{n}_2)} dx,$$

where $\hat{q}_0(\xi)$ is the Fourier transform of $q_0(x)$ with $\xi = k(\vec{n}_1 + \vec{n}_2)$. Choose

$$\vec{n}_j = (\sin\theta_j \cos\phi_j, \sin\theta_j \sin\phi_j, \cos\theta_j), \quad j = 1, 2,$$

where θ_j, ϕ_j are the latitudinal and longitudinal angles, respectively. It is obvious that the domain $[0,\pi]\times[0,2\pi]$ of (θ_j,ϕ_j), $j=1,2$, corresponds to the ball $\{\xi\in\mathbb{R}^3 : |\xi|\le 2k\}$. Thus, the Fourier modes of $\hat{q}_0(\xi)$ in the ball $\{\xi : |\xi|\le 2k\}$ can be determined. The scattering data with the higher wavenumber must be used in order to recover more modes of the true scatterer.

Define the data

$$G(\zeta)=\begin{cases} \frac{1}{(\vec{p}_1\cdot\vec{p}_2)k^2}\int_S(\nu\times E)\cdot(\vec{n}_2\times\vec{p}_2+\nu\times\vec{p}_2)e^{\mathrm{i}kx\cdot\vec{n}_2}ds & \text{for } |\zeta|\le 2k,\\ 0, & \text{otherwise,}\end{cases}$$

where $\zeta=\zeta(k,\theta_1,\phi_1,\theta_2,\phi_2)\in\mathbb{R}^3$. The linear integral equation (3.3) can then be formally reformulated as

$$\int_{\mathbb{R}^3} q_0(x)e^{\mathrm{i}x\cdot\zeta}dx = G(\zeta). \tag{3.4}$$

Taking the inverse Fourier transform of (3.4) leads to

$$\frac{1}{(2\pi)^3}\int_{\mathbb{R}^3} e^{-\mathrm{i}x\cdot\zeta}\left[\int_{\mathbb{R}^3} q_0(y)e^{\mathrm{i}y\cdot\zeta}dy\right]d\zeta = \frac{1}{(2\pi)^3}\int_{\mathbb{R}^3} e^{-\mathrm{i}x\cdot\zeta}G(\zeta)d\zeta.$$

By the Fubini theorem, we have

$$\frac{1}{(2\pi)^3}\int_{\mathbb{R}^3} q_0(y)\left[\int_{\mathbb{R}^3} e^{\mathrm{i}(y-x)\cdot\zeta}d\zeta\right]dy = \frac{1}{(2\pi)^3}\int_{\mathbb{R}^3} e^{-\mathrm{i}x\cdot\zeta}G(\zeta)d\zeta.$$

Using the inverse Fourier transform of the Dirac delta function

$$\frac{1}{(2\pi)^3}\int_{\mathbb{R}^3} e^{\mathrm{i}(y-x)\cdot\zeta}d\zeta = \delta(y-x),$$

we deduce

$$\int_{\mathbb{R}^3} q_0(y)\delta(y-x)dy = \frac{1}{(2\pi)^3}\int_{\mathbb{R}^3} e^{-\mathrm{i}x\cdot\xi}G(\xi)d\xi,$$

which gives

$$q_0(x) = \frac{1}{(2\pi)^3}\int_{\mathbb{R}^3} e^{-\mathrm{i}x\cdot\zeta}G(\zeta)d\zeta. \tag{3.5}$$

In practice, the integral equation (3.5) is implemented by using the FFT.

3.2. Recursive linearization. As discussed in the previous section, when the wavenumber is small, the Born approximation allows a reconstruction of those Fourier modes less than or equal to $2k$ for the function $q(x)$. We now describe a procedure that recursively determines q_k at $k=k_j$ for $j=1,2,\ldots$ with increasing wavenumbers. Suppose now that the scatterer $q_{\tilde{k}}$ has been recovered at some wavenumber $\tilde{k}$, and that the wavenumber k is slightly larger than $\tilde{k}$. We wish to determine q_k, or equivalently, to determine the perturbation

$$\delta q = q_k - q_{\tilde{k}}.$$

For the reconstructed scatterer $q_{\tilde{k}}$, we solve at the wavenumber k the forward scattering problem

$$\nabla \times (\nabla \times \tilde{E}) - k^2(1 + q_{\tilde{k}})\tilde{E} = k^2 q_{\tilde{k}} E^i, \quad x \in B, \tag{3.6}$$

$$\nu \times (\nabla \times \tilde{E}) + \mathrm{i}k\nu \times (\nu \times \tilde{E}) = 0 \quad \text{on } S. \tag{3.7}$$

For the scatterer q_k, we have

$$\nabla \times (\nabla \times E) - k^2(1 + q_k)E = k^2 q_k E^i, \quad x \in B, \tag{3.8}$$

$$\nu \times (\nabla \times E) + \mathrm{i}k\nu \times (\nu \times E) = 0 \quad \text{on } S. \tag{3.9}$$

Subtracting (3.6), (3.7) from (3.8), (3.9) and omitting the second order smallness in δq and in $\delta E = E - \tilde{E}$, we obtain

$$\nabla \times (\nabla \times \delta E) - k^2(1 + q_{\tilde{k}})\delta E = k^2 \delta q (E^i + \tilde{E}), \quad x \in B, \tag{3.10}$$

$$\nu \times (\nabla \times \delta E) + \mathrm{i}k\nu \times (\nu \times \delta E) = 0 \quad \text{on } S. \tag{3.11}$$

For the scatterer q_k and the incident wave E^i, we define the map $S(q_k, E^i)$ by

$$S(q_k, E^i) = E,$$

where E is the scattered field at the wavenumber k. Let γ be the trace operator to the boundary S of the ball B. Define the scattering map

$$M(q_k, E^i) = \gamma S(q_k, E^i).$$

It is easily seen that the scattering map $M(q_k, E^i)$ is linear with respect to E^i but is nonlinear with respect to q_k. For simplicity, denote $M(q_k, E^i)$ by $M(q_k)$. By the definition of the trace operator, we have

$$M(q_k) = \nu \times E|_S.$$

We refer to [1] for the Fréchet differentiability of the scattering map. Let $DM(q_{\tilde{k}})$ be the Fréchet derivative of $M(q_k)$, and denote the residual operator

$$R(q_{\tilde{k}}) = \nu \times \delta E|_S.$$

It follows from [1] that

$$DM(q_{\tilde{k}})\delta q = R(q_{\tilde{k}}). \tag{3.12}$$

The regularized least-squares solution of (3.12) is

$$\delta q = [\alpha I + DM^*(q_{\tilde{k}})DM(q_{\tilde{k}})]^{-1} DM^*(q_{\tilde{k}})R(q_{\tilde{k}}),$$

where $DM^*(q_{\tilde{k}})$ is the adjoint operator of $DM(q_{\tilde{k}})$, I is the identity operator, and α is some suitable positive number. In practice, the main difficulty is the enormous computational cost of solving linear systems with huge full matrix. Here, we consider an alternative way of solving (3.12) which is much less computationally demanding.

To state the approach, we first examine the boundary data $\nu \times E(x; \theta, \phi; k)$. Here, the variable x is the observation point, which has two degrees of freedom since it is on the sphere S. The terms θ, ϕ are latitudinal and longitudinal angles, respectively, of the incident wave E^i. At each frequency, we have four degrees of freedom, and thus

data redundancy, which may be addressed by fixing one of the incident angles, say θ. Define $\phi_j = (j-1) * \frac{2\pi}{m}, j = 1, \ldots, m$, and the residual operator

$$R_j(q_{\tilde{k}}) = \nu \times E(x; \theta, \phi_j; k)|_S - \nu \times \tilde{E}(x; \theta, \phi_j; k)|_S,$$

where m is the total number of the incident waves or sweeps, and $\tilde{E}(x; \theta, \phi_j; k)$ is the solution of (3.6), (3.7) with the incident wave of longitudinal angle ϕ_j and the scatterer $q_{\tilde{k}}$. Instead of solving (3.12) for all incident waves simultaneously, we may solve it for one incident wave at a time while updating the residual operator after each determination of the incremental correction δq. Thus, for each incident wave with incident angle ϕ_j, we consider the equation

$$M_j(q_k) = \nu \times E(x; \theta, \phi_j; k)|_S, \tag{3.13}$$

where $M_j(q_k)$ is the scattering map corresponding to the incident wave with longitudinal angle ϕ_j. It follows from [1] that

$$DM_j(q_{\tilde{k}})\delta q_j = R_j(q_{\tilde{k}}), \tag{3.14}$$

where $DM_j(q_{\tilde{k}})$ is the Fréchet derivative of the scattering map $M_j(q_k)$. The nonlinear Landweber iteration for (3.13) yields

$$\delta q_j = \beta_k DM_j^*(q_{\tilde{k}}) R_j(q_{\tilde{k}}), \tag{3.15}$$

where $DM_j^*(q_{\tilde{k}})$ is the adjoint operator of $DM_j(q_{\tilde{k}})$, and β_k is some relaxation parameter [7].

Remark 3.1. For a fixed wavenumber, the stopping index of nonlinear Landweber iteration (3.15) could be determined from the discrepancy principle. However, in practice, it is not necessary to do many iterations. Our numerical results indicate that the iterative process for different incident angles $\phi_j, j = 1, \ldots, m$, is sufficient to obtain reasonable accuracy.

Next, we discuss the role of the relaxation parameter β_k in the iteration (3.15), which may be understood more clearly by considering the iteration from a different point of view.

Consider the optimization problem of (3.13),

$$\min_{q_k} \parallel M_j(q_k) - \nu \times E(x; \theta, \phi_j; k) \parallel^2_{(L^2(S))^3}. \tag{3.16}$$

The first order optimality condition for the problem (3.16) is given by

$$DM_j^*(q_{\tilde{k}}) \left(M_j(q_k) - \nu \times E(x; \theta, \phi_j; k)\right)|_S = 0. \tag{3.17}$$

To solve the optimality equation (3.17), the time marching scheme proposed in [22] consists of finding the steady state of the following parabolic equation:

$$\frac{dq_k}{dt} = DM_j^*(q_{\tilde{k}}) \left(\nu \times E(x; \theta, \phi_j; k) - M_j(q_k)\right)|_S.$$

The numerical solution could be computed from the explicit method

$$\delta q_j = \tau DM_j^*(q_{\tilde{k}}) R_j(q_{\tilde{k}}),$$

where τ is the discretized time step. Thus, the relaxation parameter β_k is essentially the step size of time marching, whose length is restricted by the stability of the explicit method.

In order to compute the correction δq_j, we need some efficient way to compute $DM_j^*(q_{\tilde{k}})R_j(q_{\tilde{k}})$, which is given by the following theorem.

THEOREM 3.1. *Given the residual $R_j(q_{\tilde{k}})$, there exits a function F_j satisfying the adjoint equations*

$$\nabla \times (\nabla \times F_j) - k^2(1 + \overline{q_{\tilde{k}}})F_j = 0, \quad x \in B, \tag{3.18}$$

$$\nabla \times F_j - \mathrm{i}k\nu \times F_j = R_j(q_{\tilde{k}}) \quad \text{on } S, \tag{3.19}$$

such that the adjoint Fréchet derivative $DM_j^*(q_{\tilde{k}})$ satisfies

$$\big[DM_j^*(q_{\tilde{k}})R_j(q_{\tilde{k}})\big](x) = k^2(\overline{E_j^i}(x) + \overline{\tilde{E}_j}(x)) \cdot F_j(x), \tag{3.20}$$

where E_j^i is the incident wave with the longitudinal angle ϕ_j and $\tilde{E}_j$ is the solution of (3.6), (3.7) *with the incident wave E_j^i.*

Proof. Let $\tilde{E}_j$ be the solution of (3.6), (3.7) with the incident wave E_j^i. Consider the equations as follows,

$$\nabla \times (\nabla \times \delta E) - k^2(1 + q_{\tilde{k}})\delta E = k^2 \delta q(E_j^i + \tilde{E}_j), \quad x \in B, \tag{3.21}$$

$$\nu \times (\nabla \times \delta E) + \mathrm{i}k\nu \times (\nu \times \delta E) = 0 \quad \text{on } S, \tag{3.22}$$

and the adjoint equations (3.18) and (3.19), which take the variational form

$$(\nabla \times F_j, \nabla \times \phi) - k^2((1 + \overline{q_{\tilde{k}}})F_j, \phi) - \mathrm{i}k\langle \nu \times F_j, \nu \times \phi\rangle = \langle R_j(q_{\tilde{k}}), \nu \times \phi\rangle \quad \forall \phi \in H_{\mathrm{imp}}(\mathrm{curl}, B).$$

The existence and uniqueness of the weak solution for the adjoint equations may be proved in the same way as for the scattered field. The proof is omitted.

Multiplying (3.21) with the complex conjugate of F_j and integrating over B on both sides, we obtain

$$\int_B \overline{F}_j \cdot \big[\nabla \times (\nabla \times \delta E)\big]dx - k^2 \int_B (1 + q_{\tilde{k}})\overline{F}_j \cdot \delta E dx = k^2 \int_B \delta q(E_j^i + \tilde{E}_j) \cdot \overline{F}_j dx.$$

Integration by parts yields

$$\int_S \big[\delta E \times (\overline{\nabla \times F_j}) - \overline{F}_j \times (\nabla \times \delta E)\big] \cdot \nu ds = k^2 \int_B \delta q(E_j^i + \tilde{E}_j) \cdot \overline{F}_j dx.$$

Using the boundary condition (3.22), we deduce

$$\int_S (\nu \times \delta E) \cdot (\overline{\nabla \times F_j + \mathrm{i}k\nu \times F_j})ds = k^2 \int_B \delta q(E_j^i + \tilde{E}) \cdot \overline{F}_j dx.$$

It follows from (3.14) and the boundary condition (3.19) that

$$\int_S [DM_j(q_{\tilde{k}})\delta q] \cdot \overline{R_j(q_{\tilde{k}})}ds = k^2 \int_B \delta q(E_j^i + \tilde{E}_j) \cdot \overline{F}_j dx.$$

We know from the adjoint operator $DM_j^*(q_{\tilde{k}})$ that

$$\int_B \delta q\, \overline{DM_j^*(q_{\tilde{k}})R_j(q_{\tilde{k}})}dx = k^2 \int_B \delta q(E_j^i + \tilde{E}_j) \cdot \overline{F}_j dx.$$

TABLE 1
Recursive linearization reconstruction algorithm for inverse medium scattering.

Initialization:
  $k = k_{\min}$ smallest $k_{\min}$
  q_0 Born approximation
Reconstruction loop:
  FOR $k = k_{\min} : k_{\max}$ march along wavenumbers
    FOR $j = 1 : m$ perform m sweeps over incident angles
      solve (3.6)–(3.7) for $\tilde{E}_j$ one forward problem
      solve (3.18)–(3.19) for F_j one adjoint problem
      $\delta q_k^j = \beta_k k^2 (\overline{E_j^i} + \overline{\tilde{E}_j}) \cdot F_j$
      $q_k^j := q_k^j + \delta q_k^j$
    END
    $q_k := q_k^m$
  END
  $q := q_{k_{\max}}$ final reconstruction

Since this holds for any δq, we have

$$\overline{DM_j^*(q_{\tilde{k}}) R_j(q_{\tilde{k}})} = k^2 (E_j^i + \tilde{E}_j) \cdot \overline{F}_j.$$

Taking the complex conjugate of the above equation yields the result. □

Using this theorem, we can rewrite (3.15) as

$$\delta q_j = \beta_k k^2 (\overline{E_j^i}(x) + \overline{\tilde{E}_j}(x)) \cdot F_j(x). \tag{3.23}$$

Thus, for each incident wave with a longitudinal angle ϕ_j, we solve one forward problem (3.6), (3.7) and one adjoint problem (3.18), (3.19). Since the adjoint problem has a variational form similar to that of the forward problem, we need to compute essentially two forward problems at each sweep. Once δq_j is determined, $q_{\tilde{k}}$ is updated by $q_{\tilde{k}} + \delta q_j$. After completing the mth sweep, we get the reconstructed scatterer q_k at the wavenumber k.

The recursive linearization for inverse medium scattering of Maxwell's equations can be summarized in Table 1.

4. Numerical experiments. In this section, we discuss the numerical solution of the forward scattering problem and the computational issues of the recursive linearization algorithm.

As for the forward solver, we adopt the edge elements which were developed originally for the finite element solution of Maxwell's equations [20, 12] in the early 1980s. From the mathematical point of view, these are natural approximation spaces for the Hilbert space $H(\mathrm{curl}, B)$, which is the adequate functional space for the variational formulation of Maxwell's equations. Vector fields in such finite element (FE) spaces have continuous tangential traces, which is consistent with the physics. Therefore, the natural degrees of freedom for these elements are related to tangential traces along edges or faces. Here, we take the symmetric second order edge elements for tetrahedral edge elements [13]. When the unknowns are ordered according to the reverse Cuthill–McKee (RCM) ordering [9], the profile of FE matrix is highly banded, which improves the condition number of the FE coefficient matrix. The sparse large scale linear system can be most efficiently solved if the zero elements of the coefficient matrix are not stored. We use the commonly used compressed row storage (CRS)

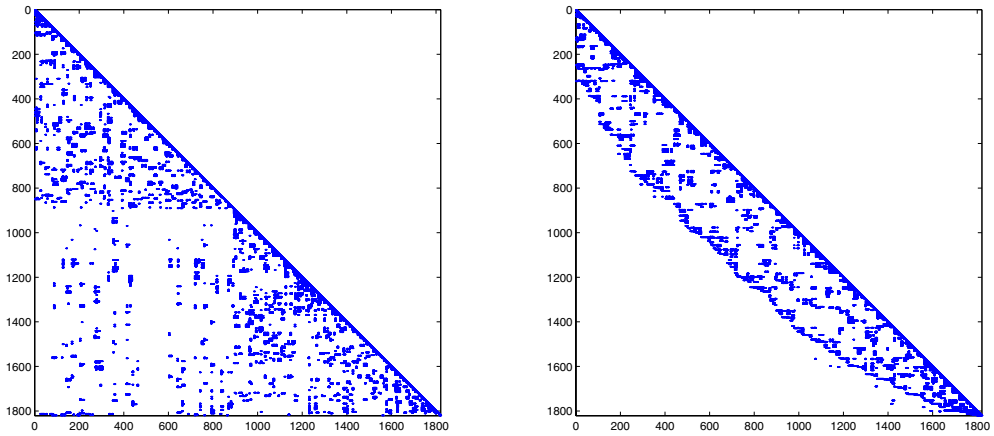


FIG. 1. *Sparsity pattern of an FE matrix with* 1820 *unknowns:* (a) *original ordering,* (b) *RCM ordering.*

format, which makes no assumptions about the sparsity structure of the matrix and does not store any unnecessary elements. In fact, from the variational formula of our direct problem (2.1), the coefficient matrix is complex symmetric. Hence, only the lower triangular portion of the matrix needs to be stored. Figure 1 shows a typical sparsity pattern of an FE matrix with 1820 unknowns from the symmetric second order edge element. Regarding the linear solver, either biconjugate gradient (BiCG) or quasi-minimal residual (QMR) algorithms with diagonal preconditioning may be employed to solve the sparse, symmetric, and complex system of the equations. It appears for our experiments that the QMR is more efficient.

In the following, we present two numerical examples where the number of the incident wave $m = 20$, the incident latitudinal angle $\theta = 0$, and the incident longitudinal angle $\phi_j = (j-1) * \frac{2\pi}{m}, j = 1, \ldots, m$. The relaxation parameter β_k is taken to be $0.1/k$ for the tested examples. For stability analysis, some relative random noise is added to the data; i.e., the tangential trace of the electric field takes the form

$$\nu \times E|_S := (1 + \sigma \,\mathrm{rand}) \cdot (\nu \times E|_S).$$

Here, rand gives uniformly distributed random numbers in $[-1, 1]$, and σ is a noise level parameter taken to be 0.02 in our numerical experiments. Define the relative error by

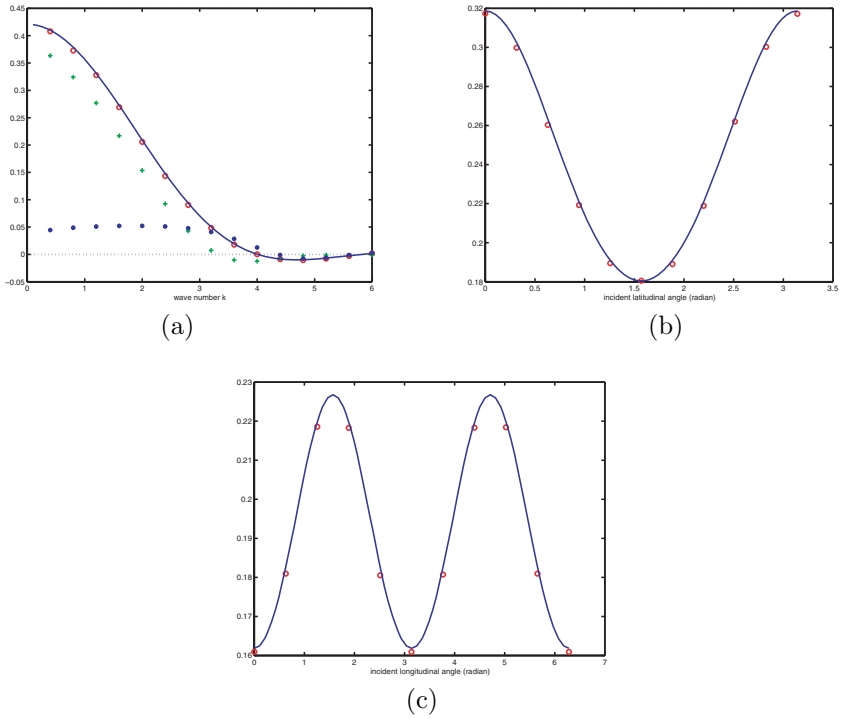
$$e_2 = \frac{\left(\sum_{i,j,k} |q_{ijk} - \bar{q}_{ijk}|^2\right)^{\frac{1}{2}}}{\left(\sum_{i,j,k} |q_{ijk}|^2\right)^{\frac{1}{2}}},$$

where $\bar{q}$ is the reconstructed scatter and q is the true scatterer.

Example 4.1. Reconstruct a scatterer defined by

$$q(x, y, z) = \begin{cases} 1 - \sqrt{\frac{x^2}{1^2} + \frac{y^2}{0.8^2} + \frac{z^2}{0.5^2}} & \text{for } \frac{x^2}{1^2} + \frac{y^2}{0.8^2} + \frac{z^2}{0.5^2} \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

The compact support of this scatterer is an ellipsoid contained in the unit ball. For simplicity, we take $\vec{n}_1 = \vec{n}_2$ and $\vec{p}_1 = \vec{p}_2$ to test the forward solver. The numerical



(a) (b) (c)

FIG. 2. (a) *Integrals at different wavenumbers for the fixed incident angle* $\theta = \frac{\pi}{3}$ *and* $\phi = \frac{\pi}{3}$. *Solid curve: the exact integral value of the left-hand side of* (3.2), $+$*: the computed integral value of the first term of the right-hand side of* (3.2), $*$*: the computed integral value of the second term of right-hand side of* (3.2), $\circ$*: the computed integral value of the right-hand side of* (3.2). (b) *Integrals with different* θ *for the fixed wavenumber* $k = 2.0$ *and* $\phi = \frac{\pi}{3}$. *Solid curve: the exact integral value of the left-hand side of* (3.2), $\circ$*: the computed integral value of the right hand-side of* (3.2). (c) *Integrals with different* ϕ *for the fixed wavenumber* $k = 2.0$ *and* $\theta = \frac{\pi}{3}$. *Solid curve: the exact integral value of the left-hand side of* (3.2), $\circ$*: the computed integral value of the right-hand side of* (3.2).

TABLE 2
Relative error at different wavenumbers.

k	1	2	3	4	5	6
e_2	0.5494	0.4876	0.3197	0.1856	0.1534	0.0895

results are shown in Figure 2. In Figure 2(a), for the fixed incident latitudinal angle $\theta = \frac{\pi}{3}$ and the longitudinal angle $\phi = \frac{\pi}{3}$, the forward problem is solved at different wavenumbers. In Figure 2(b) and 2(c), for the fixed wavenumber $k = 2$, the numerical results are shown with different latitudinal angles $\theta \in [0, \pi]$ (fix $\phi = \frac{\pi}{3}$) and $\phi \in [0, 2\pi]$ (fix $\theta = \frac{\pi}{3}$), respectively. It is easily seen from Figure 2(a) that the first term of the right-hand side of the integral equation (3.2) is dominant compared with the second (nonlinear) term when the wavenumber is small, which validates the Born approximation. Figure 3 shows the slices of the true scatterer, and Figure 4 gives the reconstruction at the wavenumber $k = 6$. The relative errors are shown in Table 2 at different wavenumbers.

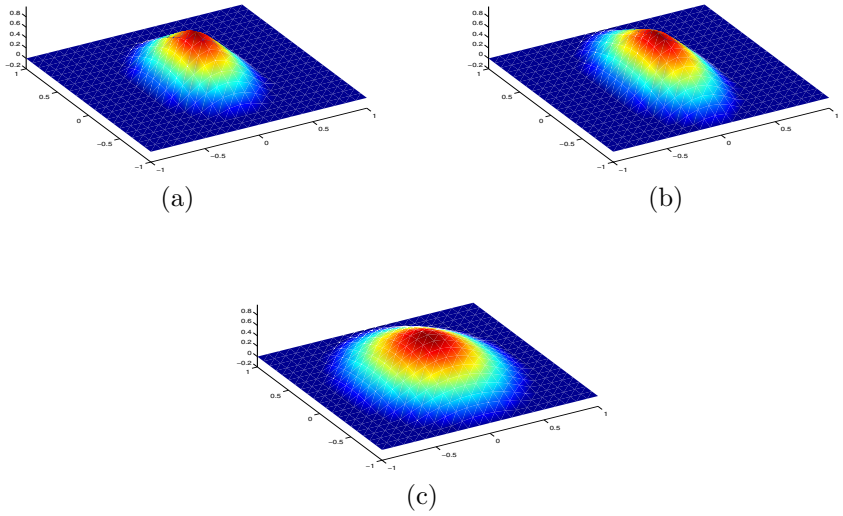


FIG. 3. *True scatterer of Example* 4.1: (a) *the slice* $x = 0$*;* (b) *the slice* $y = 0$*;* (c) *the slice* $z = 0$.

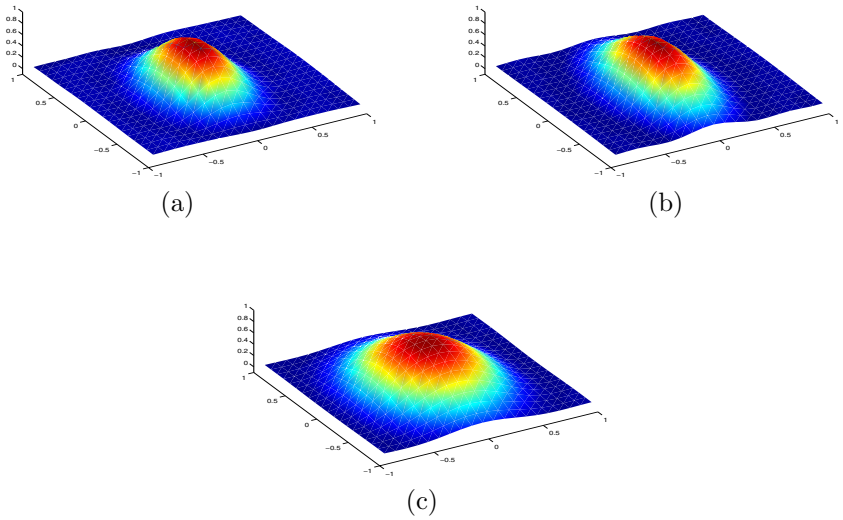
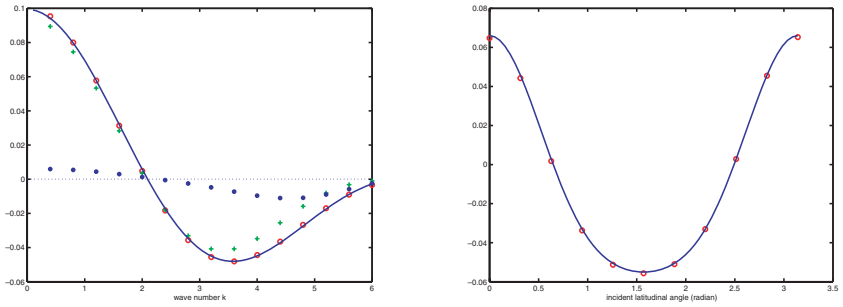


FIG. 4. *Reconstruction of Example* 4.1: (a) *the slice* $x = 0$*;* (b) *the slice* $y = 0$*;* (c) *the slice* $z = 0$.

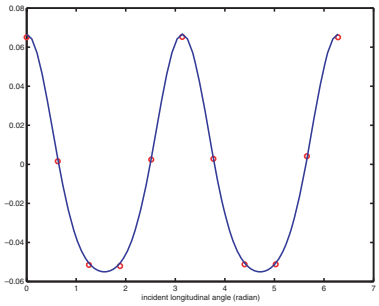
Example 4.2. Reconstruct a scatterer defined by

$$q(x,y,z)=\begin{cases}\sin(\frac{4\pi}{25})-\sin\big((x^2+(y+0.5)^2+z^2)\pi\big) & \text{for } x^2+(y+0.5)^2+z^2\le 0.4^2,\\ \sin(\frac{4\pi}{25})-\sin\big((x^2+(y-0.5)^2+z^2)\pi\big) & \text{for } x^2+(y-0.5)^2+z^2\le 0.4^2,\\ 0, & \text{otherwise.}\end{cases}$$

The compact support of this scatterer is two isolated balls with the same radius of 0.4 and the centers at $(0,-0.5,0)$ and $(0,0.5,0)$. For simplicity, we take $\vec{n}_1 = \vec{n}_2$



(a) (b)



(c)

FIG. 5. (a) *Integrals with different wavenumbers for the fixed incident angle* $\theta = \frac{\pi}{3}$ *and* $\phi = \frac{\pi}{3}$. *Solid curve: the exact integral value of the left-hand side of* (3.2), $+$*: the computed integral value of the first term of the right-hand side of* (3.2), $*$*: the computed integral value of the second term of right-hand side of* (3.2), $\circ$*: the computed integral value of the right-hand side of* (3.2). (b) *Integrals with different* θ *for the fixed wavenumber* $k = 3.0$ *and* $\phi = \frac{\pi}{3}$. *Solid curve: the exact integral value of the left-hand side of* (3.2), $\circ$*: the computed integral value of the right hand-side of* (3.2). (c) *Integrals with different* ϕ *for the fixed wavenumber* $k = 3.0$ *and* $\theta = \frac{\pi}{3}$. *Solid curve: the exact integral value of the left-hand side of* (3.2), $\circ$*: the computed integral value of the right-hand side of* (3.2).

TABLE 3
Relative error at different wavenumbers.

k	1	2	3	4	5	6	7
e_2	0.6963	0.6479	0.5891	0.4951	0.3376	0.2568	0.2221

and $\vec{p}_1 = \vec{p}_2$ in the test of the forward solver. The numerical results are given in Figure 5. In Figure 5(a), for the fixed incident latitudinal angle $\theta = \frac{\pi}{3}$ and the longitudinal angle $\phi = \frac{\pi}{3}$, the forward problem is solved at different wavenumbers. In Figure 5(b) and 5(c), for the fixed wavenumber $k = 3$, the numerical results are shown with different latitudinal angles $\theta \in [0, \pi]$ (fix $\phi = \frac{\pi}{3}$) and $\phi \in [0, 2\pi]$ (fix $\theta = \frac{\pi}{3}$), respectively. It is easily seen from Figure 5(a) that the first term of the right-hand side of the integral equation (3.2) is dominant compared with the second (nonlinear) term when the wavenumber k is small, which once again validates the Born approximation. Figure 6 shows the slices of the true scatterer and Figure 7 gives the reconstruction at the wavenumber $k = 7$. The relative errors are shown in Table 3 at different wavenumbers.

5. Concluding remarks. The proposed recursive linearization algorithm is stable and efficient for solving the inverse medium scattering problem with multiple

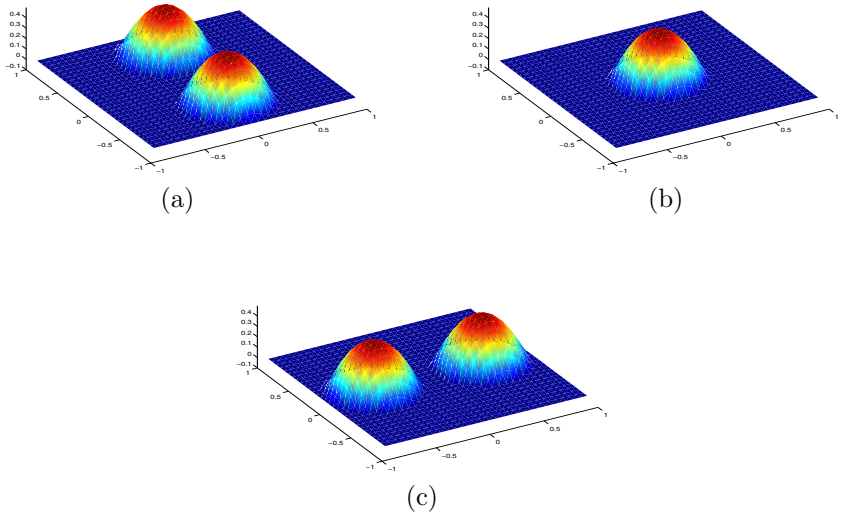


FIG. 6. *True scatterer of Example* 4.2: (a) *the slice* $x = 0$*;* (b) *the slice* $y = -0.5$*;* (c) *the slice* $z = 0$.

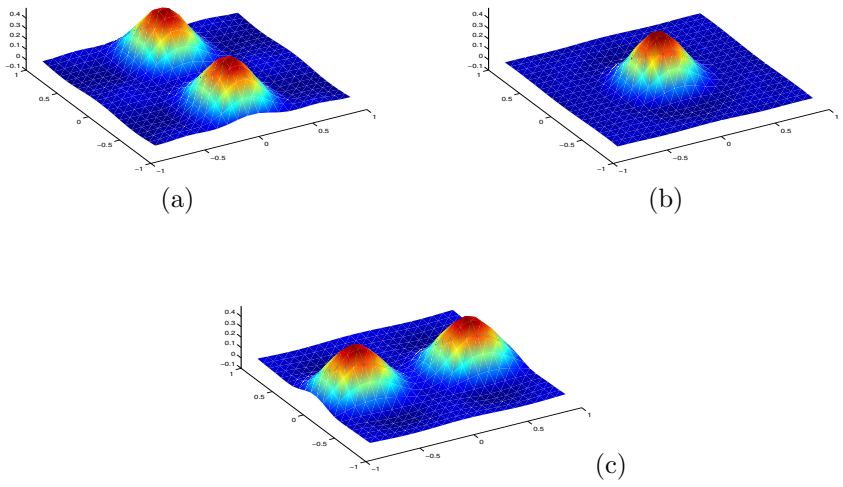


FIG. 7. *Reconstruction of Example* 4.2: (a) *the slice* $x = 0$*;* (b) *the slice* $y = -0.5$*;* (c) *the slice* $z = 0$.

frequency scattering data in three dimensions. Theoretically, scattering data with even higher wavenumbers could be used to recover more complicated scatterers which contain higher-frequency features, i.e., more Fourier modes. However, the difficulty lies in the fact that the forward model becomes difficult to solve due to the highly oscillatory nature of the solution. For a larger k, the mesh size has to be smaller, which makes numerical solution more expensive. Finally, we point out two important future directions of this research. The first concerns the convergence analysis of the

recursive linearization algorithm, which is currently in progress and will be reported elsewhere. Another challenging project is to develop an efficient algorithm for the inverse medium scattering with fixed frequency scattering data.

REFERENCES

[1] H. Ammari and G. Bao, *Analysis of the scattering map of a linearized inverse medium problem for electromagnetic waves*, Inverse Problems, 17 (2001), pp. 219–234.

[2] G. Bao and J. Liu, *Numerical solution of inverse scattering problems with multi-experimental limited aperture data*, SIAM J. Sci. Comput., 25 (2003), pp. 1102–1117.

[3] G. Bao and P. Li, *Inverse medium scattering for three-dimensional time harmonic Maxwell equations*, Inverse Problems, 20 (2004), pp. L1–L7.

[4] Y. Chen, *Inverse scattering via Heisenberg uncertainty principle*, Inverse Problems, 13 (1997), pp. 253–282.

[5] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Appl. Math. Sci. 93, Springer-Verlag, Berlin, 1998.

[6] O. Dorn, H. Bertete-Aguirre, J. G. Berryman, and G. C. Papanicolaou, *A nonlinear inversion method for 3D electromagnetic imaging using adjoint fields*, Inverse Problems, 15 (1999), pp. 1523–1558.

[7] H. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publisher, Dordrecht, The Netherlands, 1996.

[8] M. Eller, V. Isakov, G. Nakamura, and D. Tataru, *Uniqueness and stability in the Cauchy problem for Maxwell's and elasticity system*, in Proceedings of Nonlinear Partial Differential Equations and Their Applications 14 (College de France Seminar), D. Cioranescu and J. L. Lions, eds., Stud. Math. Appl., 31, North–Holland, Amsterdam, 2002, pp. 329–350.

[9] N. E. Gibbs, W. G. Poole, Jr., and P. K. Stockmeyer, *An algorithm for reducing the bandwidth and profile of a sparse matrix*, SIAM J. Numer. Anal., 13 (1976), pp. 236–250.

[10] T. Hohage, *On the numerical solution of a three-dimensional inverse medium scattering problem*, Inverse Problems, 17 (2001), pp. 1743–1763.

[11] H. Haddar and P. Monk, *The linear sampling method for solving the electromagnetic inverse medium problem*, Inverse Problems, 18 (2002), pp. 891–906.

[12] J. Jin, *The Finite Element Methods in Electromagnetics*, John Wiley & Sons, New York, 2002.

[13] A. Kameari, *Symmetric second order edge elements for triangle and tetrahedra*, IEEE Trans. Magn., 35 (1999) pp. 1394–1397.

[14] A. Kirsch and P. Monk, *A finite element/spectral method for approximating the time-harmonic Maxwell system in* $\mathbb{R}^3$, SIAM J. Appl. Math., 55 (1995), pp. 1324–1344.

[15] A. Kirsch and P. Monk, *Corrigendum: A finite element/spectral method for approximating the time-harmonic Maxwell system in* $\mathbb{R}^3$, SIAM J. Appl. Math., 58 (1998), pp. 2024–2028.

[16] A. Kirsch and P. Monk, *A finite element method for approximating electro-magnetic scattering from a conducting object*, Numer. Math., 92 (2002), pp. 501–534.

[17] P. Monk, *Finite Element Methods for Maxwell's Equations*, Oxford University Press, Oxford, U.K., 2003.

[18] F. Natterer, *The Mathematics of Computerized Tomography*, Teubner, Stuttgart, 1986.

[19] F. Natterer and F. Wübbeling, *A propagation-backpropagation method for ultrasound tomography*, Inverse Problems, 11 (1995), pp. 1225–1232.

[20] J. C. Nédélec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.

[21] J. C. Nédélec, *Acoustic and Electromagnetic Equations: Integral Representations for Harmonic Problems*, Springer, New York, 2000.

[22] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.

[23] M. Vögeler, *Reconstruction of the three-dimensional refractive index in electromagnetic scattering by using a propagation-backpropagation method*, Inverse Problems, 19 (2003), pp. 739–753.

SIAM J. APPL. MATH.
Vol. 65, No. 6, pp. 2067–2092

STIMULUS-LOCKED TRAVELING WAVES AND BREATHERS IN AN EXCITATORY NEURAL NETWORK*

STEFANOS E. FOLIAS† AND PAUL C. BRESSLOFF*

Abstract. We analyze the existence and stability of stimulus-locked traveling waves in a one-dimensional synaptically coupled excitatory neural network. The network is modeled in terms of a nonlocal integro-differential equation, in which the integral kernel represents the spatial distribution of synaptic weights, and the output firing rate of a neuron is taken to be a Heaviside function of activity. Given an inhomogeneous moving input of amplitude I_0 and velocity v, we derive conditions for the existence of stimulus-locked waves by working in the moving frame of the input. We use this to construct existence tongues in (v, I_0)-parameter space whose tips at $I_0 = 0$ correspond to the intrinsic waves of the homogeneous network. We then determine the linear stability of stimulus-locked waves within the tongues by constructing the associated Evans function and numerically calculating its zeros as a function of network parameters. We show that, as the input amplitude is reduced, a stimulus-locked wave within the tongue of an unstable intrinsic wave can undergo a Hopf bifurcation, leading to the emergence of either a traveling breather or a traveling pulse emitter.

Key words. traveling waves, traveling breathers, inhomogeneous neural media

AMS subject classification. 92C20

DOI. 10.1137/040615171

1. Introduction.

Understanding the conditions under which traveling waves of activity can propagate in cortical neural tissue is becoming an increasingly active area of research. Experimentally, these waves can be induced by a brief electrical stimulation of a disinhibited in vitro cortical slice [7, 14, 39, 29, 30]. The underlying mechanism for the propagation of such waves appears to be synaptic in origin rather than diffusive, with action potentials traveling along the axons of individual neurons. Axonal waves are modeled in terms of reaction diffusion equations based on either the four-variable Hodgkin–Huxley equations [20] or the reduced two-variable FitzHugh–Nagumo equations [12]. On the other hand, synaptic waves are typically modeled in terms of nonlocal integro-differential equations of the form [27]

$$\begin{aligned} \tau \frac{\partial u(x,t)}{\partial t} &= -u(x,t) + \int_{-\infty}^{\infty} w(x|x') f(u(x',t)) dx' - \beta q(x,t) + I(x,t), \\ \frac{1}{\epsilon} \frac{\partial q(x,t)}{\partial t} &= -q(x,t) + u(x,t), \end{aligned} \tag{1.1}$$

where τ is a membrane or synaptic time constant, $u(x,t)$ is a neural field that represents the local activity of a population of excitatory neurons at position $x \in \mathbb{R}$, $I(x,t)$ is an external input current, $f(u)$ denotes the output firing rate function, and $w(x|x')$ is the strength of connections from neurons at x' to neurons at x. The neural field $q(x,t)$ represents some form of local negative feedback mechanism such as spike frequency adaptation or synaptic depression, with β, ε determining the relative strength and rate of feedback. This form of inhibitory feedback is distinct from nonlocal synaptic inhibition, which tends to favor the formation of stationary bumps of

*Received by the editors September 15, 2004; accepted for publication (in revised form) February 24, 2005; published electronically August 9, 2005.
http://www.siam.org/journals/siap/65-6/61517.html
†Department of Mathematics, University of Utah, 155 S 1400 E, Salt Lake City, UT 84112 (sfolias@math.utah.edu, bressloff@math.utah.edu).

activity rather than traveling waves [38, 1, 28]. The nonlinear function f is typically taken to be a sigmoid function $f(u) = 1/(1 + \mathrm{e}^{-\gamma(u-\kappa)})$ with gain γ and threshold κ. Since there is strong vertical coupling between cortical layers, it is possible to treat a thin vertical cortical slice as an effective one-dimensional medium. Analysis of the model provides valuable information regarding how the speed of a traveling wave, which is relatively straightforward to measure experimentally, depends on various features of the underlying neural tissue [27]. Indeed, one prediction of the model, concerning how the speed of the wave depends on the firing threshold of the neurons, has recently been confirmed experimentally in disinhibited rat cortical slices [32]. External electric fields are used to modulate the threshold and thus control wave propagation.

One of the common assumptions in the analysis of traveling wave solutions of (1.1) is that the system is spatially homogeneous, that is, that the external input $I(x,t)$ is independent of both x and t and the synaptic weights depend only on the distance between presynaptic and postsynaptic cells, $w(x|x') = w(x - x')$. The existence of traveling waves can then be established for a class of positive, bounded weight distributions $w(x)$ that includes the exponential function $(2d)^{-1}\mathrm{e}^{-x/d}$, where d determines the range of synaptic coupling. For appropriate choices of network parameters, one finds that a single right- or left-moving traveling front exists in the absence of any feedback [4, 9, 21], whereas a pair of right- or left-moving traveling pulses exists when there is significant feedback [27]; numerically it is found that the faster pulse is stable, whereas the slower pulse is unstable. Following the original work of Amari [1], exact traveling wave solutions can be constructed by taking the high gain limit $\gamma \to \infty$, for which $f(u) = H(u - \kappa)$, where H is the Heaviside step function; that is, $H(u) = 1$ if $u \geq 0$ and $H(u) = 0$ if $u < 0$. The stability of traveling wave solutions of (1.1) in the case of a Heaviside firing rate function has recently been analyzed by Zhang [42, 43] using an Evans function approach. This is a technique for analyzing wave stability in unbounded domains that was originally developed within the context of reaction diffusion equations describing the axonal propagation of action potentials [10, 11, 22]. The basic idea is to linearize the full nonlinear equations about the traveling wave solution and to construct a complex analytic function known as the Evans function, whose zeros correspond to the point spectrum of the associated linear operator. Having established that the essential spectrum lies in the left-half complex plane, the wave is linearly stable if no eigenvalues have a positive real part and the zero eigenvalue is simple; the existence of the latter reflects the translation invariance of the system. Evans functions have now been applied to a variety of dissipative and Hamiltonian PDE systems [35], as well as a number of nonlocal integrodifferential equations [42, 43, 23, 34]. In the case of traveling wave solutions of (1.1), Zhang [42] derived an analytical expression for the Evans function using a variation of the parameters method to solve the inhomogeneous ordinary differential equation arising from linearization about the traveling wave solution. In the scalar case (zero feedback), the eigenvalues can be calculated explicitly and the associated front shown to be stable. On the other hand, for the full vector equation (1.1), it has been possible to prove stability of the fast traveling pulse only in the singular limit of slow feedback (small ε). However, one can still numerically evaluate the zeros of the Evans function outside this regime. This has been implemented by Coombes and Owen [8], who have extended the Evans function approach of Zhang [42] to a more general class of network models that incorporates discrete axonal delays and dendritic processing.

We have recently been interested in the effects of stationary inhomogeneous inputs on wave propagation and its failure in excitatory networks described by (1.1). As one

might expect intuitively, a sufficiently large variation in input blocks wave propagation (in one dimension) by spatially pinning the activity of the network. In particular, a step input or ramp results in a stationary front, whereas a local Gaussian input induces a stationary pulse. We have analyzed the stability of these stationary solutions for a Heaviside firing rate function, and shown how reducing the amplitude of the input can induce a Hopf bifurcation leading to the formation of a stable, spatially localized oscillatory solution, or *breather* [5, 13]. In the case of fronts, we have further shown that there is a critical level of negative feedback at which the homogeneous system undergoes a symmetry-breaking front bifurcation, whereby a stationary front loses stability and bifurcates into a pair of stable counterpropagating fronts. The front bifurcation acts as an organizing center for the formation of a breather in the presence of a weak input inhomogeneity [13]. Analogous results have been found for fronts [36, 18, 19, 2, 31] and pulses [33] in reaction diffusion systems. One of the potential difficulties in experimentally testing our predictions regarding input-induced coherent oscillations in cortical slices is that persistent currents tend to destroy the neurons. Although it might be possible to circumvent this problem using other forms of stimulation such as external electric fields [32], an alternative strategy is to consider the effects of moving stimuli. This is also more realistic from the perspective of the intact cortex, which is constantly being bombarded by nonstationary sensory inputs.

In this paper we extend the Evans function approach of Zhang [42] and our own previous work on stationary inhomogeneous inputs, in order to analyze the existence and stability of traveling waves locked to a moving input of constant speed v. In order to construct exact traveling wave solutions, we follow previous treatments [1, 27, 42] by considering a Heaviside firing rate function and a homogeneous weight distribution, for which (1.1) becomes

$$\tau \frac{\partial u(x,t)}{\partial t} = -u(x,t) - \beta q(x,t) + \int_{-\infty}^{\infty} w(x-x')H(u(x',t)-\kappa)dx' + I(x-vt),$$

$$\frac{1}{\epsilon}\frac{\partial q(x,t)}{\partial t} = -q(x,t) + u(x,t). \tag{1.2}$$

We assume throughout that $w(x)$ is a positive symmetric function that is monotonically decreasing on $[0,\infty)$ and satisfies the normalization condition $\int_{-\infty}^{\infty} w(x)dx < \infty$. The input is written as $I(x-vt) = I_0\,\chi(x-vt)$, with χ a fixed spatial profile that is either a bounded monotonically decreasing function in the case of fronts, or a unimodal Gaussian-like function in the case of pulses. The input amplitude I_0 and velocity v are treated as bifurcation parameters. Working in the moving frame of the input, we derive threshold-crossing conditions for the existence of a stimulus-locked wave, and use this to construct existence tongues in (v, I_0)-parameter space whose tips at $I_0 = 0$ correspond to the intrinsic waves of the homogeneous network, assuming that the latter exist. In the particular case of an exponential weight distribution, we show that there are two tongues in the positive v domain, corresponding to an unstable/stable pair of right-moving intrinsic waves. We determine the stability of the waves within these existence tongues by first constructing the Evans function for a general weight distribution w satisfying the properties listed below (1.2) and then numerically calculating the zeros of the Evans function for the exponential weight distribution. We show that as the input is reduced, a stimulus-locked wave within the tongue of the unstable intrinsic wave can undergo a Hopf bifurcation leading to the emergence of a traveling oscillatory wave. The latter takes the form of a breather or a pulse emitter in the moving frame of the stimulus. In the limit $v \to 0$ our results reduce to those previously obtained for stationary inputs [6, 13].

Note that analogous wave instabilities have been found in a scalar network with asymmetric lateral inhibition [40]. Such a network consists of a Mexican hat weight function $w_\circ$ that models short-range excitation and long-range inhibition, which is shifted asymmetrically from the center such that $w(x|x') = w_\circ(x - x' - s)$ for some fixed displacement s. This displacement introduces a form of directional selectivity, in which the network responds preferentially to stimuli moving in a particular direction, and has thus been suggested as a possible recurrent mechanism for the directional selectivity of neurons in visual cortex [37, 25]. Xie and Giese [40] have analyzed the existence and stability of stimulus-locked pulses in an asymmetric lateral inhibition network. They effectively construct the associated Evans function, although they do not identify it as such, and show how the pulse can destabilize when the stimulus velocity differs significantly from the natural velocity of unidirectional intrinsic waves; this instability generates a transition to a so-called lurching wave. Yet another neural system in which a traveling pulse can undergo a Hopf bifurcation leading to the formation of lurching waves is a synaptically coupled integrate-and-fire network with discrete axonal delays [15, 16]. Here a pulse consists of a single propagating spike, and the instability is due to fluctuations in the sequence of neuronal firing times, which start to grow at a critical value of the delay [3]. This example applies to intrinsic waves in a homogeneous network.

The structure of the paper is as follows. In order to illustrate the general approach, we begin by considering the simpler case of zero negative feedback ($\beta = 0$), for which (1.1) reduces to a scalar equation in u (section 2). The corresponding existence tongues for stimulus-locked fronts and their stability can be completely determined analytically. We next consider the existence of stimulus-locked pulses in the full vector system (1.1), numerically solving a set of nonlinear functional equations in order to construct the associated tongues (section 3). We then develop the linear stability analysis of stimulus-locked pulses in order to determine the stability of solutions within the tongues (section 4). Finally, we present numerical simulations illustrating the formation of traveling breathers and pulse emitters. Although we focus on traveling pulses rather than fronts in the case of the full system (1.1), it is straightforward to carry over our results to the case of stimulus-locked fronts, as briefly reported elsewhere [6]. Throughout the paper we work with dimensionless units. The fundamental time scale is taken to be the membrane time constant τ, which is assumed to be of the order 10 msec. The fundamental length scale is taken to be in the range d of synaptic coupling, which can vary from a few hundred micrometers to a few millimeters.

2. Stimulus-locked traveling fronts in a scalar equation. In this section we carry out a complete analysis of the existence and stability of stimulus-locked fronts in a scalar version of (1.2). As an illustrative example, we construct tongue diagrams for an exponential weight distribution, showing how the existence regions of fronts in the (v, I_0)-plane deform as the threshold κ is varied. We also establish that the fronts within the existence tongues are always stable.

2.1. Existence of stimulus-locked fronts. Consider

$$\frac{\partial u(x,t)}{\partial t} = -u(x,t) + \int_{-\infty}^{\infty} w(x-y)H(u(y,t)-\kappa)dy + I(x-vt), \tag{2.1}$$

where the input is taken to be a positive bounded monotonic function. We seek traveling front solutions of the form $u(x,t) = U(\xi)$, where $\xi = x - vt$ and

$$U(\xi) > \kappa, \quad \xi < \xi_0; \qquad U(\xi_0) = \kappa; \qquad U(\xi) < \kappa, \quad \xi > \xi_0,$$

for some $\xi_0 \in \mathbb{R}$. The wave of excitation is assumed to travel at the same velocity as the input, though the relative positions of the active region (above threshold) and the input may vary with respect to the velocity and the input strength. Thus, the active region is locked to the input but may precede or succeed the input in position. We take $U \in \mathcal{C}^1(\mathbb{R},\mathbb{R})$, where $\mathcal{C}^n(\mathbb{R},\mathbb{R})$ denotes the set of all n-times continuously differentiable functions $f : \mathbb{R} \longrightarrow \mathbb{R}$ that are bounded with respect to the sup norm. If $I_0 = 0$, then the system is translationally invariant and ξ_0 becomes a free parameter. In this case we refer to traveling waves as intrinsic or *natural* waves. The profile of the front is determined according to

$$-v\frac{dU(\xi)}{d\xi} = -U(\xi) + \int_{-\infty}^{\xi_0} w(\xi-\eta)d\eta + I(\xi). \tag{2.2}$$

Setting

$$W(\xi) = \int_{-\infty}^{\xi} w(\eta)d\eta,$$

we can integrate (2.2) over $[\xi,\infty)$ for $v > 0$ to obtain

$$U(\xi) = \frac{1}{v}\int_{\xi}^{\infty} e^{(\xi-\eta)/v} N_e(\eta;\xi_0)d\eta,$$

where

$$N_e(\xi;\xi_0) = 1 - W(\xi-\xi_0) + I(\xi).$$

We are assuming that w is normalized such that $\int_{-\infty}^{\infty} w(\eta)d\eta = 1$. Similarly, for $v < 0$ we integrate over $(-\infty,\xi]$ to find

$$U(\xi) = -\frac{1}{v}\int_{-\infty}^{\xi} e^{(\xi-\eta)/v} N_e(\eta;\xi_0)d\eta.$$

The threshold condition for the existence of a stimulus-locked front is $\kappa = U(\xi_0)$.

As a specific example, we consider a Heaviside input $I(\zeta) = I_0 H(-\zeta)$ and an exponential weight function

$$w(x) = \frac{1}{2d}e^{-|x|/d}, \tag{2.3}$$

with the length scale fixed by setting $d = 1$. The resulting threshold condition is

$$\kappa = \begin{cases} \dfrac{1}{2(1+v)} + \begin{cases} 0, & \xi_0 \geq 0, \\ I_0(1-e^{\xi_0/v}), & \xi_0 < 0, \end{cases} & v > 0, \\ \dfrac{1+2|v|}{2(1+|v|)} + \begin{cases} I_0 e^{\xi_0/v}, & \xi_0 > 0, \\ I_0, & \xi_0 \leq 0, \end{cases} & v < 0. \end{cases} \tag{2.4}$$

In the absence of an input ($I_0 = 0$), the threshold condition reduces to

$$\kappa = \begin{cases} \dfrac{1}{2(1+v_\circ)}, & v \geq 0, \\ \dfrac{1+2|v_\circ|}{2(1+|v_\circ|)}, & v < 0, \end{cases}$$

where $v_\circ$ is the natural speed of the wave. Solving for $v_\circ$ in terms of κ, we find that $v_\circ$ is a sigmoidal function of κ:

$$v_\circ(\kappa) = \begin{cases} \dfrac{\frac{1}{2}-\kappa}{\kappa}, & 0 < \kappa \leq \frac{1}{2}, \\ \dfrac{\frac{1}{2}-\kappa}{(\kappa-1)}, & \frac{1}{2} < \kappa < 1. \end{cases}$$

The homogeneous network supports a stationary natural front ($v_\circ = 0$) when $\kappa = \frac{1}{2}$, a front moving to the right for $0 < \kappa < \frac{1}{2}$, and front moving to the left for $\frac{1}{2} < \kappa < 1$. Moreover, $v_\circ \to \infty$ as $\kappa \to 0$ and $v_\circ \to -\infty$ as $\kappa \to 1$. It does not support a natural front when $\kappa > 1$, as any heteroclinic orbit joining the equilibrium $\{0, 1\}$ at infinity does not satisfy the threshold behavior used to define a traveling front solution. This recovers a result from [9].

We now analyze (2.4) for $I_0 > 0$ in order to determine the regions of the (v, I_0)-parameter subspace for which stimulus-locked waves exist. We first consider the case $v > 0$. For $\xi_0 \geq 0$ we have the threshold condition

$$\kappa = \frac{1}{2(1+v)},$$

and hence there are infinitely many waves parameterized by $\xi_0 \in [0, \infty)$, all of which travel with the natural speed $v = \frac{1-2\kappa}{2\kappa}$ for $0 < \kappa < \frac{1}{2}$. This degeneracy is a consequence of using the Heaviside input and would not occur if a continuous strictly monotonic input were used; however, the analysis is considerably more involved. For $\xi_0 < 0$ we have instead

$$\kappa = \frac{1}{2(1+v)} + I_0(1 - e^{\xi_0/v}).$$

As the right-hand side is monotonic in ξ_0, we can solve for ξ_0 as a function of v to obtain

$$\xi_0(v) = v \ln \left[1 - \frac{1}{I_0} \left(\kappa - \frac{1}{2(1+v)} \right) \right].$$

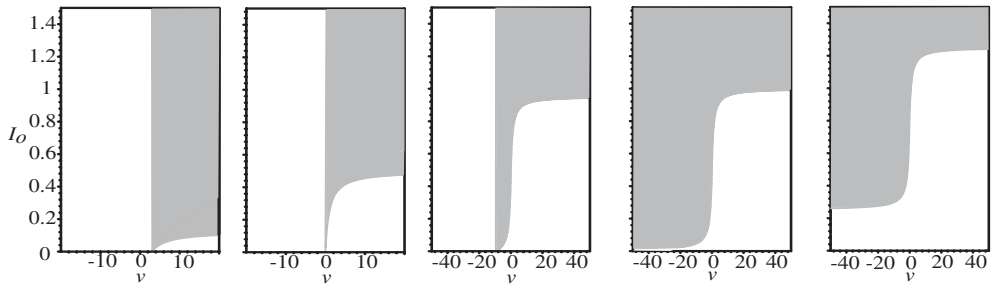
Since $\xi_0 < 0$ and $v > 0$, we see that solutions exist only if

$$0 < 1 - \frac{1}{I_0} \left(\kappa - \frac{1}{2(1+v)} \right) \leq 1$$

or, equivalently,

$$2(\kappa - I_0) \ < \ \frac{1}{1+v} \ \leq \ 2\kappa. \tag{2.5}$$

The right inequality of (2.5) implies that, if $\kappa < \frac{1}{2}$, then $v > v_\circ(\kappa)$, where $v_\circ$ is the corresponding natural velocity. Similarly, the left inequality implies that, if $I_0 < \kappa$, then $0 < v < v_1(\kappa - I_0)$, with $v_1(s) = \frac{1}{2s} - 1$. Hence, for $0 < \kappa \leq \frac{1}{2}$ we obtain the existence regions in the (v, I_0)-plane shown in Figure 2.1(a)–(b). The left boundary is given by $v = v_\circ(\kappa)$ and the right boundary by $v = v_1(\kappa - I_0)$. The two boundaries form a tongue that emerges from the natural speed $v_\circ(\kappa)$ at $I_0 = 0$.



(a) $0 < \kappa < 1/2$ (b) $\kappa = 1/2$ (c) $1/2 < \kappa < 1$ (d) $\kappa = 1$ (e) $\kappa > 1$

FIG. 2.1. *Deformation of existence regions (gray) for stimulus-locked traveling fronts as* κ *varies in the scalar equation. Particular values of* κ *are as follows:* (a) $\kappa = 0.125$, (c) $\kappa = 0.95$, (e) $\kappa = 1.25$.

Now consider $v < 0$. For $\xi_0 < 0$ we have the threshold condition

$$\kappa = \frac{1 + 2|v|}{2(1 + |v|)} + I_0,$$

which implies

$$|v| = \frac{1 - 2(\kappa - I_0)}{2(\kappa - I_0 - 1)} \equiv v_2(\kappa - I_0).$$

Again we have an infinite family of waves corresponding to a single speed. Since $|v| \geq 0$, such solutions exist only for

$$\kappa - 1 \ < \ I_0 \ < \ \kappa - \frac{1}{2}.$$

On the other hand, for $\xi_0 \geq 0$ we have the threshold condition

$$\kappa = \frac{1 + 2|v|}{2(1 + |v|)} + I_0 e^{\xi_0/v}.$$

Monotonicity of the right-hand side again allows us to solve for $\xi_0(v)$ to find

$$\xi_0(v) = v \ln \left[\frac{1}{I_0} \left(\kappa - \frac{1 + 2|v|}{2(1 + |v|)} \right) \right],$$

and, since $v < 0$ and $\xi_0 \geq 0$, it follows that waves exist only for v satisfying

$$\kappa - I_0 \leq \frac{1 + 2|v|}{2(1 + |v|)} < \kappa. \tag{2.6}$$

The right inequality of (2.6) implies that if $\frac{1}{2} < \kappa < 1$, then $v_\circ(\kappa) < v < 0$. Thus, for $\frac{1}{2} < \kappa < 1$ we obtain the existence region shown in Figure 2.1(c); the left boundary is given by $v = v_0(\kappa)$ and the right boundary by $v = v_2(\kappa - I_0)$ for $v < 0$ and $v = v_1(\kappa - I_0)$ for $v > 0$. Again there is a tongue with tip at the natural speed. For $\kappa > 1$ the left boundary disappears, and one finds stimulus-locked waves only when $I_0 > \kappa - 1$, i.e., when there no longer exist natural waves. The left inequality of (2.6) implies that if $\frac{1}{2} < \kappa - I_0 < 1$, then $v < v_2(\kappa - I_0) < 0$, whereas if $\kappa - I_0 > 1$, then no solution exists. For all $\kappa > 1$ the region of existence is identical to that for $\kappa = 1$, though it is shifted vertically by $\kappa - 1$, as shown in Figure 2.1(d)–(e).

2.2. Stability of stimulus-locked fronts. Consider the evolution of small smooth perturbations $\bar{\varphi}$ of the stimulus-locked front solution U. Linearizing (2.1) about the wave, the perturbations evolve according to

$$\frac{\partial \bar{\varphi}}{\partial t} - v\frac{\partial \bar{\varphi}}{\partial \xi} + \bar{\varphi} = \int_{\mathbb{R}} w(\xi-\eta)H'(U(\eta)-\kappa)\bar{\varphi}(\eta)d\eta. \tag{2.7}$$

Separating variables, $\bar{\varphi}(\xi,t) = \varphi(\xi)e^{\lambda t}$, we find that $\varphi \in \mathcal{C}^1(\mathbb{R},\mathbb{C})$ satisfies the eigenvalue problem

$$(\mathcal{L}+\mathcal{N}_s)\,\varphi = \lambda\,\varphi, \tag{2.8}$$

where

$$\mathcal{L}\varphi = v\frac{\partial \varphi}{\partial \xi} - \varphi, \qquad \mathcal{N}_s\varphi\,(\xi) = \frac{w(\xi-\xi_0)}{|U'(\xi_0)|}\,\varphi(\xi_0). \tag{2.9}$$

We need to characterize the spectrum of the linear operator $\mathcal{L}+\mathcal{N}_s : \mathcal{C}^1(\mathbb{R},\mathbb{C}) \longrightarrow \mathcal{C}^0(\mathbb{R},\mathbb{C})$ in order to determine the linear stability of the traveling pulse. The following definitions concern linear operators $\mathcal{T} : \mathcal{D}(\mathcal{T}) \longrightarrow \mathcal{B}$, where $\mathcal{B}$ is a Banach space and the domain $\mathcal{D}(\mathcal{T})$ of $\mathcal{T}$ is dense in $\mathcal{B}$ [41]. In our case $\mathcal{D}(\mathcal{L}+\mathcal{N}_s) = \mathcal{C}^1(\mathbb{R},\mathbb{C})$, which is dense in $\mathcal{C}^0(\mathbb{R},\mathbb{C})$. λ is in the resolvent set ρ if $\lambda \in \mathbb{C}$ is such that $\mathcal{T}-\lambda$ has a range dense in $\mathcal{B}$ and a continuous inverse $(\mathcal{T}-\lambda)^{-1}$; otherwise λ is in the spectrum σ. We decompose the spectrum into the following disjoint sets: λ is an element of the point spectrum σ_p if $\mathcal{T}-\lambda$ is not invertible; λ is an element of the continuous spectrum σ_c if $\mathcal{T}-\lambda$ has an unbounded inverse with domain dense in $\mathcal{B}$; λ is an element of the residual spectrum σ_r if $\mathcal{T}-\lambda$ has an inverse (bounded or not) whose domain is not dense in $\mathcal{B}$. We refer to elements of the point spectrum as eigenvalues and the union of the continuous and residual spectra as the essential spectrum.

Regarding the essential spectrum, we mention that $\mathcal{N}_s$ is a compact linear operator. The consequence is that, since $\mathcal{N}_s$ is compact, the operators $\mathcal{L}+\mathcal{N}_s$ and $\mathcal{L}$ have the same essential spectra [24, 23]. To see that the operator is compact, we define $\mathcal{N}_s$ by the composition $\mathcal{TS}$, where

$$\mathcal{S}\,\varphi = \varphi(\xi_0), \qquad (\mathcal{T}\,z)(\xi) = \frac{w(\xi-\xi_0)}{|U'(\xi_0)|}z.$$

Since $\mathcal{S} : \mathcal{C}^1(\mathbb{R},\mathbb{C}) \longrightarrow \mathbb{C}$ has a finite-dimensional range, it is a compact linear operator. Moreover, since $\mathcal{T} : \mathbb{C} \longrightarrow \mathcal{C}^0(\mathbb{R},\mathbb{C})$ is a bounded linear operator, it follows that the composition $\mathcal{TS}$ is a compact linear operator.

Resolvent and the point spectrum. We seek to construct a bounded inverse by solving the inhomogeneous equation

$$(\mathcal{L}+\mathcal{N}_s-\lambda)\varphi = -f, \tag{2.10}$$

where $f \in \mathcal{C}^0(\mathbb{R},\mathbb{C})$, using a variation of parameters approach along the lines of Zhang [42]. We write (2.10) as

$$\frac{\partial}{\partial \xi}\Big(e^{-\left(\frac{1+\lambda}{v}\right)\xi}\,\varphi(\xi)\Big) = -\frac{1}{v}\,e^{-\left(\frac{1+\lambda}{v}\right)\xi}\,\Big(\mathcal{N}_s\varphi(\xi)+f(\xi)\Big). \tag{2.11}$$

For $\frac{\mathcal{R}e(\lambda)+1}{v} > 0$, integrating (2.11) over $[\xi,\infty)$ yields

$$\varphi(\xi) - \Lambda_+(\lambda;\xi)\,\varphi(\xi_0) \;=\; \mathcal{H}_f(\xi), \tag{2.12}$$

where

$$\Lambda_+(\lambda;\xi) = \frac{1}{v|U'(\xi_0)|}\int_\xi^\infty w(\eta-\xi_0)e^{\left(\frac{1+\lambda}{v}\right)(\xi-\eta)}d\eta,$$

$$\mathcal{H}_f(\xi) = \frac{1}{v}\int_\xi^\infty e^{\left(\frac{1+\lambda}{v}\right)(\xi-\eta)}f(\eta)d\eta.$$

Using the Hölder inequality, it can be shown that both $\Lambda_+(\lambda;\xi)$ and $\mathcal{H}_f(\xi)$ are bounded for all $\xi \in \mathbb{R}$ and $f \in \mathcal{C}^0(\mathbb{R},\mathbb{C})$. It is then seen from (2.12) that $\varphi(\xi)$ is determined by its restriction $\varphi(\xi_0)$, in which case we obtain

$$\left(1-\Lambda_+(\lambda;\xi_0)\right)\varphi(\xi_0) = \frac{1}{v}\int_{\xi_0}^\infty e^{\left(\frac{1+\lambda}{v}\right)(\xi-\eta)}f(\eta)d\eta.$$

This can be solved for $\varphi(\xi_0)$ and hence for $\varphi(\xi)$ if and only if

$$1-\Lambda_+(\lambda;\xi_0) \neq 0.$$

This results in a bounded inverse which is defined on all of $\mathcal{C}^0(\mathbb{R},\mathbb{C})$, and therefore all corresponding λ are in the resolvent set. On the other hand, we cannot invert the operator for λ such that

$$1-\Lambda_+(\lambda;\xi_0) = 0.$$

In this case

$$(\mathcal{L}+\mathcal{N}_s-\lambda)\varphi = 0 \tag{2.13}$$

has nontrivial solutions, indicating that λ is in the point spectrum. Moreover, if we define the function

$$\mathcal{E}_+(\lambda;\xi_0) = 1-\Lambda_+(\lambda;\xi_0), \qquad \frac{\mathcal{R}e(\lambda)+1}{v} > 0,$$

we see that eigenvalues form the zero set. Similarly for $\frac{\mathcal{R}e(\lambda)+1}{v} < 0$, integrating (2.11) over $(-\infty,\xi_0]$ yields a similar condition for the existence of eigenfunctions

$$1 = \Lambda_-(\lambda,\xi_0), \qquad \frac{\mathcal{R}e(\lambda)+1}{v} < 0,$$

where

$$\Lambda_-(\lambda;\xi) = -\frac{1}{v|U'(\xi_0)|}\int_{-\infty}^\xi w(\eta-\xi_0)e^{\left(\frac{1+\lambda}{v}\right)(\xi-\eta)}d\eta. \tag{2.14}$$

The Evans function is then defined as

$$\mathcal{E}(\lambda;\xi_0) = 1-\Lambda_\pm(\lambda;\xi_0), \qquad \frac{\mathcal{R}e(\lambda)+1}{v} \gtrless 0.$$

Essential spectrum. Since $\mathcal{N}_s$ does not contribute to the essential spectrum of $\mathcal{L}+\mathcal{N}_s$, we need only calculate the essential spectrum of the linear operator $\mathcal{L}$. The essential spectrum is the set of $\lambda = -1 + iv\rho$, where $\rho \in \mathbb{R}$. Since this has negative real part, the essential spectrum does not contribute to any wave instabilities. We demonstrate that, for these values of λ, there exist bounded functions for which the inverse operator $(\mathcal{L}-\lambda)^{-1}$ becomes unbounded, indicating that λ is a member of the continuous spectrum.

Suppose that $\lambda = -1 + iv\rho$, and consider the sequence of bounded functions [43]

$$\varphi_m(\xi) = (1 - e^{-\xi^2/2m^2})e^{i\rho\xi}, \qquad m \in \mathbb{N},$$

for which

$$\|\varphi_m\|_\infty = 1 \qquad \forall \ \ m \in \mathbb{N},\ \rho \in \mathbb{R}.$$

However,

$$(\mathcal{L}-\lambda)\varphi_m(\xi) = \frac{v}{m^2}\ \xi e^{-\xi^2/2m^2}\ e^{i\rho\xi},$$

which implies that

$$\left\|(\mathcal{L}-\lambda)\varphi_m\right\|_\infty = \frac{v}{m^2}\left\|\xi e^{-\xi^2/2m^2}\right\|_\infty \longrightarrow 0 \quad \text{as} \quad m \longrightarrow \infty.$$

Hence, $(\mathcal{L}-\lambda)^{-1}$ is unbounded, and the set of $\lambda = -1 + iv\rho$, where $\rho \in \mathbb{R}$, form the essential spectrum. The residual spectrum in this case is empty, though we shall see that the vector system does, in fact, have a nonempty residual spectrum.

Evans function for an exponential weight distribution. We now explicitly calculate the zeros of the Evans functions for a Heaviside input and exponential weight distribution. The region in the complex plane $\mathbf{D} = \{z \ : \ \mathcal{R}e(z) > -1\}$ is the domain of the Evans function $\mathcal{E}_+$, and we need only consider this region to determine the stability of the wave. For $v > 0$ and $\lambda \in \mathbf{D}$,

$$\begin{aligned}\mathcal{E}_+(\lambda,\xi_0) &= 1 - \frac{1}{v|U'(\xi_0)|}\int_{\xi_0}^{\infty} w(\eta-\xi_0)e^{\left(\frac{1+\lambda}{v}\right)(\xi_0-\eta)}d\eta \\ &= 1 - \frac{1}{2(1+\lambda+v)}\frac{1}{|U'(\xi_0)|},\end{aligned}$$

and similarly for $v < 0$ and $\lambda \in \mathbf{D}$,

$$\begin{aligned}\mathcal{E}_-(\lambda,\xi_0) &= 1 + \frac{1}{v|U'(\xi_0)|}\int_{-\infty}^{\xi_0} w(\eta-\xi_0)e^{\left(\frac{1+\lambda}{v}\right)(\xi_0-\eta)}d\eta \\ &= 1 + \frac{1}{2(1+\lambda+v)}\frac{1}{|U'(\xi_0)|}.\end{aligned}$$

Note that this recovers the Evans function obtained by Zhang [42] in the case of a homogeneous input. From this we can directly solve $\mathcal{E}_\pm(\lambda;\xi_0) = 0$ for λ:

$$\lambda = -(1+|v|) + \frac{1}{2|U'(\xi_0)|}, \qquad v \in \mathbb{R}, \tag{2.15}$$

with $U'(\xi_0)$ determined from (2.2),

$$U'(\xi_0) = \frac{1}{v}\left(U(\xi_0) - \int_{-\infty}^{\xi_0} w(\xi_0 - \eta)d\eta - I(\xi_0)\right)$$
$$= \frac{1}{v}\left(\kappa - \frac{1}{2} - I(\xi_0)\right)$$

and κ satisfying the self-consistency conditions (2.4).

In the case $I_0 = 0$ the eigenvalues are given by

$$\lambda = -(1 + |v|) + \frac{|v|}{2\left|\kappa - \frac{1}{2}\right|}, \qquad v \in \mathbb{R}, \tag{2.16}$$

where v is the natural wave speed. Substituting (2.4) into (2.16), we find that the only eigenvalue in $\mathbf{D}$ is the zero eigenvalue $\lambda = 0$. Moreover, it can be shown that the eigenvalue is simple [42] and hence that the natural front is linearly stable, modulo uniform translations.

In the case of an inhomogeneous input ($I_0 > 0$), we have to deal with each of the separate subdomains of the threshold conditions (2.4). First, for $v > 0,\ \xi_0 > 0$ we notice that $I(\xi_0) = 0$ and κ is identical to the case of a natural wave; hence, $\lambda = 0$ is the only eigenvalue in $\mathbf{D}$. If $v > 0,\ \xi_0 < 0$, substituting (2.4) for κ into (2.15) yields the eigenvalue

$$\lambda = -1 - v + \frac{v}{2|\kappa - \frac{1}{2} - I_0|}$$
$$= (1 + v)\left[-1 + \frac{v}{\left|v + 2(1 + v)I_0(1 - e^{\xi_0/v})\right|}\right].$$

Since $I_0(1 - e^{\xi_0/v}) > 0$ for all $v > 0,\ \xi_0 < 0,\ I_0 > 0$, it follows that $\lambda < 0$ and the corresponding front is always stable. On the other hand, if $v < 0$ and $\xi_0 < 0$, we find $\lambda = 0$, again indicating stability with respect to the degenerate family of waves corresponding to the boundary of the tongue. For $\xi_0 > 0$ we similarly calculate

$$\lambda = (1 + |v|)\left[-1 + \frac{|v|}{\left||v| + 2(1 + |v|)I_0 e^{\xi_0/v}\right|}\right].$$

Since $2(1 + |v|)I_0 e^{\xi_0/v} > 0$ for $v < 0,\ \xi_0 > 0,\ I_0 > 0$, it again follows that $\lambda < 0$ and the corresponding front is always stable.

3. Stimulus-locked traveling pulses in the vector system. In this section we construct stimulus-locked traveling pulse solutions of (1.2) in the case of a unimodal input moving with constant velocity v. We first derive the formal solution for a general weight distribution w, and then use this to construct existence tongues in the (v, I_0)-plane for an exponential weight distribution and a Gaussian input of amplitude I_0.

3.1. Existence of stimulus-locked pulses. Consider a traveling pulse that is generated by, and locked to, an inhomogeneous input I traveling with constant speed v. Such a wave has permanent or *stationary* form; i.e., it translates as a rigid structure. Define the traveling wave coordinates (ξ, t), where $\xi = x - vt$ and v is

the velocity associated with the input. A *stimulus-locked traveling pulse* is a pair of functions (U,Q), with $U,Q \in \mathcal{C}^1(\mathbb{R},\mathbb{R})$, which in traveling wave coordinates satisfy the conditions

$$\begin{aligned} &U(\xi_i)=\kappa, \qquad i=1,2; && U(\xi)\longrightarrow 0 \qquad \text{as } \xi\longrightarrow\pm\infty;\\ &U(\xi)>\kappa, \qquad \xi_1<\xi<\xi_2; && U(\xi) < \kappa, \qquad \text{otherwise,}\end{aligned}$$

with ξ_1,ξ_2 defining the points at which the activity U crosses threshold. Taking $u(x,t)=U(x-vt)$ and $q(x,t)=Q(x-vt)$, the profile of the pulse is governed by

$$\begin{aligned} -v\,U_\xi &= -U-\beta Q+\int_{\xi_1}^{\xi_2} w(\xi-\eta)d\eta + I(\xi),\\ -\frac{v}{\epsilon}\,Q_\xi &= -Q \;+\; U.\end{aligned}$$

In general, we take the excitatory weight function $w(x)$ to be nonnegative, continuous, symmetric in x, and monotonically decreasing in $|x|$. Let $\mathbf{s}=(U,Q)^T$ and W denote an antiderivative of w; we can rewrite the system more compactly as

$$\mathcal{L}\mathbf{s} \equiv \begin{pmatrix} vU_\xi - U - \beta Q \\ vQ_\xi + \epsilon U - \epsilon Q\end{pmatrix} = -\begin{pmatrix} N_e \\ 0\end{pmatrix}, \tag{3.1}$$

where

$$N_e(\xi) = W(\xi-\xi_1) - W(\xi-\xi_2) + I(\xi). \tag{3.2}$$

We use variation of parameters to solve this linear equation. The homogeneous problem $\mathcal{L}\mathbf{s}=\mathbf{0}$ has the two linearly independent solutions,

$$\mathbf{S}_+(\xi) = \begin{pmatrix} \beta \\ m_+ - 1\end{pmatrix}\exp(\mu_+\xi), \qquad \mathbf{S}_-(\xi) = \begin{pmatrix} \beta \\ m_- - 1\end{pmatrix}\exp(\mu_-\xi),$$

where

$$\mu_\pm = \frac{m_\pm}{v}, \qquad m_\pm = \frac{1}{2}\Big(1+\epsilon \pm \sqrt{(1-\epsilon)^2-4\epsilon\beta}\Big).$$

We set

$$\mathbf{s}(\xi) = \Big[\mathbf{S}_+\big|\mathbf{S}_-\Big]\begin{pmatrix} a(\xi)\\ b(\xi)\end{pmatrix},$$

where $a,b\in\mathcal{C}^1(\mathbb{R},\mathbb{R})$ and $[A|B]$ denotes the matrix whose first column is defined by the vector A and whose second column is defined by the vector B. Since $\mathcal{L}\mathbf{S}_\pm = 0$, (3.1) becomes

$$\Big[\mathbf{S}_+\big|\mathbf{S}_-\Big]\,\frac{\partial}{\partial\xi}\begin{pmatrix} a(\xi)\\ b(\xi)\end{pmatrix} = -\frac{1}{v}\begin{pmatrix} N_e(\xi)\\ 0\end{pmatrix}. \tag{3.3}$$

Since $[\mathbf{S}_+|\mathbf{S}_-]$ is invertible, we find

$$\frac{\partial}{\partial\xi}\begin{pmatrix} a(\xi)\\ b(\xi)\end{pmatrix} = -\frac{1}{v\beta(m_+-m_-)}\Big[\mathbf{Z}_+\big|\mathbf{Z}_-\Big]^T\begin{pmatrix} N_e(\xi)\\ 0\end{pmatrix},$$

where

$$\mathbf{Z}_+(\xi) = \begin{pmatrix} 1-m_- \\ \beta \end{pmatrix} \exp(-\mu_+\xi), \qquad \mathbf{Z}_-(\xi) = -\begin{pmatrix} 1-m_+ \\ \beta \end{pmatrix} \exp(-\mu_-\xi).$$

For $v > 0$, we integrate over $[\xi, \infty)$ to obtain

$$\begin{pmatrix} a(\xi) \\ b(\xi) \end{pmatrix} = \begin{pmatrix} a_\infty \\ b_\infty \end{pmatrix} + \frac{1}{v\beta(m_+ - m_-)} \int_\xi^\infty \Big[\mathbf{Z}_+ \big| \mathbf{Z}_-\Big]^T \begin{pmatrix} N_e(\eta) \\ 0 \end{pmatrix} d\eta,$$

where a_∞, b_∞ denote the values of $a(\xi), b(\xi)$ as $\xi \longrightarrow \infty$. Thus

(3.4)
$$\mathbf{s}(\xi) = \Big[\mathbf{S}_+ \big| \mathbf{S}_-\Big] \begin{pmatrix} a_\infty \\ b_\infty \end{pmatrix} + \frac{1}{v\beta(m_+ - m_-)} \Big[\mathbf{S}_+ \big| \mathbf{S}_-\Big] \int_\xi^\infty \Big[\mathbf{Z}_+ \big| \mathbf{Z}_-\Big]^T \begin{pmatrix} N_e(\eta) \\ 0 \end{pmatrix} d\eta.$$

Using the Hölder inequality and that $N_e \in \mathcal{C}^0(\mathbb{R}, \mathbb{R})$, it is straightforward to show that the integral term in (3.4) is bounded for all $\xi \in \mathbb{R}$; hence, a bounded solution $\mathbf{s}$ exists only if $a_\infty = b_\infty = 0$. The general stimulus-locked pulse is given by

$$\mathbf{s}(\xi) = \frac{1}{v\beta(m_+ - m_-)} \Big[\mathbf{S}_+ \big| \mathbf{S}_-\Big] \int_\xi^\infty \Big[\mathbf{Z}_+ \big| \mathbf{Z}_-\Big]^T \begin{pmatrix} N_e(\eta) \\ 0 \end{pmatrix} d\eta.$$

Furthermore, if we define the functions

$$\mathcal{M}_\pm(\xi) = \frac{1}{v(m_+ - m_-)} \int_\xi^\infty e^{\mu_\pm(\xi-\eta)} N_e(\eta) d\eta,$$

we can express the solution (U, Q) as follows:

$$U(\xi) = (1-m_-)\mathcal{M}_+(\xi) - (1-m_+)\mathcal{M}_-(\xi), \tag{3.5}$$

$$Q(\xi) = \beta^{-1}(m_+ - 1)(1-m_-)\big[\mathcal{M}_+(\xi) - \mathcal{M}_-(\xi)\big]. \tag{3.6}$$

Since $N_e(\xi)$ is dependent upon ξ_1, ξ_2, the threshold conditions $U(\xi_i) = \kappa$, where $i = 1, 2$ and $\xi_1 < \xi_2$, determine the relationship between the input strength I_0 and the position of the pulse relative to the input I. This provides the following consistency conditions for the existence of a stimulus-locked traveling pulse, which, we note, reduce to the case of natural waves for $I_0 = 0$:

$$\kappa = (1-m_-)\mathcal{M}_+(\xi_1) - (1-m_+)\mathcal{M}_-(\xi_1), \tag{3.7}$$

$$\kappa = (1-m_-)\mathcal{M}_+(\xi_2) - (1-m_+)\mathcal{M}_-(\xi_2). \tag{3.8}$$

3.2. Pulses for an exponential weight distribution. Consider, in particular, an exponential weight distribution given by (2.3) with $d = 1$ and a Gaussian input

$$I(x) = I_0\, e^{-(x/\sigma)^2}. \tag{3.9}$$

Existence conditions determined from (3.7) and (3.8) yield the following system of nonlinear equations that determines the relationship between the input parameters (v, I_0) and the threshold points (ξ_1, ξ_2):

$$\kappa = K(\xi_1 - \xi_2) + T_+(\xi_1) - T_-(\xi_1), \tag{3.10}$$

$$\kappa = J(\xi_1 - \xi_2) + T_+(\xi_2) - T_-(\xi_2), \tag{3.11}$$

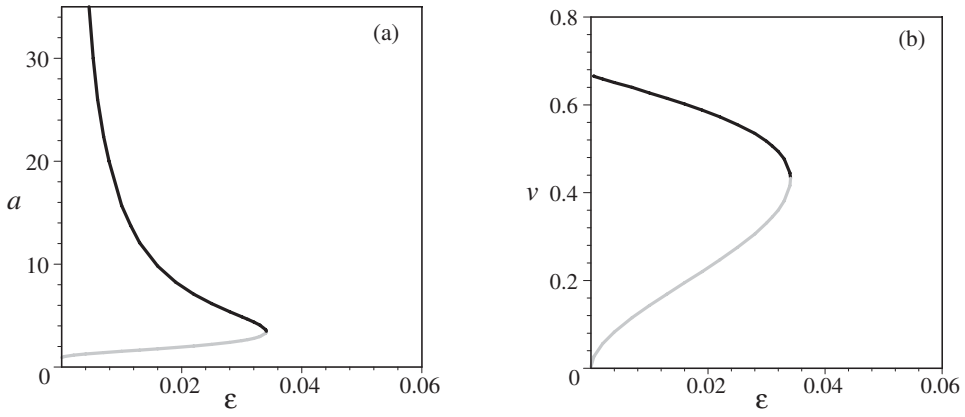


FIG. 3.1. *Bifurcation curves for the existence of natural traveling pulses ($I_0 = 0$) for the vector system* (1.2) *in* (a) *the (ϵ, a)-plane and* (b) *the (ϵ, v)-plane, illustrating that natural waves exist only for small ϵ. Here $a = \xi_2 - \xi_1$ denotes the width of a pulse. The stable branch (black), characterized by wide (large a), fast pulses, and the unstable branch (gray), characterized by narrow, slow pulses, annihilate in a saddle-node bifurcation at a critical value ϵ_c. In this case $\kappa = 0.3$, $\beta = 2.5$, and $\epsilon_c \approx 0.341$.*

where

$$K(\zeta) = K_0 + K_1 e^{\zeta} - K_+ e^{\mu_+ \zeta} + K_- e^{\mu_- \zeta}, \qquad J(\zeta) = \frac{v+\epsilon}{2(v+m_+)(v+m_-)}\Big(1 - e^{\zeta}\Big),$$

$$K_1 = \frac{1}{2}\frac{v-\epsilon}{(v-m_+)(v-m_-)}, \qquad K_\pm = \frac{v^2(1-m_\mp)}{m_\pm(v^2 - {m_\pm}^2)(m_+ - m_-)},$$

$$K_0 = \left(\frac{(1-m_-)(2v+m_+)}{2m_+(v+m_+)(m_+-m_-)}\right) - \left(\frac{(1-m_+)(2v+m_-)}{2m_-(v+m_-)(m_+-m_-)}\right),$$

$$T_\pm(\zeta) = \frac{\sqrt{\pi}\ \sigma I_0}{2\ v}\ \left(\frac{1-m_\mp}{m_+ - m_-}\right) \exp\left(\frac{(\mu_\pm \sigma)^2}{4} + \mu_\pm \zeta\right) \mathbf{erfc}\left(\frac{\zeta}{\sigma} + \frac{\mu_\pm \sigma}{2}\right),$$

and $\mathbf{erfc}(z)$ denotes the complementary Error function.

Natural traveling pulses ($I_0 = 0$). Numerically solving (3.10) and (3.11) for $I_0 = 0$, we find that for sufficiently small ϵ there exists a pair of traveling pulses arising from a saddle-node bifurcation. Numerical simulations suggest that the larger and faster pulse is stable while the smaller slower pulse is unstable and acts as a separatrix between the fast pulse and the rest state [27]. Zhang's analysis has shown the fast pulse to be stable in the singular limit $\epsilon \longrightarrow 0$ [42]. In Figure 3.1 we present bifurcation diagrams using ϵ as a bifurcation parameter to demonstrate the existence and stability of natural waves; stability is determined by numerically solving for the zero set of the Evans function, constructed in section 4.2. It is found that the larger, faster wave is stable (black), while the smaller, slower wave is unstable (gray).

Stimulus-locked traveling pulses. Numerically solving (3.10) and (3.11) for $I_0 > 0$, we can determine the regions in the (v, I_0)-plane where one or more stimulus-locked waves exist. First, performing a continuation from the pair of natural waves, we generate a corresponding pair of existence tongues with tips at $I_0 = 0$. These are illustrated in Figure 3.2 with the left-hand (right-hand) tongue emerging from the unstable (stable) natural wave. We then note that the left-hand tongue includes

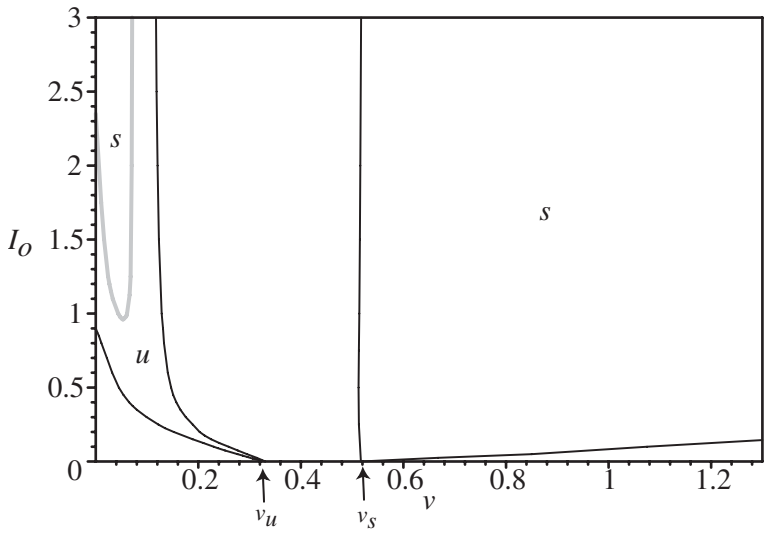


FIG. 3.2. *Regions of existence of the stimulus-locked traveling pulses in the* (v, I_0)*-plane for* $\sigma = 1.0$*,* $\kappa = 0.3$*,* $\epsilon = 0.03$*, and* $\beta = 2.5$*. The left and right regions form tongues that issue from the unstable* v_u *and stable* v_s *natural traveling pulses, respectively. The Hopf curve within the left-hand tongue is shown in gray. Stationary pulses correspond to the intersection of the tongue and the line* $v = 0$*.*

stationary pulses at $v = 0$. In previous work we have shown how a stationary unimodal input can generate a stable stationary pulse that bifurcates to a stable breather via a Hopf bifurcation as the input amplitude is reduced [5, 13]. In section 4 we construct the associated Evans function for traveling pulses within the tongue regions and use this to determine the stability of stimulus-locked pulses. We find that there is a Hopf curve within the left-hand tongue that is a continuation of the Hopf bifurcation point for stationary pulses ($v = 0$); this is shown in Figure 3.2 by the gray curve. Above the Hopf curve the pulse is stable, while it is unstable below. On the other hand the pulse within the right-hand tongue is always stable. Finally, note that there also exist additional stimulus-locked pulse solutions in certain subregions inside as well as outside of the tongues; however, these are found to be always unstable.

4. Stability of the stimulus-locked traveling pulse. We begin by analyzing the resolvent and the spectrum of the operator associated with the linearization of the vector system (1.2) about the general stimulus-locked traveling pulse constructed in section 3.1. This analysis indicates that potential instabilities arise only due to the behavior of eigenvalues, which can be determined by calculation of the zero set of the Evans function. We then present the explicit construction of the Evans function for the stimulus-locked traveling pulse in the particular case of the exponential weight distribution, and calculate the zero sets of this Evans function for the pulse existence tongues shown in Figure 3.2, thereby determining their stability.

4.1. Spectral analysis of the linearized operator. Consider the evolution of small smooth perturbations of the stimulus-locked traveling pulse with stationary form (U, Q),

$$u = U + \bar{\varphi},$$
$$q = Q + \bar{\psi}.$$

Substituting into the system expressed in traveling wave coordinates and linearizing, we find that the perturbations, to first order, satisfy

$$\frac{\partial \bar{\varphi}}{\partial t} - v\frac{\partial \bar{\varphi}}{\partial \xi} + \bar{\varphi} + \beta\bar{\psi} = \int_{\mathbb{R}} w(\xi-\eta)H'(U(\eta)-\kappa)\bar{\varphi}(\eta)d\eta, \tag{4.1}$$

$$\frac{\partial \bar{\psi}}{\partial t} - v\frac{\partial \bar{\psi}}{\partial \xi} - \epsilon\bar{\varphi} + \epsilon\bar{\psi} = 0. \tag{4.2}$$

Separating variables,

$$\begin{pmatrix} \bar{\varphi}(\xi,t) \\ \bar{\psi}(\xi,t) \end{pmatrix} = \begin{pmatrix} \varphi(\xi) \\ \psi(\xi) \end{pmatrix} e^{\lambda t}, \tag{4.3}$$

the spatial components $\varphi, \psi \in \mathcal{C}^1(\mathbb{R},\mathbb{C})$ satisfy the spectral problem

$$(\mathcal{L}+\mathcal{N}_s)\begin{pmatrix} \varphi \\ \psi \end{pmatrix} = \lambda\begin{pmatrix} \varphi \\ \psi \end{pmatrix}, \tag{4.4}$$

where

$$\mathcal{L} = v\frac{\partial}{\partial \xi} - \mathcal{A}, \qquad \mathcal{A} = \begin{bmatrix} 1 & \beta \\ -\epsilon & \epsilon \end{bmatrix}, \tag{4.5}$$

$$\mathcal{N}_s\begin{pmatrix} \varphi \\ \psi \end{pmatrix} = \begin{pmatrix} \frac{w(\xi-\xi_1)}{|U'(\xi_1)|}\,\varphi(\xi_1) + \frac{w(\xi-\xi_2)}{|U'(\xi_2)|}\,\varphi(\xi_2) \\ 0 \end{pmatrix}. \tag{4.6}$$

Resolvent and the point spectrum. Letting $\mathbf{z} = (\varphi,\psi)^T$, we seek to construct a bounded inverse by solving

$$(\mathcal{L}+\mathcal{N}_s-\lambda)\mathbf{z} = -\mathbf{f},$$

where $\mathbf{f} = (f_1,f_2)^T$ and $f_1, f_2 \in \mathcal{C}^0(\mathbb{R},\mathbb{C})$. Following the variation of parameters approach of Zhang [42], we find that the linearly independent solutions of the homogeneous problem $(\mathcal{L}-\lambda)\phi = 0$ are

$$\Phi_+(\xi,\lambda) = \begin{pmatrix} \beta \\ m_+ - 1 \end{pmatrix} e^{\left(\frac{\lambda+m_+}{v}\right)\xi},$$

$$\Phi_-(\xi,\lambda) = \begin{pmatrix} \beta \\ m_- - 1 \end{pmatrix} e^{\left(\frac{\lambda+m_-}{v}\right)\xi},$$

in which case we set

$$\mathbf{z}(\xi) = \Big[\Phi_+\Big|\Phi_-\Big]\begin{pmatrix} \bar{a}(\xi) \\ \bar{b}(\xi) \end{pmatrix}.$$

Subsequently, the coefficient functions are determined according to

$$\Big[\Phi_+\Big|\Phi_-\Big]\frac{\partial}{\partial \xi}\begin{pmatrix} \bar{a} \\ \bar{b} \end{pmatrix} = -\frac{1}{v}\Big(\mathcal{N}_s\mathbf{z}+\mathbf{f}\Big). \tag{4.7}$$

Inversion of $[\Phi_+|\Phi_-]$ leads to

$$\frac{\partial}{\partial \xi}\begin{pmatrix} \bar{a} \\ \bar{b} \end{pmatrix} = -\frac{1}{v\beta(m_+-m_-)}\Big[\Psi_+\Big|\Psi_-\Big]^T\Big(\mathcal{N}_s\mathbf{z}+\mathbf{f}\Big), \tag{4.8}$$

where

$$\Psi_+(\xi,\lambda) = \begin{pmatrix} 1-m_- \\ \beta \end{pmatrix} e^{-\left(\frac{\lambda+m_+}{v}\right)\xi},$$
$$\Psi_-(\xi,\lambda) = -\begin{pmatrix} 1-m_+ \\ \beta \end{pmatrix} e^{-\left(\frac{\lambda+m_-}{v}\right)\xi}.$$

For $\mathcal{R}e(\lambda) > -m_-$, we integrate over $[\xi,\infty)$ to obtain

$$\begin{pmatrix} \bar{a}(\xi) \\ \bar{b}(\xi) \end{pmatrix} = \begin{pmatrix} \bar{a}_\infty \\ \bar{b}_\infty \end{pmatrix} + \frac{1}{v\beta(m_+ - m_-)} \int_\xi^\infty \Big[\Psi_+\big|\Psi_-\Big]^T \big(\mathcal{N}_s\mathbf{z} + \mathbf{f}\big) d\eta,$$

where $\bar{a}_\infty, \bar{b}_\infty$ denote the values of $a(\xi), b(\xi)$ as $\xi \longrightarrow \infty$. Thus

$$\mathbf{z}(\xi) = \Big[\Phi_+\big|\Phi_-\Big] \begin{pmatrix} \bar{a}_\infty \\ \bar{b}_\infty \end{pmatrix} + \frac{1}{v\beta(m_+ - m_-)} \Big[\Phi_+\big|\Phi_-\Big] \int_\xi^\infty \Big[\Psi_+\big|\Psi_-\Big]^T \big(\mathcal{N}_s\mathbf{z} + \mathbf{f}\big) d\eta.$$

As we shall discuss, the integral term is bounded for all ξ, and, consequently, for a bounded solution to exist, we must require that $\bar{a}_\infty = \bar{b}_\infty = 0$. Thus

$$\mathbf{z}(\xi) = \frac{1}{v\beta(m_+ - m_-)} \Big[\Phi_+\big|\Phi_-\Big] \int_\xi^\infty \Big[\Psi_+\big|\Psi_-\Big]^T \big(\mathcal{N}_s\mathbf{z} + \mathbf{f}\big) d\eta,$$

which can be rewritten as

$$\begin{pmatrix} \varphi(\xi) \\ \psi(\xi) \end{pmatrix} - \Lambda_1(\lambda,\xi) \begin{pmatrix} \varphi(\xi_1) \\ 0 \end{pmatrix} - \Lambda_2(\lambda,\xi) \begin{pmatrix} \varphi(\xi_2) \\ 0 \end{pmatrix} = \mathcal{H}(\xi), \tag{4.9}$$

where

$$\Lambda_i(\lambda,\xi) = \frac{1}{v\beta(m_+ - m_-)} \Big[\Phi_+\big|\Phi_-\Big] \int_\xi^\infty \Big[\Psi_+\big|\Psi_-\Big]^T \frac{w(\eta-\xi_i)}{|U'(\xi_i)|}\, d\eta,$$
$$\mathcal{H}(\xi) = \frac{1}{v\beta(m_+ - m_-)} \Big[\Phi_+\big|\Phi_-\Big] \int_\xi^\infty \Big[\Psi_+\big|\Psi_-\Big]^T \mathbf{f}(\eta)\, d\eta.$$

Elements of Λ_i and $\mathcal{H}$ are finite sums of terms of the forms

$$\int_\xi^\infty e^{\left(\frac{\lambda+m_\pm}{v}\right)(\xi-\eta)} w(\eta-\xi_i) d\eta, \qquad \int_\xi^\infty e^{\left(\frac{\lambda+m_\pm}{v}\right)(\xi-\eta)} f_i(\eta) d\eta.$$

Using the Hölder inequality, it is straightforward to show that these terms, and hence Λ_i and $\mathcal{H}$, are bounded for all $\xi \in \mathbb{R}$ and for all $f_i \in \mathcal{C}^0(\mathbb{R},\mathbb{C})$. Now we must determine the conditions under which (4.9) has a unique solution. Since the solution $\mathbf{z}(\xi)$ is determined completely by the restrictions $\mathbf{z}(\xi_1)$ and $\mathbf{z}(\xi_2)$, we obtain the following finite-dimensional system by substituting $\xi = \xi_1, \xi_2$ into (4.9):

$$\Big(\mathbf{I} - \Delta(\lambda)\Big) \begin{pmatrix} \varphi(\xi_1) \\ \varphi(\xi_2) \end{pmatrix} = \begin{pmatrix} \mathcal{H}_1(\xi_1) \\ \mathcal{H}_1(\xi_2) \end{pmatrix},$$

where $\mathcal{H} = (\mathcal{H}_1, \mathcal{H}_2)^T$, $\bar{\Lambda}_i(\lambda,\xi) = (\,1\ 0\,)\ \Lambda_i(\lambda,\xi)\ (\,1\ 0\,)^T$, and

$$\Delta(\lambda,\xi_1,\xi_2) = \begin{pmatrix} \bar{\Lambda}_1(\lambda,\xi_1) & \bar{\Lambda}_2(\lambda,\xi_1) \\ \bar{\Lambda}_1(\lambda,\xi_2) & \bar{\Lambda}_2(\lambda,\xi_2) \end{pmatrix}.$$

This system has a unique solution if and only if $\mathbf{det}(\mathbf{I}-\Delta(\lambda))\neq 0$, resulting in a bounded inverse defined on all of $\mathcal{C}^0(\mathbb{R},\mathbb{C})\times\mathcal{C}^0(\mathbb{R},\mathbb{C})$. All such λ are elements of the resolvent set. Conversely, we cannot invert the operator for λ such that

$$\mathbf{det}\big(\mathbf{I}-\Delta(\lambda,\xi_1,\xi_2)\big)=0,$$

in which case

$$(\mathcal{L}+\mathcal{N}_s-\lambda)\mathbf{z}=0$$

has nontrivial solutions and λ is an element of the point spectrum. As in the scalar front case, the function

$$\mathcal{E}(\lambda,\xi_1,\xi_2)=\mathbf{det}\big(\mathbf{I}-\Delta(\lambda,\xi_1,\xi_2)\big),\qquad \mathcal{R}e(\lambda)>-m_- \tag{4.10}$$

identifies eigenvalues with its zero set, indicating that $\mathcal{E}$ is an Evans function for the set for which $\mathcal{R}e(\lambda)>-m_-$. In a similar fashion, a resolvent and an Evans function can be defined on the set for which $\mathcal{R}e(\lambda)<-m_+$; however, we do not pursue the explicit construction, as it does not reflect an instability of the stimulus-locked wave.

Continuous spectrum. Using arguments similar to those of the case of the scalar equation, it can be shown that the operator $\mathcal{N}_s:\mathcal{C}^1(\mathbb{R},\mathbb{C})\times\mathcal{C}^1(\mathbb{R},\mathbb{C})\longrightarrow\mathcal{C}^0(\mathbb{R},\mathbb{C})\times\mathcal{C}^0(\mathbb{R},\mathbb{C})$ is compact. Again this implies that the essential spectrum of $\mathcal{L}+\mathcal{N}_s$ is identical to that of $\mathcal{L}$. In the case of the vector operator $\mathcal{L}$, the continuous spectrum is the union of the disjoint sets of $\lambda=-m_\pm+iv\rho$, where $\rho\in\mathbb{R}$. To see this, assume such λ and consider the sequence of functions $\phi_n^\pm\in\mathcal{C}^1(\mathbb{R},\mathbb{C})\times\mathcal{C}^1(\mathbb{R},\mathbb{C})$, where n is a positive integer; $\mathcal{Y}_\pm$ are the eigenvectors of the matrix $\mathcal{A}$ defined in (4.5), corresponding to the eigenvalues $m_\pm$; and

$$\phi_n^\pm(\xi)=e^{i\rho\xi}\big(1-e^{-\xi^2/2n^2}\big)\mathcal{Y}_\pm.$$

If $\mathcal{Y}_\pm$ are normalized to unity, then $\big\|\phi_n^\pm\big\|_\infty=1$ for all n; however,

$$\Big\|\mathcal{L}\phi_n^\pm\Big\|=\frac{v}{n^2}\Big\|\xi e^{-\frac{\xi^2}{2n^2}}\Big\|\longrightarrow 0\qquad\text{as}\quad n\longrightarrow\infty.$$

Hence, $(\mathcal{L}-\lambda)^{-1}$ is unbounded, and λ is a member of the continuous spectrum of $\mathcal{L}+\mathcal{N}_s$.

Residual spectrum. To complete the characterization of the spectrum, we demonstrate that the set $\{\lambda\in\mathbb{C}:\mathcal{R}e(\lambda)\in(-m_+,-m_-)\}$ defines the residual spectrum of $\mathcal{L}+\mathcal{N}_s$. We must show that for such λ there exists a bounded inverse whose domain is not dense in $\mathcal{C}^0(\mathbb{R},\mathbb{C})\times\mathcal{C}^0(\mathbb{R},\mathbb{C})$. Consider our previous construction of the inverse operator $(\mathcal{L}+\mathcal{N}_s-\lambda)^{-1}$. Since we need calculate only the residual spectrum of $\mathcal{L}$, we integrate (4.8) over $[c,d]$, neglecting $\mathcal{N}_s$, to obtain

$$\begin{pmatrix}\bar a(d)\\ \bar b(d)\end{pmatrix}-\begin{pmatrix}\bar a(c)\\ \bar b(c)\end{pmatrix}=-\frac{1}{v\beta(m_+-m_-)}\int_c^d\Big[\Psi_+\big|\Psi_-\Big]^T f(\eta)d\eta.$$

There are only two cases to consider. First, taking $c=\xi$ and $d=\infty$, we examine the integral term of $\mathbf{z}(\xi)$, components of which have the form

$$\int_\xi^\infty e^{\left(\frac{\lambda+m_\pm}{v}\right)(\xi-\eta)}\Big[(1-m_\mp)f_1(\eta)+\beta f_2(\eta)\Big]d\eta.$$

Since $\lambda + m_- < 0$ and $v > 0$, all components are bounded, and hence $\mathcal{L} + \mathcal{N}_s - \lambda$ is bounded only if f either decays sufficiently fast such that

$$\int_\xi^\infty e^{\left(\frac{\lambda+m_+}{v}\right)(\xi-\eta)}\Big[(1-m_-)f_1(\eta)+\beta f_2(\eta)\Big]d\eta < \infty, \qquad \xi \in \mathbb{R},$$

or satisfies $(1-m_-)f_1(\eta)+\beta f_2(\eta) = 0$ for all η. Similarly, for $c = -\infty$ and $d = \xi$, we must require that

$$\int_{-\infty}^\xi e^{\left(\frac{\lambda+m_-}{v}\right)(\xi-\eta)}\Big[(1-m_+)f_1(\eta)+\beta f_2(\eta)\Big]d\eta < \infty, \qquad \xi \in \mathbb{R},$$

or $(1-m_+)f_1(\eta)+\beta f_2(\eta) = 0$ for all η. Since the union of all such f is not dense in $\mathcal{C}^0(\mathbb{R},\mathbb{C}) \times \mathcal{C}^0(\mathbb{R},\mathbb{C})$, we conclude that λ lies in the residual spectrum.

4.2. Evans function for stimulus-locked traveling pulses. The following gives the explicit construction of the Evans function for stimulus-locked waves in the case of a Gaussian input, Heaviside firing rate function, and exponential weight distribution. Note that this includes natural waves where $I_0 = 0$. After a lengthy calculation,

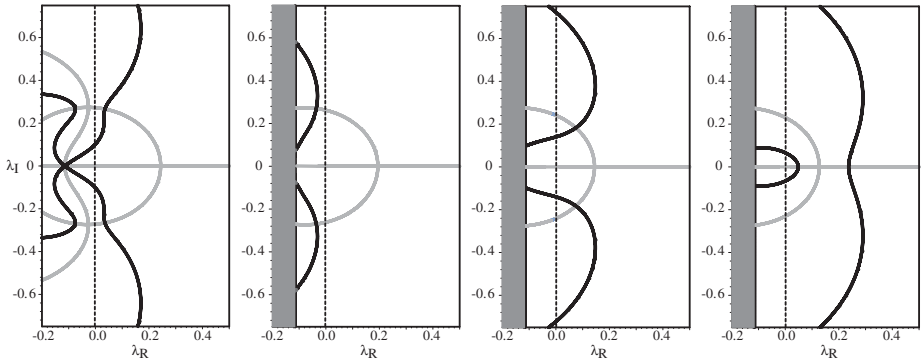
$$\begin{aligned}\mathcal{E}(\lambda,\xi_1,\xi_2) &= \mathbf{det}\big(\mathbf{I}-\Delta(\lambda,\xi_1,\xi_2)\big) \qquad\qquad \Re e(\lambda) > -m_-\\ &= \left(1-\frac{\Theta_+(\lambda)}{\big|U'(\xi_1)\big|}\right)\left(1-\frac{\Theta_+(\lambda)}{\big|U'(\xi_2)\big|}\right) - \frac{\Theta_+(\lambda)\Gamma(\lambda)}{\big|U'(\xi_1)U'(\xi_2)\big|}e^{(\xi_1-\xi_2)},\end{aligned} \tag{4.11}$$

where

$$\begin{aligned}\Gamma_\pm(\lambda) &= \frac{(1-m_\mp)v}{(m_+-m_-)(v^2-(\lambda+m_\pm)^2)},\\ \Theta_\pm(\lambda) &= \frac{1}{2(m_+-m_-)}\left(\frac{1-m_-}{\lambda+m_+\pm v}-\frac{1-m_+}{\lambda+m_-\pm v}\right),\\ \Gamma(\lambda) &= \Theta_-(\lambda)e^{(\xi_1-\xi_2)}+\Gamma_+(\lambda)e^{\left(\frac{\lambda+m_+}{v}\right)(\xi_1-\xi_2)}-\Gamma_-(\lambda)e^{\left(\frac{\lambda+m_-}{v}\right)(\xi_1-\xi_2)}.\end{aligned} \tag{4.12}$$

Since the zero set of the Evans function (4.11) comprises solutions of a transcendental equation, we solve for the eigenvalues numerically by finding the intersection points of the zero sets of the real and complex parts of the Evans function. This leads to the stability results shown in Figure 3.2, namely, that pulses within the right-hand tongue are stable whereas pulses within the left-hand tongue are stable only if they lie inside the region enclosed by the Hopf curve. An example of a zero set construction is shown in Figure 4.1 for fixed I_0 and various values of v.

Linear stability of the traveling pulse solution is characterized by all eigenvalues of the linearization having negative real part, with the possible exception that $\lambda = 0$ is a simple eigenvalue. Moreover, Hopf bifurcations may be identified by a pair of complex eigenvalues crossing the imaginary axis from the left-half plane. It has been found in many infinite-dimensional dynamical systems, such as semilinear parabolic equations, that the criterion for a Hopf bifurcation carries over from ordinary differential equations. Although smoothness properties of the flow are required for its proof using invariant manifold theory, the result is *essentially* based on the behavior of eigenvalues of the linearized operator [26]. We shall assume this and use numerical



(a) $v = 0.0$ (b) $v = 0.05$ (c) $v = 0.1$ (d) $v = 0.1175$

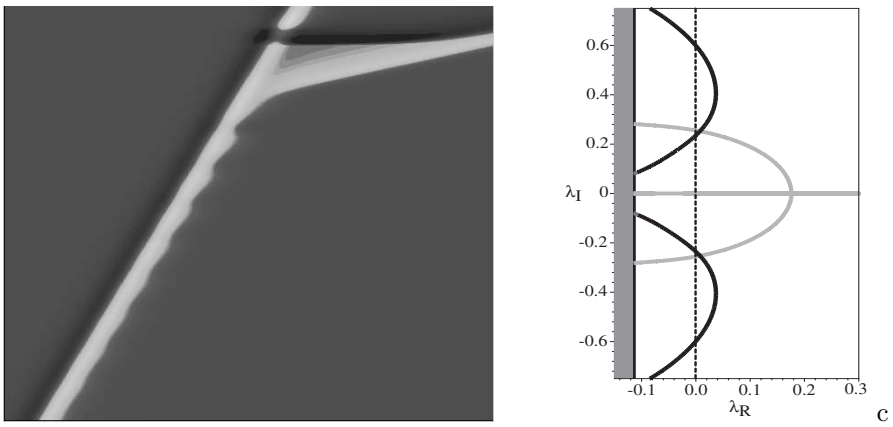
FIG. 4.1. *Graphs of the zero sets of the real (dark curves) and imaginary (light curves) parts of the Evans functions for* $I_0 = 2.0$ *and a sequence of stimulus speeds* v*; intersection points indicate eigenvalues. Note that the horizontal gray line is part of the zero set of the imaginary part. The vertical shaded region* $\mathcal{R}e(\lambda) \leq -m_-$ *indicates the essential spectrum. This sequence of plots indicates that two Hopf bifurcation points occur, thus defining the boundary of the stable region within the left tongue depicted in Figure* 3.2. *Case* (a) *is associated with the existence of a stable stationary breather, case* (b) *with a stable traveling pulse, and cases* (c) *and* (d) *with a traveling emitter. See text for more details.*

simulations, as discussed in the following section, to explore the behavior of the model near these bifurcation points. Note, for $I_0 > 0$, $\lambda = 0$ is not an eigenvalue and does not complicate the eigenvalue criteria of the standard Hopf bifurcation theorem, as would be the case with natural waves.

4.3. Numerical simulations. In this section we explore the behavior of the vector system (1.2) in all regions of the (v, I_0)-plane shown in Figure 3.2. In particular, we describe the various types of solutions that emerge beyond the Hopf bifurcation curve, as well as beyond the existence tongues.

For parameter values supporting natural traveling waves, and in the absence of an input ($I_0 = 0$), an initial sufficiently large local displacement of the activity u from rest induces a locally excited region of activity, which rapidly develops into a pair of diverging natural traveling pulses, as in the reaction diffusion analogue. Similarly, for parameter values supporting stable stimulus-locked waves in the presence of an input ($I_0 > 0$), an initial displacement of u near the input (or no initial displacement in the case of sufficiently large input strength I_0) rapidly approaches the stable traveling pulse. For certain speeds v the initial transient can generate an additional single or pair of traveling waves that propagate away from the input. As expected, the speed and width of the stimulus-locked traveling pulse closely match those of the theory.

Interestingly, for the parameter values in Figure 3.2, numerical simulations suggest that the left-hand branch of the Hopf curve (gray) corresponds to a supercritical bifurcation, while the right-hand branch is subcritical without a sharp transition to a breathing pulse. We first characterize the nature of solutions obtained by crossing the subcritical branch of the Hopf curve. We find a region of activity moving with the input whose right boundary oscillates with increasing amplitude. After a critical point, the system emits a natural traveling pulse, whose speed is faster than that of the input, as shown in Figure 4.2. The region between the one excited by the

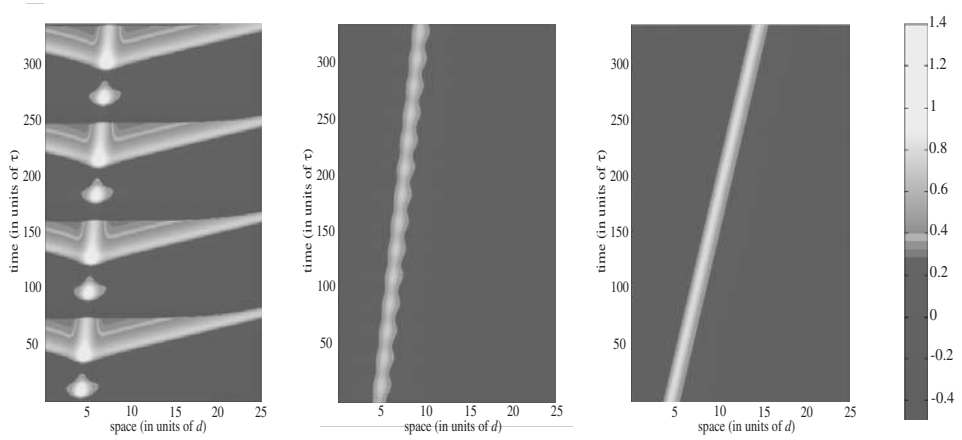


(a) (b)

FIG. 4.2. *Instability of the stimulus-locked traveling pulse in the presence of two complex conjugate eigenvalues with positive real part for* $I_0 = 1.0$, $v = 0.07$, $\sigma = 1$, $\kappa = 0.3$, $\epsilon = 0.03$, *and* $\beta = 2.5$. *In this case the bifurcation appears subcritical with the absence of a sharp jump to a stable breathing pulse. Instead, instability manifests itself as a periodic cycling of an initial phase of periodically modulated growth of the active region, followed ultimately by the shedding of a natural traveling pulse.* (a) *Space-time plot showing one cycle of the instability, where the vertical axis represents time and the horizontal axis represents space.* (b) *Graph of the corresponding zero set of the Evans function. The periodic process of shedding or emitting natural traveling pulses becomes more rapid as the real part of the eigenvalue increases.*

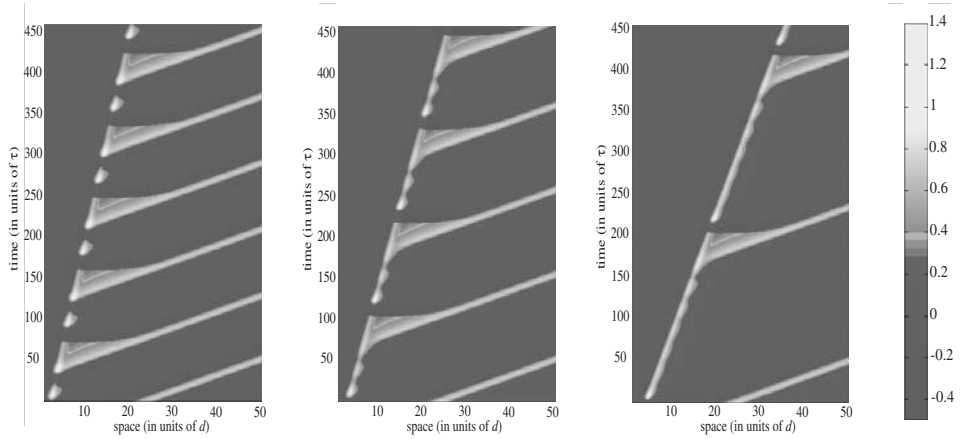
input and the new natural wave recovers, and the process repeats periodically. Such solutions we refer to as *pulse-emitters*. The smaller the real part of the eigenvalue, the slower the instability grows and the more time is required for the wave to be emitted. As v is increased, the real part of the eigenvalue grows and the number of oscillations occurring before the shedding of natural waves decreases, until the eigenvalues become real, as illustrated in the figure sequence 4.1(b)–(d), and the pulse rapidly emits natural pulses. This behavior continues until v is increased to the boundary of the right-hand tongue where there is a smooth transition to a stable stimulus-locked pulse.

When the left-hand supercritical branch of the Hopf curve is crossed by reducing I_0 or v, we find a smooth transition to a stimulus-locked traveling breather. In the special case of a stationary stimulus ($v = 0$), reducing I_0 generates a stationary breather, as we have shown previously [5, 13]. The breathing solutions continue to persist in a subregion of the (v, I_0)-plane bounded to the right by the left (supercritical) branch of the Hopf curve in 3.2. As one moves in this subregion away from the left Hopf branch, the amplitude of the oscillations grows. After some point, the breathing solution disappears, and a new type of temporally periodic solution appears, each cycle of which is characterized by one or more breathing pulse oscillations followed by the emission of a pair of natural waves, possibly intermixed with interludes of subthreshold behavior. An example of such a transition is illustrated in Figure 4.3. This type of pulse-emitting solution appears to be part of a family of related responses of the system to a localized input, which also includes the pulse-emitting behavior associated with the region between the subcritical Hopf curve and the stable right



(a) $v = 0.01$ (b) $v = 0.014$ (c) $v = 0.03$

FIG. 4.3. *Sequence of space-time plots for fixed input* $I_0 = 1.5$*, illustrating the transition from pulse emitter, to breather, to stimulus-locked pulse as* v *increases through the supercritical branch of the Hopf curve shown in Figure* 3.2*. Other parameters are* $\epsilon = 0.03$*,* $\kappa = 0.3$*,* $\beta = 2.5, \sigma = 1$*.*

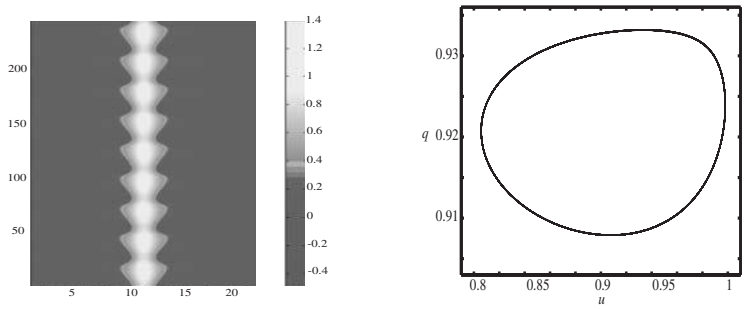


(a) $v = 0.04$ (b) $v = 0.05$ (c) $v = 0.07$

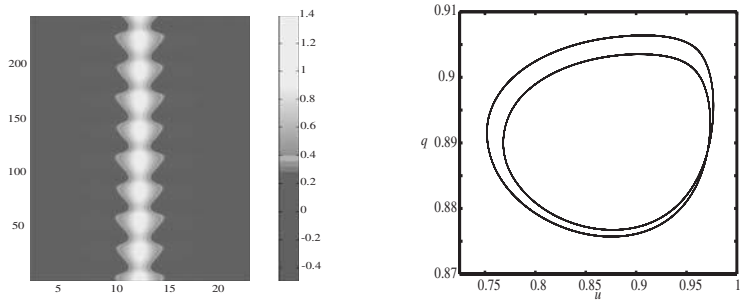
FIG. 4.4. *Transitions between various pulse-emitting solutions for fixed* $I_0 = 0.9$ *as* v *is increased. These solutions exist within the unstable part of the left-hand tongue of Figure* 3.2*, sufficiently below the Hopf curve such that stable breathers no longer exist. Other parameters are* $\epsilon = 0.03$*,* $\kappa = 0.3$*,* $\beta = 2.5$*,* $\sigma = 1$*.*

tongue shown in Figure 3.2. Furthermore, there is a smooth transition of behaviors joining the two regions, as shown in Figure 4.4.

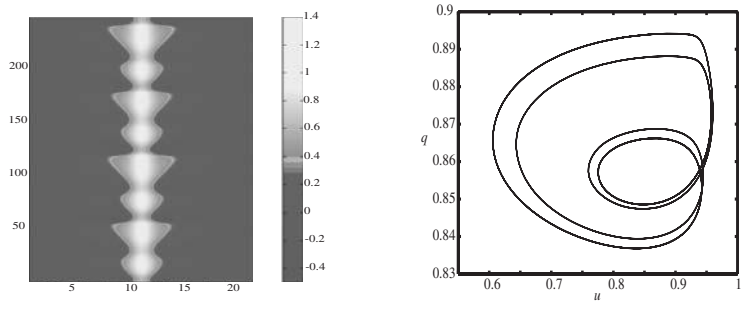
Although the above account applies to the case $\sigma = 1$, most features are valid for more general σ. One main point of difference lends insight into the disappearance of the breather. If we consider stationary pulses for $\sigma = \sqrt{2}$ and explore the evolution of the breathing pulse as we further decrease I_0 beyond the bifurcation point, we find that a secondary bifurcation occurs, giving rise to two modes of breathing rather than



(a) $I_0 = 2.4$



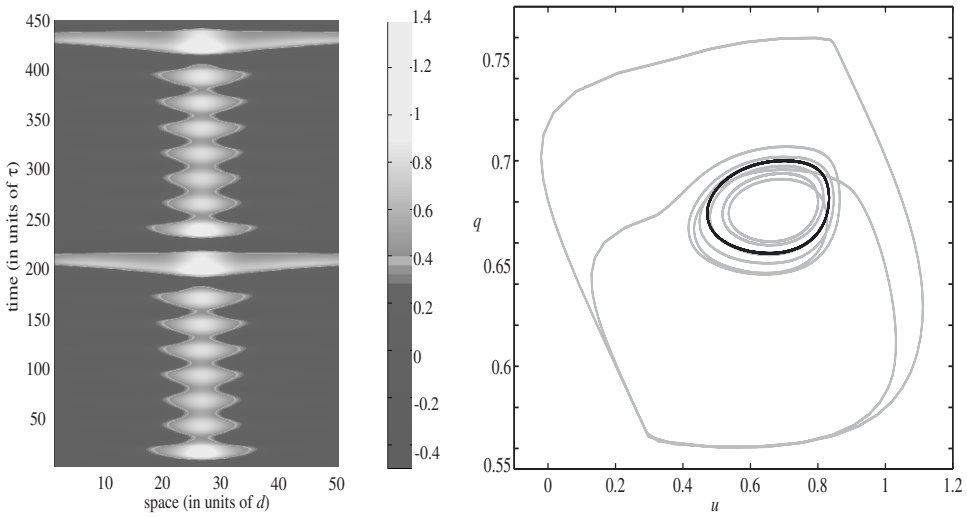
(b) $I_0 = 2.3$



(c) $I_0 = 2.2$

FIG. 4.5. *Sequence of period-doubling bifurcations of a breathing pulse for* $\sigma = \sqrt{2}$. *The left-hand column shows space-time plots for different values of current amplitude beyond the initial Hopf bifurcation point, with an orbit corresponding to the center spatial point plotted in the* (u, q)*-phase plane in the right-hand column; other spatial points are qualitatively similar. Other parameter values are* $\kappa = 0.3$, $\beta = 2.5$, $\epsilon = 0.03$, $v = 0$. *(Note that at higher resolution each loop in* (c) *is actually a pair of closely spaced loops, indicating that it corresponds to the third doubling in the sequence.)*

one. By graphing, in the (u, q)-phase plane, the orbit corresponding to a spatial point at the center of the input, we find that the evolution of the orbit, as I_0 is decreased, strongly resembles that of a period-doubling bifurcation, as shown in Figure 4.5(a)–(b). Decreasing I_0 leads to additional period doublings, as illustrated in Figure 4.5(c). Ultimately, decreasing I_0 leads to behavior similar to that found for $\sigma = 1$. This suggests that for $\sigma = 1$ the first period-doubling bifurcation may be subcritical, and the orbit instead weaves its way around the unstable limit cycle giving rise to the sequence of breathing pulses and emission, as shown in Figure 4.6.



(a) (b)

FIG. 4.6. (a) *Space-time plot of a stationary ($v = 0$) pulse-emitter for $\sigma = 1$, $I_0 = 1.35$, $\kappa = 0.3$, $\beta = 2.5$ and $\epsilon = 0.03$.* (b) *Corresponding phase portrait showing the orbit (gray trajectory) of the center spatial point plotted in the (u, q)-phase plane. Also shown is the corresponding orbit (black trajectory) of the stable breather that exists when $I_0 = 1.4$.*

5. Discussion. In this paper we have shown how to extend the analysis of the existence and stability of pulses arising from a stationary stimulus input to that of a input moving with constant speed. We described the continuation from the unstable/stable pair of natural waves by constructing a corresponding pair of existence tongues emerging from the natural waves at $I_0 = 0$, with the left-hand tongue including stationary pulses at $v = 0$, for a particular choice of parameter values supporting natural waves. We have extended Zhang's analysis of stability of natural waves to that of stimulus-locked waves and numerically evaluated the Evans function to determine eigenvalues away from the singular limit $\epsilon \to 0$. This allowed us to analyze the stability of the existence tongues in the (v, I_0)-plane and show the continuation of the Hopf bifurcation found for stationary pulses. Numerically this Hopf curve was found to have a supercritical branch, from which breathing pulses emerge and a subcritical branch from which no breathing pulse emerges. In general for parameter values that do not support either stimulus-locked pulses or breathers, the system generates more complicated behavior, including the emission of natural traveling waves when such waves exist.

It would be interesting to contrast the type of local inhibition analyzed in this paper, which is primarily due to intrinsic neuronal properties, with that of nonlocal inhibition, arising from the ubiquitous inhibitory populations of neurons found in cortex. From previous work [1, 28], we know that the two-population, excitatory-inhibitory system supports stable stationary pulses which, moreover, can undergo a subcritical Hopf bifurcation. In this case no breathing pulse emerges; however, it is possible that the presence of a localized input is capable of stabilizing such a breathing pulse solution. In addition, it would be interesting to provide a more thorough analysis

of the scalar model considered by Xie and Giese [40], by constructing tongue diagrams and Hopf bifurcation curves and, furthermore, considering the effect of varying the degree of nonlocal inhibition.

From a more general perspective, the analysis presented here and in related work [6, 13] has established that the combined effect of local inhomogeneities and recurrent synaptic interactions can result in nontrivial forms of coherent oscillations and waves. Although we have focused on rather abstract neural field equations, we expect our results to carry over (at least qualitatively) to more biophysically realistic conductance-based models. Indeed, elsewhere we have confirmed the existence of stationary breathers and pulse emitters in the case of a modified Traub model [13]. One of the advantages of studying simplified models is that it can generate predictions regarding how dynamical properties such as wave speed depend on characteristic features of neural tissue. One striking demonstration of this is the recent study of wave propagation in disinhibited cortical slices, where the speed of the wave was controlled by external electric fields, confirming predictions based on homogeneous neural field equations [32]. Our own work predicts that coherent oscillations can be induced by local inhomogeneities. Such inhomogeneities could arise from external stimuli or reflect changes in the excitability of local populations of neurons. The former suggests a network mechanism for stimulus-induced oscillations, which may play an important role in visual processing [17], whereas the latter suggests a network mechanism for generating epileptiform activity.

REFERENCES

[1] S. Amari, *Dynamics of pattern formation in lateral inhibition type neural fields*, Biol. Cybernet., 27 (1977), pp. 77–87.

[2] M. Bode, *Front bifurcations in reaction-diffusion systems with inhomogeneous parameter distributions*, Phys. D, 106 (1997), pp. 270–286.

[3] P. C. Bressloff, *Traveling waves and pulses in a one-dimensional network of excitable integrate-and-fire neurons*, J. Math. Biol., 40 (2000), pp. 169–198.

[4] P. C. Bressloff, *Traveling fronts and wave propagation failure in an inhomogeneous neural network*, Phys. D, 155 (2001), pp. 83–100.

[5] P. C. Bressloff, S. E. Folias, A. Prat, and Y-X. Li, *Oscillatory waves in inhomogeneous neural media*, Phys. Rev. Lett., 91 (2003), 78101.

[6] P. C. Bressloff and S. E. Folias, *Front bifurcations in an excitatory neural network*, SIAM J. Appl. Math., 65 (2004), pp. 131–151.

[7] R. D. Chervin, P. A. Pierce, and B. W. Connors, *Periodicity and directionality in the propagation of epileptiform discharges across neocortex*, J. Neurophysiol., 60 (1988), pp. 1695–1713.

[8] S. Coombes and M. R. Owen, *Evans functions for integral neural field equations with Heaviside firing rate function*, SIAM J. Appl. Dynam. Syst., 3 (2004), pp. 574–600.

[9] G. B. Ermentrout and J. B. Mcleod, *Existence and uniqueness of travelling waves for a neural network*, Proc. Roy. Soc. Edinburgh A, 123 (1993), pp. 461–478.

[10] J. W. Evans, *Nerve axon equations:* IV. *The stable and unstable impulse*, Indiana Univ. Math. J., 24 (1975), pp. 1169–1190.

[11] J. W. Evans and J. Feroe, *Local stability theory of the nerve impulse*, Math. Biosci., 37 (1977), pp. 23–50.

[12] R. FitzHugh, *Impulses and physiological states in theoretical models of nerve membrane*, Biophys. J., 1 (1961), pp. 445–466.

[13] S. E. Folias and P. C. Bressloff, *Breathing pulses in an excitatory neural network*, SIAM J. Appl. Dynam. Syst., 3 (2004), pp. 378–407.

[14] D. Golomb and Y. Amitai, *Propagating neuronal discharges in neocortical slices: Computational and experimental study*, J. Neurophysiol., 78 (1997), pp. 1199–1211.

[15] D. Golomb and G. B. Ermentrout, *Continuous and lurching traveling pulses in neuronal networks with delay and spatially decaying connectivity*, Proc. Natl. Acad. Sci. USA, 96 (1999), pp. 13480–13485.

[16] D. Golomb and G. B. Ermentrout, *Bistability in pulse propagation in networks of excitatory and inhibitory populations*, Phys. Rev. Lett., 86 (2001), pp. 4179–4182.
[17] C. M. Gray, *Synchronous oscillations in neuronal systems: Mechanisms and functions*, J. Comput. Neurosci., 1 (1994), pp. 11–38.
[18] A. Hagberg and E. Meron, *Pattern formation in non-gradient reaction-diffusion systems: The effects of front bifurcations*, Nonlinearity, 7 (1994), pp. 805–835.
[19] A. Hagberg, E. Meron, I. Rubinstein, and B. Zaltzman, *Controlling domain patterns far from equilibrium*, Phys. Rev. Lett., 76 (1996), pp. 427–430.
[20] A. L. Hodgkin and A. F. Huxley, *A quantitative description of membrane current and its application to conduction and excitation in nerve*, J. Physiol., 117 (1952), pp. 500–544.
[21] M. A. P. Idiart and L. F. Abbot, *Propagation of excitation in neural network models*, Network, 4 (1993), pp. 285–294.
[22] C. K. R. T. Jones, *Stability of the traveling wave solution of the FitzHugh–Nagumo system*, Trans. Amer. Math. Soc., 286 (1984), pp. 431–469.
[23] T. Kapitula, N. Kutz, and B. Sandstede, *The Evans function for nonlocal equations*, Indiana Univ. Math. J., 53 (2004), pp. 1095–1126.
[24] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
[25] R. Maex and G. A. Orban, *Model circuit of spiking neurons generating direction selectivity in simple cells*, J. Neurophysiol., 75 (1996), pp. 1515–1545.
[26] J. Marsden and M. McCracken, *Hopf Bifurcation and Its Applications*, Appl. Math. Sci. 10, Springer-Verlag, New York, 1976.
[27] D. J. Pinto and G. B. Ermentrout, *Spatially structured activity in synaptically coupled neuronal networks:* I. *Traveling fronts and pulses*, SIAM J. Appl. Math., 62 (2001), pp. 206–225.
[28] D. J. Pinto and G. B. Ermentrout, *Spatially structured activity in synaptically coupled neuronal networks:* II. *Lateral inhibition and standing pulses*, SIAM J. Appl. Math, 62 (2001), pp. 226–243.
[29] D. Pinto, S. L. Patrick, W. C. Huang, and B. W. Connors, *The fine structure of epileptiform activity in neocortex in vitro*, submitted.
[30] D. Pinto, S. L. Patrick, W. C. Huang, and B. W. Connors, *Mechanisms of initiation, propagation, and termination of epileptiform activity in neocortex in vitro*, submitted.
[31] A. Prat and Y.-X. Li, *Stability of front solution in inhomogeneous media*, Phys. D, 186 (2003), pp. 50–68.
[32] K. A. Richardson, S. J. Schiff, and B. J. Gluckman, *Control of traveling waves in the mammalian cortex*, Phys. Rev. Lett., 94 (2005), paper 028103.
[33] J. Rinzel and J. P. Keener, *Hopf bifurcation to repetitive activity in nerve*, SIAM J. Appl. Math., 43 (1983), pp. 907–922.
[34] J. Rubin, *A nonlocal eigenvalue problem for the stability of a traveling wave in a neuronal medium*, Discrete Contin. Dynam. Syst. A, 4 (2004), pp. 925–940.
[35] B. Sandstede, *Stability of travelling waves*, in Handbook of Dynamical Systems II, B. Fiedler, ed., North–Holland, Amsterdam, 2002, pp. 983–1055.
[36] P. Schutz, M. Bode, and H.-G.- Purwins, *Bifurcations of front dynamics in a reaction-diffusion equation system with spatial inhomogeneities*, Phys. D, 82 (1995), pp. 382–397.
[37] H. Suarez, C. Koch, and R. Douglas, *Modeling direction selectivity of simple cells in striate visual cortex within the framework of the canonical microcircuit*, J. Neurosci., 15 (1995), pp. 6700–6719.
[38] H. R. Wilson and J. D. Cowan, *A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue*, Kybernetik, 13 (1973), pp. 55–80.
[39] J.-Y. Wu, L. Guan, and Y. Tsau, *Propagating activation during oscillations and evoked responses in neocortical slices*, J. Neurosci., 19 (2001), pp. 5005–5015.
[40] X. Xie and M. Giese, *Nonlinear dynamics of direction-selective recurrent neural media*, Phys. Rev. E, 65 (2002), paper 051904.
[41] K. Yosida, *Functional Analysis*, Springer-Verlag, New York, 1968.
[42] L. Zhang, *On stability of traveling wave solutions in synaptically coupled neuronal networks*, Differential Integral Equations, 16 (2003), pp. 513–536.
[43] L. Zhang, *Existence, uniqueness, and exponential stability of traveling wave solutions of some integral differential equations arising from neuronal networks*, J. Differential Equations, 197 (2004), pp. 162–196.

SIAM J. APPL. MATH.
Vol. 65, No. 6, pp. 2093–2106

ENERGY MAXIMIZERS, NEGATIVE TEMPERATURES, AND ROBUST SYMMETRY BREAKING IN VORTEX DYNAMICS ON A NONROTATING SPHERE*

CHJAN C. LIM†

Dedicated to Bro. Felix Donohue, La Sallian, with friendship and respect.

Abstract. This paper relates the existence and uniqueness of constrained energy maximizers to the occurrence of negative temperatures in a recent statistical mechanics model of the energy-enstrophy theory. We construct examples of steady state solutions of the vorticity equation which break $SO(3)$ symmetry from the negative temperature vorticity distributions in the spherical model. These vortex states correspond to solid-body rotation flows at rotation rates Θ, which depend only on the fixed value of enstrophy Γ, that is, $\Theta = \sqrt{\Gamma/(4\int_{S^2}\cos^2\theta\, dx)}$. They are robust in the sense that they constitute most probable states in a spherical model of the statistical energy-enstrophy theory at negative temperatures, and have exponentially large Gibbs probability relative to any other macrostates. The existence and uniqueness of energy maximizers in a variational formulation of the new energy-enstrophy theory also give a necessary condition for the spherical model energy-enstrophy theory to be well defined at all temperatures.

Key words. energy maximizers, energy-enstrophy model on a sphere, spherical model, negative temperature

AMS subject classifications. 76B47, 49S05, 82B23

DOI. 10.1137/040605916

1. Introduction. Negative temperatures exist in many physical systems including lasers. The work of Onsager [3], Montgomery and Joyce [2], and Eyink and Spohn [4] shows that negative temperatures arise in two-dimensional (2d) point vortex statistics. We have extended this to vortex statistics on the nonrotating sphere by constructing a convergent family of lattice spin models H_N. In each member of the family parametrized by the number N of lattice nodes placed on the sphere S^2, we construct a well-defined equilibrium statistic, using a new version of the energy-enstrophy theory known as the spherical model:

$$Z_N = \int (\Pi ds_j)\delta\left(N\Gamma - \sum s_j^2\right)\exp\left[-\beta H_N\right]. \tag{1.1}$$

One can regard the spin state $\vec{s}(N) = (s_1, \ldots, s_N)$ as a macrostate or coarse-grained vorticity distribution for which there are many equivalent microstates, each of which can be viewed as a rearrangement of distinguishable vorticity parcels of similar strength between lattice cells/sites. A spin state $\vec{s}$ tends in the continuum/thermodynamic limit to a vorticity function $w(\theta, \phi) \in L_2(S^2)$, where $L_2(S^2)$ denotes the Hilbert space of square-integrable measurable functions on the sphere S^2 with coordinates given by co-latitude θ and longitude ϕ. We sketch one derivation of the spherical model in this paper.

*Received by the editors March 29, 2004; accepted for publication (in revised form) December 9, 2004; published electronically August 9, 2005. This work was supported by ARO grant W911NF-05-1-0001 and DOE grant DE-FG02-04ER25616.
http://www.siam.org/journals/siap/65-6/60591.html
†Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180 (limc@rpi.edu).

This new version was done for two specific reasons: the classical energy-enstrophy theory based on a Gibbs canonical ensemble in the energy and enstrophy [8] is equivalent to the Gaussian model [13], [12], [1] and has a vanishingly small temperature range of validity in the continuum limit (only for $\beta = 0$); changing the canonical constraint on the enstrophy to a microcanonical constrain converted the resulting statistics into that of Kac's spherical model, which is exactly solvable in the continuum limit of N tending to infinity and, more importantly, it is valid for all values of inverse temperature β, including negative ones. The exact solution in the continuum/thermodynamic limit of the spherical model for energy-enstrophy theory has been discussed in Lim [13], [12] and simulated in Lim and Nebus [29]. One of the main conclusions from this work is that for large values of kinetic energy H relative to the enstrophy Γ, when $\beta < 0$, there is only one most probable vorticity distribution, namely, the solid-body rotation state [13], [12], [29]. This macrostate is directly related to the spherical harmonic Y_{10} corresponding to the smallest positive eigenvalue of the Laplace–Beltrami operator $-\Delta$ on the sphere S^2.

In this paper, we go deeper into the relationship between negative temperature states in the spherical model for energy-enstrophy theory and the existence of unique maximal energy states in a variational formulation of the constrained enstrophy problem. The main results of the paper are the existence and uniqueness of these maximal energy states modulo a symmetry group in a variational formulation of the problem on the surface of the sphere, and two applications. First, we discuss an important application to statistical physics in the following subsection, concerning a necessary condition for a statistical mechanics energy-enstrophy theory to be well defined, namely, bounds on the energy. The second application concerns the existence and robustness of symmetry-breaking steady-states of the Euler equations and the ergodicity of inviscid vortex dynamics on a nonrotating sphere. This application is motivated by Shepherd's discovery that on a rapidly rotating sphere, inviscid single layer vortex dynamics is highly anisotropic and nonergodic. His proof of nonergodicity does not extend to the nonrotating sphere. We have implicitly assumed that inviscid vortex dynamics on the nonrotating sphere are ergodic when formulating the spherical model energy-enstrophy theory.

1.1. Unboundedness in the classical energy-enstrophy theory. What then is the main advantage of the spherical model over the older Kraichnan energy-enstrophy theory? The results in this paper establish a necessary condition for the spherical model to be well defined for all temperatures; that is, the energy $H[w]$ has a finite upper bound for fixed enstrophy $\Gamma[w]$. This is not the case in the Kraichnan energy-enstrophy theory where, because enstrophy is constrained in a canonical manner, the augmented energy has the form

$$E[w] = H[w] + \frac{\mu}{\beta}\Gamma[w].$$

We have recently shown in [25] that $E[w]$ has no upper bound for $\frac{\mu}{\beta} < 0$ when

$$-2\frac{\mu}{\beta} < \|G\|,$$

where G is the integral operator whose kernel is the Green's function of the Laplace–Beltrami operator on the sphere S^2 when $\int_{S^2} w\,dx = 0$. A more serious difficulty arises for the classical energy-enstrophy theory for inviscid flows on a nonrotating

sphere when the temperature is positive. Namely, for $\beta > 0$ and $\mu > 0$, the unique global minimizer w_m of $E[w]$ is the trivial vorticity $w_m = 0$ [25]; that is

$$E[w_m] = \min_{\int_{S^2} w dx = 0} E[w] = 0.$$

This positive temperature difficulty is not as serious in Kraichnan's original energy-enstrophy theory for inviscid flows on the plane as it is for the sphere S^2, because $\int w dx$ need not be zero for planar flows.

General dissatisfaction with the classical energy-enstrophy theory has led to several new directions for research. Works criticizing the classical energy-enstrophy theory include the Miller–Robert theory [19], [18], which invokes an infinite number of vorticity constraints; the Turkington–Majda model [21], [22], which is based on information theoretic Bayesian statistics; and the spherical model of energy-enstrophy-circulation theory [13], [12], [14], which is based on a microcanonical constraint on enstrophy rather than the canonical one in Kraichnan's theory [8]. Like the Miller–Robert [19], [18] and Turkington–Majda [21], [22] theories, the spherical model has predicted much the same results obtained by the older energy-enstrophy theories. They are all capable of supporting negative temperatures and the inverse energy cascade, which are ubiquitous in 2d flows, when the effective rotation rates are small enough or zero. The advantages of the spherical model over the other recent theories lie in its relative simplicity of constraining only energy and enstrophy—which follows from the deep relationships between them and has important consequences further discussed below—and the fact that there is an exact analytical solution for its partition function [14].

1.2. Ergodicity and robust symmetry breaking. This paper also examines the phenomena of symmetry breaking in inviscid vortex dynamics on a nonrotating sphere in relation to its ergodicity. We will discuss the following questions: (a) Do inviscid 2d flows on a nonrotating sphere exhibit anisotropy strong enough to imply nonergodicity? (b) Is this symmetry breaking robust in some other suitable sense? The answer to the first question is yes, there is symmetry breaking or anisotropy, but there is no evidence of nonergodicity of 2d inviscid flows on a nonrotating sphere. The answer to the second question is also yes, the symmetry breaking that we discover for vortex dynamics on the nonrotating sphere is robust in the sense of having exponentially high probability in a Gibbs canonical ensemble.

Invoking statistical mechanics to prove the robustness of the symmetry breaking basic states on the nonrotating sphere is more than just a convenient and physically valid tool. There is no known Lyapunov stability result for the zonally symmetric basic states in this problem, unlike the case of a strongly rotating sphere [6]. Moreover, rigorous arguments prove that in most high dimensional Hamiltonian systems, some trajectories near these steady states must escape along a chain of whiskered tori [28]. Without recourse to results on dynamical stability, the tools of equilibrium statistical mechanics are indispensable. Provided we are satisfied that the problem is ergodic, only the issue of which statistical mechanics model to use matters, and is one of the main topics of discussion in this paper.

An elegant paper by Shepherd [6] argued convincingly that inviscid 2d flows on the beta plane and rotating sphere are nonergodic for sufficiently large beta and rotation rates. His arguments are based on a version of Arnold's stability theorems for zonally symmetric basic solutions. This method requires the beta or rotation rate to be not only nonzero but large. Numerical experiments confirm this anisotropy in unforced

flows on the sphere [15], [16] and forced flows [20]. The anisotropy discovered by Shepherd [6] and the resulting nonergodic property of the (large beta) beta plane and strongly rotating sphere models put into question whether equilibrium statistical mechanics is valid for these models.

On the other hand, there is no known proof that vortex dynamics on the nonrotating sphere is nonergodic. But like many other parts of physics, there is no proof of ergodicity either. In light of this situation, we have the usual physicist's justifications to use equilibrium statistical mechanics to demonstrate the robustness of symmetry breaking.

There are two special relationships between energy and enstrophy which make them natural constraints in any statistical mechanics theory for inviscid flows. First and foremost is the fact that the enstrophy is just the L_2 norm of the vorticity w, and the kinetic energy is the quadratic form whose kernel is the Green's function of the Laplace–Beltrami operator in the problem. Poincaré's inequality then implies that the Dirichlet quotient of enstrophy over energy is bounded below by the spectrum of the Laplace–Beltrami operator in the domain. This yields a very natural variational framework which is further exploited in this paper.

Next is the powerful theory called the principle of selective decay [7] or minimum enstrophy, which tells us that the Dirichlet quotient tends to a nontrivial and fixed minimum in the unforced periodic 2d Navier–Stokes equations, even as the energy and the enstrophy separately tend to zero under the effect of viscosity. This principle, which provides a valuable link between inviscid models and the long time dynamics of damped 2d flows, is again based only on the two quadratic forms of energy and enstrophy.

As far as we know, there are no equally compelling results for higher vorticity moments as there are for the energy and enstrophy. Several important issues for inviscid vortex dynamics have been discussed in relation to the number of vorticity constraints to keep in any physically relevant statistical theories of 2d flows. Kraichnan [8], Leith [9], Chorin [10], [11], and Majda and Holen [17] have all argued for the sufficiency of the first two in equilibrium statistical mechanics, namely total circulation and enstrophy, in addition to energy. Nonetheless, there are subtle issues concerning the role of higher order vorticity moments in the variational theory of the Barotropic vorticity equation, which are discussed in a recent work [30].

Our third and most important reason for using a statistical mechanics model with only the energy and enstrophy constraints is the simple fact that we are interested in the robustness of truly global energy maximizers which break $SO(3)$ symmetry. Adding more constraints to the canonical probability measure will increase the number of terms and Lagrange multipliers in the corresponding enthalpy functional. Likewise, adding more microcanonical constraints is equivalent to the more constrained variational problem of finding energy extrema on the intersection of more than one manifold in phase space, which must necessarily yield suboptimal extrema instead of the global energy maximizer we are looking at.

1.3. Summary of content. We discuss the relationships between energy, enstrophy and total circulation of inviscid vortex dynamics on a nonrotating sphere in a deterministic variational setting to show that unique energy maximizers are related to negative temperatures in a family of lattice spin models for energy-enstrophy theory on a sphere. To show that symmetry breaking solutions are robust in a physically meaningful sense, we use this equilibrium statistical mechanics energy-enstrophy model to prove that the energy maximizers are in fact most probable macrostates

in a Gibbs ensemble. For this argument to work, negative temperatures must be supported in this model. The rigorous proof of negative temperatures in the spherical model for energy-enstrophy theory is given in exact solutions of the spherical model [13], [12], [14]. The main results on which these applications are based are the existence and uniqueness of constrained energy maximizers on the intersection of iso-enstrophy and iso-circulation manifolds. Finally, we end with a discussion of the relationship between the constrained energy maximizer (shown to exist in this paper) and the maximizer of the free energy.

2. Inviscid vortex dynamics on a nonrotating sphere. Without loss of generality we will discuss only the case of flows on the unit sphere S^2. Inviscid flows on the nonrotating sphere are governed by the equation

$$\frac{D}{Dt}w = 0, \tag{2.1}$$

where w is the absolute vorticity, which further satisfies the zero total circulation condition

$$K[w] = \int_{S^2} w dx = 0.$$

The Casimir symmetries of (2.1) imply that in addition to the kinetic energy of flow (from time invariance)

$$H[w] = \frac{1}{4}\int_{S^2} dx w(x) \int_{S^2} dx' w(x') \ln \frac{1}{|x - x'|} \tag{2.2}$$

and the usual momenta (from $SO(3)$ symmetry), the inviscid vortex dynamics governed by (2.1) conserves an infinite number of vorticity moments [10]. The first two of these moments are total circulation $K[w]$ and enstrophy

$$\Gamma[w] = \int_{S^2} w^2 dx.$$

We recall that any spherical harmonic Y_{lm} (that is, any eigenfunction of the Laplace–Beltrami operator on the unit sphere) is a steady solution of the equation of motion. Thus, any zonally symmetric harmonic such as $Y_{10} = k\cos\theta$, where θ is the co-latitude, is a steady solution of (2.1). These are symmetry breaking solutions. In other words, instead of the full $SO(3)$ symmetry of (2.1), the zonal harmonics Y_{l0} are only $SO(2)$ symmetric.

Recall that for any given zonal basic solution in Shepherd's proof, there is a small enough rotation rate Ω of the sphere which results in the vanishing of the meridional gradient of the total vorticity $Q_{(\cos\theta)}$ somewhere on the sphere. This spoils the upper bound in the Lyapunov stability theorem, which so effectively controlled the growth of the enstrophy of the disturbance vorticity q in the problem of the strongly rotating sphere [6]. Thus Shepherd's proof does not work here since the rate of rotation $\Omega = 0$.

Without results as strong as those in [6], a generic stability analysis of these zonal basic states within the framework of high-dimensional Hamiltonian dynamics reveals that there are nearby solutions that escape to infinity via the well-known phenomena of Arnold's whiskered tori and Nekhoroshev's theory [28]. There is, as far as we know, no proof of nonergodicity in the case of the nonrotating sphere. That is not to say that there is a known proof of ergodicity, which is a notoriously difficult result to

obtain. Thus, without any definite evidence to the contrary, we bow to the wisdom of physicists and assume that inviscid vortex dynamics on the surface of the nonrotating sphere is ergodic [10]. We will construct an equilibrium statistical mechanics theory for this problem. But before we do that, we will formulate a pair of optimization problems related to the definition of the statistical theory and to symmetry breaking in vortex dynamics on a sphere.

3. Variational analysis. We will show that kinetic energy, enstrophy, and total circulation conservation yield a dual pair of related optimization problems for inviscid vortex flows on the nonrotating sphere. The first problem is to maximize the energy H for fixed enstrophy Γ; i.e.,

$$\max H[w] = H_{\max}(\Gamma) < \infty,$$
$$w \in W(\Gamma), \quad K[w] = \int_{S^2} w dx = 0,$$

where $W(\Gamma)$ consists of all vorticity distributions w in $L^2(S^2)$, which has fixed L^2 norm or enstrophy. The dual problem is to minimize the enstrophy for fixed energy. Existence of a unique energy maximizer $w_0(\Gamma)$ such that $H[w_0] = H_{\max}(\Gamma) < \infty$ is an important result with subsequent applications in constructing a well-defined statistical mechanics theory.

3.1. Upper bound and energy maximizers. The kinetic energy $H[w_0]$ of a purely solid-body rotation vorticity distribution

$$w_0 = k(\Theta)\cos\theta$$

(at spin rate Θ, with θ denoting the co-latitude and assuming unit fluid density on the unit sphere S^2) is given by

$$H[w_0] = 4\Theta^2 \int_{S^2} \sin^2\theta \, dx. \tag{3.1}$$

The enstrophy of the same vorticity distribution w_0 is

$$\Gamma = 4\Theta^2 \int_{S^2} \cos^2\theta \, dx. \tag{3.2}$$

Rayleigh's (or Poincaré's) variational inequality [26] implies that

$$\frac{H[w]}{\Gamma[w]} = \frac{\frac{1}{2}\langle w, G(w)\rangle}{\langle w, w\rangle} \le D^{-1},$$

where

$$G(w)(x) = \frac{1}{2}\int_{S^2} w(x') \ln \frac{1}{|x - x'|} dx' \tag{3.3}$$

is the integral operator inverse to $-\Delta$ on S^2 with $K[w] = 0$ and D is the smallest positive eigenvalue of $-\Delta$ on S^2. The Dirichlet quotient is thus bounded below by

$$Q[w] = \frac{\Gamma[w]}{H[w]} = \frac{\int_{S^2} w^2 dx}{\frac{1}{2}\langle w, G(w)\rangle} \ge D.$$

Since the ratio

$$\Gamma/H(w_0(\Gamma)) = \frac{\int_{S^2} \cos^2\theta \, dx}{\int_{S^2} \sin^2\theta \, dx} = D$$

is a universal constant that does not depend on Γ, $w_0(\Gamma)$ is an energy maximizer for any finite enstrophy Γ; that is,

$$H[w_0] = H_{\max}(\Gamma).$$

The maximum value of the kinetic energy $H_{\max}(\Gamma)$ is therefore given by the above simple calculation to be

$$H_{\max}(\Gamma) = D\Gamma.$$

3.2. Concavity and uniqueness. Next we show that $w_0(\Gamma)$ is the unique zonally symmetric energy maximizer modulo the group $SO(3)$. That is, if $H[w'] = H_{\max}(\Gamma) = D\Gamma$, $w' \in W(\Gamma)$, $K[w'] = 0$, and w' is $SO(2)$ symmetric, then $w' = \gamma w_0$ for some $\gamma \in SO(3)$. Uniqueness of this type can be obtained from the observations that any $w \in L^2(S^2)$ can be decomposed into a linear combination of spherical harmonics Y_{lm}; the value $1/D = l(l+1) = 2$ (with $l = 1$) is the smallest positive eigenvalue of the Laplace–Beltrami operator on S^2; and the eigenspace associated with $1/D$ has dimension $2l + 1 = 3$ (with $l = 1$) modulo $SO(3)$, that is, any eigenfunction in this subspace is equal to γw_0, γw_1, or γw_{-1} for some $\gamma \in SO(3)$. Here w_0, w_1, and w_{-1} correspond to the spherical harmonics Y_{10}, Y_{11}, and $Y_{1,-1}$, respectively. Only w_0 is zonally or $SO(2)$ symmetric.

We will prove a stronger result from which uniqueness follows, namely, the strict concavity of the augmented energy functional

$$H'[w(x)] = H[w(x)] - D\Gamma[w] \tag{3.4}$$

on a convex subset; that is,

$$H'[\lambda p + (1-\lambda)q] > \lambda H'[p] + (1-\lambda)H'[q] \tag{3.5}$$

for any $\lambda \in (0,1)$ and for p and q in $M(\Gamma)$ such that

$$\begin{gathered} (1)\ K[p] = K[q] = 0, \\ (2)\ p - q \text{ is not identically zero, and} \\ (3) \text{ only one of } p \text{ or } q \text{ is in the } SO(3) \text{ orbit of } \operatorname{span}(Y_{10}, Y_{11}, Y_{1,-1}). \end{gathered} \tag{3.6}$$

It is well known (from Ekeland and Temam [27]) that the maximizer set

$$M(\Gamma) = \{w \in W(\Gamma) \mid K(w) = 0 \text{ and } H[w] = H_{\max}(\Gamma)\}$$

of the first optimization problem is a convex set if it is nonempty. The existence of the maximizer $w_0(\Gamma)$ means that this set $M(\Gamma)$ is nonempty and, thus, a convex set for any enstrophy $\Gamma < \infty$. The convexity of $M(\Gamma)$ allows the argument in the next paragraph to be carried out. The convexity of the set $W(\Gamma)$ is needed in the usual calculus of variations framework to prove the existence of energy maximizers. Fortunately for us, the existence of this maximizer was derived above by other means. The set $W(\Gamma)$ is in fact not convex.

Let w' be a different energy maximizer than w_0 that has zero total circulation and the same enstrophy Γ as w_0 and satisfies the third condition in (3.6) in the sense that w' is not in the $SO(3)$ orbit of $\mathrm{span}(w_0, w_1, w_{-1})$. Let

$$w^* = \lambda w' + (1-\lambda)w_0$$

for $\lambda \in (0,1)$. Then $H'[w^*] < 0$ because w^* is neither identically zero for any $\lambda \in (0,1)$ nor in the eigenspace containing Y_{10}. This contradicts the strict concavity theorem (3.5) which states that

$$H'[w^*] > \lambda H'[w'] + (1-\lambda)H'[w_0] = 0,$$

where the equality comes from the fact that both w_0 and w' are energy maximizers in $M(\Gamma)$, i.e.,

$$H'[w_0] = H'[w'] = 0.$$

Thus we have proven the uniqueness result.

3.3. Proof of the strict concavity of $H'[w]$. From the definition of $H'[w]$ it follows that

$$\begin{aligned}
H'[\lambda p + (1-\lambda)q] &= \frac{1}{2}\int_{S^2} d\vec{x} \int_{S^2} d\vec{x'} \\
&\quad \times [\lambda^2 p(\vec{x})p(\vec{x'}) + \lambda(1-\lambda)p(\vec{x})q(\vec{x'}) \\
&\qquad + \lambda(1-\lambda)q(\vec{x})p(\vec{x'}) + (1-\lambda)^2 q(\vec{x})q(\vec{x'})] \ln \frac{1}{|\vec{x}-\vec{x'}|} \\
&\quad - 2D\Gamma[\lambda p + (1-\lambda)q] \\
&= \frac{1}{2}\lambda^2 \int_{S^2} d\vec{x} \int_{S^2} d\vec{x'} p(\vec{x})p(\vec{x'}) \ln \frac{1}{|\vec{x}-\vec{x'}|} \\
&\quad + \frac{1}{2}(1-\lambda)^2 \int_{S^2} d\vec{x} \int_{S^2} d\vec{x'} q(\vec{x})q(\vec{x'}) \ln \frac{1}{|\vec{x}-\vec{x'}|} \\
&\quad + \lambda(1-\lambda) \int_{S^2} d\vec{x} \int_{S^2} d\vec{x'} q(\vec{x'})p(\vec{x}) \ln \frac{1}{|\vec{x}-\vec{x'}|} \\
&\quad - 2D \int_{S^2} (\lambda^2 p^2(\vec{x}) + 2\lambda(1-\lambda)q(\vec{x})p(\vec{x}) + (1-\lambda)^2 q^2(\vec{x}))d\vec{x}.
\end{aligned}$$

Defining

$$\begin{aligned}
G[p,q] &= \lambda(1-\lambda) \int_{S^2} d\vec{x} \int_{S^2} d\vec{x'} q(\vec{x'})p(\vec{x}) \ln \frac{1}{|\vec{x}-\vec{x'}|} \\
&\quad - 4D\lambda(1-\lambda) \int_{S^2} q(\vec{x})p(\vec{x})d\vec{x},
\end{aligned}$$

as in another paper by Lim and Zhu [25], we will prove

$$\lambda^2 H'[p] + (1-\lambda)^2 H'[q] + G[p,q] > \lambda H'[p] + (1-\lambda)H'[q], \tag{3.7}$$

which is equivalent to

$$\begin{aligned}
H'[p] + H'[q] &< \int_{S^2} d\vec{x} \int_{S^2} d\vec{x'} q(\vec{x'})p(\vec{x}) \ln \frac{1}{|\vec{x}-\vec{x'}|} \\
&\quad - 4D \int_{S^2} q(\vec{x})p(\vec{x})d\vec{x}.
\end{aligned}$$

By splitting the term on the right into two equal pieces, we see that we will need to prove

$$\begin{aligned}
&\frac{1}{2}\int_{S^2} d\vec{x}\int_{S^2} d\vec{x'}p(\vec{x})p(\vec{x'})\ln\frac{1}{|\vec{x}-\vec{x'}|} - 2D\int_{S^2} p^2(\vec{x})d\vec{x}\\
&\quad+\frac{1}{2}\int_{S^2} d\vec{x}\int_{S^2} d\vec{x'}q(\vec{x})q(\vec{x'})\ln\frac{1}{|\vec{x}-\vec{x'}|} - 2D\int_{S^2} q^2(\vec{x})d\vec{x}\\
&\quad-\frac{1}{2}\int_{S^2} d\vec{x}\int_{S^2} d\vec{x'}q(\vec{x'})p(\vec{x})\ln\frac{1}{|\vec{x}-\vec{x'}|} + 2D\int_{S^2} q(\vec{x})p(\vec{x})d\vec{x}\\
&\quad-\frac{1}{2}\int_{S^2} d\vec{x}\int_{S^2} d\vec{x'}q(\vec{x'})p(\vec{x})\ln\frac{1}{|\vec{x}-\vec{x'}|} + 2D\int_{S^2} q(\vec{x})p(\vec{x})d\vec{x}\\
&< 0.
\end{aligned}$$

The left side of the inequality is just $H'[p-q]$, as seen in the following rearrangement:

$$\begin{aligned}
&\frac{1}{2}\int_{S^2} d\vec{x}\int_{S^2} d\vec{x'}p(\vec{x'})(p(\vec{x})-q(\vec{x}))\ln\frac{1}{|\vec{x}-\vec{x'}|}\\
&\quad+\frac{1}{2}\int_{S^2} d\vec{x}\int_{S^2} d\vec{x'}q(\vec{x'})(q(\vec{x})-p(\vec{x}))\ln\frac{1}{|\vec{x}-\vec{x'}|}\\
&\quad-2D\int_{S^2} d\vec{x}p(\vec{x})(p(\vec{x})-q(\vec{x})) + 2D\int_{S^2} d\vec{x}q(\vec{x})(p(\vec{x})-q(\vec{x}))\\
&=\frac{1}{2}\int_{S^2} d\vec{x}\int_{S^2} d\vec{x'}(p(\vec{x})-q(\vec{x}))(p(\vec{x'})-q(\vec{x'}))\ln\frac{1}{|\vec{x}-\vec{x'}|}\\
&\quad-2D\int_{S^2} d\vec{x}(p(\vec{x})-q(\vec{x}))^2\\
&=H'[p-q].
\end{aligned}$$

In the existence of the upper bound $H_{\max}[\Gamma] = D\Gamma$, we have shown that $H'[w] \le 0$ for any nonzero $w \in W(\Gamma)$ such that $K[w] = 0$. Thus $H'[p-q] \le 0$. But for p and q satisfying the conditions in (3.6), we have

$$H'[p-q] < 0$$

because $p-q$ is not identically zero and is not in the eigenspace containing Y_{10}. The proof of the strict concavity of $H'[w]$ is now complete. The weaker statement of concavity follows from the existence of the upper bound $H'[p-q] \le 0$.

This completes the proof that for a fixed finite enstrophy, the kinetic energy $H(w)$ is bounded above by a finite positive value, i.e.,

$$H(w) \le H_{\max}(\Gamma) = D\Gamma < \infty.$$

Moreover the energy maximum is achieved by a unique zonal vorticity distribution w_0 modulo the group of rotations $SO(3)$.

4. Statistical mechanical proof of robust symmetry breaking. Next, we show that the solid-body rotation state w_0 that maximizes the energy $H[w]$ (along with the remaining spherical harmonics w_1 and w_{-1} in the first eigenspace) is robust in the sense that the most probable macrostate in an equilibrium statistical mechanics theory has overwhelming probability relative to other allowed macrostates. For this

we need a physically meaningful statistical mechanics energy-enstrophy theory of 2d inviscid flows on the nonrotating sphere, in which the solid rotating state w_0 arises naturally as its most probable macrostate.

The discussion on the relative merits of the spherical model in the introduction to this paper centers on the minimum required conditions in a statistical model for demonstrating the robustness of truly global energy maximizers such as the zonally symmetric states. A further necessary condition for any equilibrium statistical mechanics theory to be able to support such a most probable state w_0, which is, moreover, an energy maximizer for fixed enstrophy, is that negative inverse temperatures are allowed in this theory. This is because the Gibbs probability measure has the form

$$P(w) = \frac{e^{-\beta H[w]}\delta(\Gamma[w] - \Gamma')}{Z(\beta, \Gamma)}$$

for $w \in W(\Gamma)$, where β denotes the inverse temperature.

The spherical model energy-enstrophy theory with zero total circulation [13] is such a theory. It predicts that for negative inverse temperatures $\beta < 0$, there is a unique most probable macrostate (solid-body rotation)

$$w_0 = k \cos\theta = cY_{10} \tag{4.1}$$

which maximizes the kinetic energy $H[w] = H_{\max}(\Gamma)$ under the microcanonical constraint of fixed total enstrophy $\Gamma[w] = \Gamma > 0$. This can be seen by solving for the explicit form of its partition function. The somewhat lengthy derivation of the closed-form partition function from the expression

$$Z(\beta, \Gamma) = \int_{W(\Gamma)} e^{-\beta H[w]}\delta(\Gamma - \Gamma^*)dw \tag{4.2}$$

can be found in [13], [12], [14] and will not be repeated here. The value of $k = k(\Gamma)$ in (4.1) depends on the fixed value of the enstrophy Γ, which means that the rate of rotation $\Theta(w_0)$ of the solid-body rotation flow w_0 depends on the enstrophy Γ, i.e.,

$$\Theta^2(w_0) = \frac{\Gamma}{4\int_{S^2} \cos^2\theta \, dx}.$$

4.1. Derivation of the spherical model. The name spherical model comes from the fact that in a lattice approximation, the conservation of enstrophy takes the form of a hyperspherical constraint in vorticity phase space (4.3), as shown next. Using a uniform mesh M of N points $\{x_1, \ldots, x_N\}$ on S^2 and the Voronoi cells based on this mesh, approximate the vorticity by

$$w(x) \simeq \sum_{j=1}^{N} s_j H_j(x),$$

where $s_j = w(x_j)$ and H_j is the indicator function on the Voronoi cell D_j centered at x_j [29]. The truncated kinetic energy of flow in (2.2) now takes the standard form of a spin lattice model Hamiltonian,

$$H_N = -\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N} J_{jk} s_j s_k,$$

where

$$J_{jk} = \int_{S^2} dx H_j(x) \int_{S^2} dx' \ln|1 - x \cdot x'| H_k(x')$$
$$\simeq \frac{16\pi^2}{N^2} \ln|1 - x_j \cdot x_k| \quad \text{for } N \gg 1.$$

The truncated relative enstrophy is likewise given by

$$\Gamma_N = \int_{S^2} dx w^2 = \int_{S^2} dx \left(\sum_{j=1}^{N} s_j H_j(x) \right)^2 = \frac{4\pi}{N} \sum_{j=1}^{N} s_j^2. \tag{4.3}$$

Last, the truncated total circulation is given by

$$TC_N = \int_{S^2} dx w = \int_{S^2} dx \sum_{j=1}^{N} s_j H_j(x) = \frac{4\pi}{N} \sum_{j=1}^{N} s_j.$$

Thus, the microcanonical constraint $\delta(\Gamma - \Gamma^*)$ in (4.2) is approximated in this lattice formulation by a hyperspherical constraint,

$$\sum_{j=1}^{N} s_j^2 = \frac{N}{4\pi} \Gamma^*.$$

The canonical constraint on kinetic energy $H[w]$ should be taken to mean that in a numerical simulation of its equilibrium statistics, one sets the value of the inverse temperature of an infinite energy reservoir and allows the energy to flow between the vortex system and this reservoir and, thence, towards equilibrium. On the other hand, the microcanonical constraint on the enstrophy implies that in the same simulation, the value of the enstrophy is held fixed while the energy flows between the vortex system and the energy reservoir. Total circulation is held fixed at zero throughout such simulations (cf. Lim and Nebus [29]). As the simulation proceeds, the value of the energy changes and finally fluctuates around its equilibrium value $\langle H \rangle(\beta, \Gamma^*)$. Similarly, all other conserved quantities of the Euler equations on a sphere that are not explicitly built into this model, including the angular momentum of the fluid, will in general fluctuate as the simulation runs.

Indeed our experience shows Monte Carlo simulations of the spherical model have the following robust behavior: the angular momentum which is often set to zero initially changes significantly and then settles on a nonzero value given by

$$\Lambda[u] = \int_{S^2} u(\theta) \sin^2\theta d\theta d\phi,$$

where $u(\theta)$ is the zonal velocity of the solid-body rotation vorticity state $w_0(\Gamma^*)$ and θ is the co-latitude on the unit sphere S^2. Clearly, the angular momentum (per unit density) $\Lambda[u] = f(\Gamma^*)$ depends on the enstrophy Γ^* or equivalently the energy $H = 2D\Gamma^*$. From the point of view of globally maximizing the kinetic energy, the addition of a fixed angular momentum constraint into the statistical mechanics formulation must lead to undesired local energy maximizers.

This completes our demonstration that symmetry breaking of the $SO(3)$ symmetry by global energy maximizers in the problem of nonrotating spheres is robust

in a statistical sense. The same arguments can be used to prove the robustness of other symmetry breaking spherical harmonic vortex states Y_{lm} on the nonrotating sphere. The principle of selective decay then implies that the asymptotic vorticity of decaying 2d Navier–Stokes flows on S^2 is the ground state Y_{10} found, thus providing an important connection between the results on ideal flows and more realistic flows with nonzero viscosity.

5. Statistical thermodynamics of ideal flows on a sphere. We end with a discussion of the relationship between the constrained energy maximizer and the maximizer of the free energy. Since the work of Planck, it is well known that the most probable state in the statistics of the partition function (1.1) represents a minimum of the free energy $F = E - TS$, where E is the internal energy, $T = \beta^{-1}$ is the temperature, and S is the entropy. Dynamic equilibria of the underlying dynamical system, on the other hand, are related to the extremals of the energy functional E. In recent work [24], [23] on the Onsager vortex gas problem on the unbounded plane, we have asked the question, under what conditions are the minima of F well approximated by the energy extremals of E? It turns out that for low positive temperatures, this approximation is excellent, even for relatively small numbers M of point vortices, and for order one temperatures.

We first use the exact solution of the spherical model for energy-enstrophy theory to confirm the validity of this approximation in situations with negative temperatures and in systems that are based, not on a particle method like the Onsager problem, but on a lattice or spatial discretization like the family of spherical models in this paper. From the expression $F = E - TS$, one should expect from general principles that if

$$T < 0, \qquad |T| \to 0, \tag{5.1}$$

and

$$\begin{aligned} &\langle E \rangle \text{ is not small compared to } S \text{ when both are} \\ &\text{evaluated at the most probable state } w_0(T), \end{aligned} \tag{5.2}$$

then a similar approximation is again valid:

(1) The maximizers w_0 of F are the most probable states for $T < 0$;

(2) These maximizers are close to the maximizers w_0' of E.

We show that for the spherical model of ideal fluid flows on a nonrotating sphere, this approximation is actually valid for all negative temperatures, but for reasons that are different than (5.1) and (5.2). The exact solution of the spherical model for energy-enstrophy theory has the important consequence that for any negative T, large or small in numerical value, the expected value $\langle E \rangle$ of the energy has a maximal ratio to the given value of the enstrophy Γ. This ratio is given by the reciprocal of the smallest positive eigenvalue of the Laplace–Beltrami operator on the sphere S^2. This is known as Rayleigh's inequality and is related to Poincaré's inequality, and the corresponding eigenfunction is the spherical harmonic $Y_{10} = d\cos\theta$ with θ equal to the co-latitude on S^2. Thus, for all $T < 0$, E is maximally related to the enstrophy Γ.

Furthermore, the entropy

$$S(\vec{s}(N)) = k \log R(\vec{s}(N)),$$

where $R(\vec{s}(N))$ is the number of rearrangements of vorticity parcels between lattice sites corresponding to a particular spin state $\vec{s}(N) = (s_1, \ldots, s_N)$. Since, in the

continuum limit,

$$\vec{s}(N) \to w(\theta, \phi) \in L_2(S^2),$$

the family of spin states $\{\vec{s}_0(N)\}_1^\infty$ which has minimal entropy $S_N = S(\vec{s}_0(N))$ for each N, corresponds to a most probable vorticity distribution $w_0(\theta, \phi) \in L_2(S^2)$. Under generic conditions, the most probable state w_0 corresponds to the uniform function which is the spherical harmonic ψ_{00} on the sphere (cf. Lieb and Loss [5]). But in this particular problem, under the additional condition $\int_{S^2} w_0 dx = 0$ of zero total circulation which comes from Stokes's theorem, the most probable vorticity is not $d\psi_{00}$, and it may not be possible to use the method of rearrangement in [5] to find the most probable vorticity w_0. Nonetheless, the exact solution of the spherical model's partition functions Z_N implies that the most probable spin states

$$\vec{s}_0(N) \to w_0(\theta, \phi) = d\psi_{10} \in L_2(S^2)$$

for all $T < 0$. Thus, in the continuum limit, the unique most probable vorticity function $w_0 = d\psi_{10}$ is the solid-body rotation state for all negative temperatures.

Therefore, the same vorticity function $d\psi_{10}$ simultaneously maximizes the energy E and minimizes the entropy S for all $T < 0$. This certainly proves that the maximizer $w_0(T)$ of F for $T < 0$ is exactly equal to the maximizer $w_o'(T)$ of the energy E, but we had to use the exact solution of the spherical model to establish the validity of the approximation.

In problems where the validity of this approximation can be established by a priori means (that is, without solving the full partition function), one can in principle compute the most probable state $w_0(T)$, that is, maximizers of the free energy F, at $T < 0$ by performing the easier task of computing the maximizer $w_0'(T)$ of the energy E. The natural question to ask at this point is whether one can establish the validity of this approximation without using the exact solution of the partition function (1.1). It is not clear this can be done.

Future work will include further attempts to give a priori derivations of the above approximation, and the extension of the results in this paper to the barotropic vorticity equation on a rotating sphere, where the variational theory is richer and exhibits interesting bifurcations when the rate of rotation of the sphere changes relative to the total enstrophy.

Acknowledgments. The presentation of this paper benefited from the comments of two anonymous reviewers. The author would also like the acknowledge the scientific support of Dr. Gary Johnson and Dr. Robert Launer.

REFERENCES

[1] C. C. Lim, *Phase transitions and coherent structures in an energy-enstrophy theory for axisymmetric flows*, Phys. Fluids, 15 (2003), pp. 478–487.
[2] D. Montgomery and G. Joyce, *Statistical mechanics of "negative temperature" states*, Phys. Fluids, 19 (1974), pp. 1139–1145.
[3] L. Onsager, *Statistical hydrodynamics*, Nuovo Cimento Suppl., 6 (1949), pp. 279–289.
[4] G. L. Eyink and H. Spohn, *Negative-temperature states and large-scale, long-lived vortices in two-dimensional turbulence*, J. Statist. Phys., 70 (1993), pp. 833–886.
[5] E. Lieb and M. Loss, *Analysis*, 2nd ed., Grad. Stud. Math. 14, AMS, Providence, RI, 2001.
[6] T. G. Shepherd, *Non-ergodicity of inviscid two-dimensional flow on a beta-plane and on the surface of a rotating sphere*, J. Fluid Mech., 184 (1987), pp. 289–302.

[7] C. Foias and J.-C. Saut, *Asymptotic behavior, as $t \to +\infty$, of solutions of Navier-Stokes equations and nonlinear spectral manifolds*, Indiana Univ. Math. J., 33 (1984), pp. 459–477.
[8] R. H. Kraichnan, *Statistical dynamics of two-dimensional flows*, J. Fluid Mech., 67 (1975), pp. 155–175.
[9] C. Leith, *Diffusion approximation for two-dimensional turbulence*, Phys. Fluids, 11 (1968), pp. 671–673.
[10] A. Chorin, *Vorticity and Turbulence*, Springer-Verlag, New York, 1993.
[11] A. J. Chorin, *Partition functions and equilibrium measures in two-dimensional and quasi-three-dimensional turbulence*, Phys. Fluids, 8 (1996), pp. 2656–2660.
[12] C. C. Lim, *A long range spherical model and exact solutions of an energy-enstrophy theory for two-dimensional turbulence*, Phys. Fluids, 13 (2001), pp. 1961–1973.
[13] C. C. Lim, *Exact solutions of the energy-enstrophy theory for the barotropic vorticity equation on a sphere*, Phys. A, 290 (2001), pp. 131–158.
[14] C. C. Lim, *Coherent structures in an energy-enstrophy theory for axisymmetric flows*, Phys. Fluids, 15 (2003), pp. 478–487.
[15] G. Holloway and M. Hendershott, *Stochastic closure for nonlinear Rossby waves*, J. Fluid Mech., 82 (1977), pp. 747–765.
[16] P. B. Rhines, *Waves and turbulence on a beta plane*, J. Fluid Mech., 69 (1975), pp. 417–443.
[17] A. J. Majda and M. Holen, *Dissipation, topography, and statistical theories of large scale coherent structure*, Comm. Pure Appl. Math., 50 (1997), pp. 1183–1234.
[18] J. Miller, *Statistical mechanics of Euler equations in two dimensions*, Phys. Rev. Lett., 65 (1990), pp. 2137–2140.
[19] R. Robert and J. Sommeria, *Statistical equilibrium states for two-dimensional flows*, J. Fluid Mech., 229 (1991), pp. 291–310.
[20] T. Nozawa and S. Yoden, *Formation of zonal band structure in forced two-dimensional turbulence on a rotating sphere*, Phys. Fluids, 9 (1997), pp. 2081–2093.
[21] B. Turkington, *Statistical equilibrium measures and coherent states in two-dimensional turbulence*, Comm. Pure Appl. Math., 52 (1999), pp. 1–29.
[22] M. DiBattista and A. Majda, *An equilibrium statistical model for the spreading phase of open-ocean convection*, Proc. Natl. Acad. Sci. USA, 96 (1999), pp. 6009–6013.
[23] C. C. Lim and S. M. Assad, *Self containment radius for an electron plasma and point vortices in an unbounded plane*, Regul. Chaotic Dyn., (2005), to appear.
[24] S. M. Assad and C. C. Lim, *Statistical equilibrium of the Coulomb/Vortex gas in the unbounded two-dimensional plane*, Discrete Contin. Dyn. Syst. Ser. B, 5 (2005), pp. 1–14.
[25] C. C. Lim and D. Zhu, *Variational analysis of energy-enstrophy theories on the sphere*, in Proceedings of the 5th International Conference on Dynamic Systems and Differential Equations, Ponoma, CA, 2004, Discrete Contin. Dyn. Syst. Ser. B, Suppl. (2005), to appear.
[26] P. D. Lax, *Functional Analysis*, Wiley-Interscience, New York, 2002.
[27] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.
[28] C. C. Lim, *A combinatorial perturbation method and Arnold whiskered tori in vortex dynamics*, Phys. D, 64 (1993), pp. 163–184.
[29] C. C. Lim and J. Nebus, *The spherical model of logarithmic potentials as examined by Monte Carlo methods*, Phys. Fluids, 16 (2004), pp. 4020–4027.
[30] C. C. Lim and J. Shi, *The role of higher vorticity moments in a variational formulation of the barotropic vorticity model on a rotating sphere*, Arch. Ration. Mech. Anal., submitted.

SIAM J. APPL. MATH.
Vol. 65, No. 6, pp. 2107–2127

RECONSTRUCTION OF A SMALL INCLUSION IN A TWO-DIMENSIONAL OPEN WAVEGUIDE*

HABIB AMMARI†, EKATERINA IAKOVLEVA†, AND HYEONBAE KANG‡

Abstract. We consider wave propagation in a perturbed open waveguide. We provide a new asymptotic expansion for the scattered wave when the inclusion is of small diameter. We design a MUSIC (multiple signal classification) type of algorithm for locating the inclusion and illustrate its viability in numerical examples.

Key words. open waveguide, electromagnetic inclusion, reconstruction, MUSIC algorithm

AMS subject classifications. 35R20, 35B30

DOI. 10.1137/040615389

1. Introduction.

Optical waveguides are the basis of the optoelectronics and telecommunications industry. It is important in engineering design of optical communication devices not only to find out whether a defect is present or absent in a device, but also to precisely locate it and accurately characterize its size.

In this paper we discuss wave propagation in a perturbed optical waveguide. The perturbation in the electromagnetic characteristics of the waveguide is caused by a small electromagnetic inclusion. The waveguide we consider is half space ($y > 0$) with the Dirichlet boundary condition on $y = 0$. The region $0 < y < h$ is considered the core of the fiber, while the remainder is considered the cladding. The electromagnetic characteristics of the waveguide are constant in each part. The electric permittivity and the magnetic permeability are then given by

$$\varepsilon(y) = \begin{cases} \varepsilon_1 & \text{in }]0,h[, \\ \varepsilon_2 & \text{in }]h,+\infty[\end{cases}$$

and

$$\mu(y) = \begin{cases} \mu_1 & \text{in }]0,h[, \\ \mu_2 & \text{in }]h,+\infty[, \end{cases}$$

where $\varepsilon_1\mu_1 \geq \varepsilon_2\mu_2$ and $\mu_1 \neq \mu_2$.

We suppose that there is an electromagnetic inclusion D in the core of the waveguide, of the form $D = Z + \alpha B$, where $B \subset \mathbb{R}^2$ is a bounded, smooth ($\mathcal{C}^\infty$) domain containing the origin. The point $Z = (z_x, z_y) \in \mathbb{R}\times]0,h[$, which determines the location of the inclusion, is assumed to satisfy $h - d_0 \geq z_y \geq d_0 > 0$. The value of α is the order of magnitude of the diameter of the inclusion. Let μ_* and ε_* denote the magnetic permeability and the electric permittivity of the inclusion D; we shall

*Received by the editors September 21, 2004; accepted for publication (in revised form) March 11, 2005; published electronically August 9, 2005. This research was partly supported by ACI Nouvelles Interfaces des Mathématiques grant 171, CNRS-KOSEF grant 14889, and Korea Science and Engineering Foundation grant R02-2003-000-10012-0.

http://www.siam.org/journals/siap/65-6/61538.html

†Centre de Mathématiques Appliquées, CNRS UMR 7641 and Ecole Polytechnique, 91128 Palaiseau Cedex, France (ammari@cmapx.polytechnique.fr, iakov@cmapx.polytechnique.fr).

‡School of Mathematical Sciences, Seoul National University, Seoul 151-747, Republic of Korea (hkang@math.snu.ac.kr).

assume that these are positive constants. Using these notations, we introduce the piecewise constant magnetic permeability

$$\mu_\alpha(x,y) := \begin{cases} \mu_* & \text{in } D, \\ \mu_1 & \text{in } \mathbb{R}\times]0,h[\setminus \bar{D}, \\ \mu_2 & \text{in } \mathbb{R}\times]h,+\infty[\,. \end{cases}$$

If we allow the degenerate case $\alpha = 0$, then

$$\mu_0(x,y) := \begin{cases} \mu_1 & \text{in } \mathbb{R}\times]0,h[, \\ \mu_2 & \text{in } \mathbb{R}\times]h,+\infty[\,. \end{cases}$$

The piecewise constant electric permittivity $\varepsilon_\alpha(x,y)$ is defined analogously.

An incident wave u_0, in the form of a guided mode, is sent along the perturbed waveguide. It encounters the inclusion D in the core region of the waveguide and is scattered. Our first goal in this work is to provide an asymptotic formula for the scattered wave when α goes to zero. Our second goal is to use this expansion for efficiently determining the location and the shape of the inclusion D.

To set the problem mathematically, let u_α satisfy the Helmholtz equation

$$\left(\nabla\cdot\frac{1}{\mu_\alpha}\nabla+\omega^2\varepsilon_\alpha\right)u_\alpha = 0 \quad \text{in } \mathbb{R}\times]0,+\infty[\,, \tag{1}$$

and $u_\alpha - u_0$ satisfy some form of radiation condition. Unfortunately, not much is known about the exact form of this condition due to the fact that the waveguide extends from $-\infty$ to $+\infty$. We avoid this issue by first obtaining a representation of the Green's function of the homogeneous waveguide. The Green's function we give is based on the requirement that waves be outgoing and remain bounded. Using the obtained Green's function, we derive an asymptotic expansion of the solution u_α of the inhomogeneous waveguide problem. We shall mention the work by Zhang and Chandler-Wilde [23], which discusses the issue of radiation conditions for the scattering by an infinite layer. However, these conditions do not rule out the guided waves localized in the layer.

Let us emphasize here that the use of the formal equivalence between electromagnetics and linear acoustics, by term-to-term replacing permittivity and permeability by compressibility and volume density of mass, and the scalar electric field by the scalar acoustic pressure characteristic of compressional waves inside fluid media, opens up the investigation of this paper to many other applications, such as ocean-acoustics, even though the type of materials and of geometrical configurations investigated and the range of values that are allowed to be taken by the two sets of parameters in the two disciplines may differ considerably in practice. The configuration considered in this paper has also been used as a model of underwater acoustics. The area of applications is the identification of mines, submarines, or submerged obstacles in harbors and other shallow bodies of water.

The paper is organized as follows. In section 2, we construct the Green's function corresponding to the unperturbed waveguide. The main ingredient for doing this is an inverse transform formula from [5]. A similar formula was first derived by Magnanini and Santosa [12], [13]. Section 3 is devoted to the derivation of the leading-order term in the asymptotic expansion of the scattered wave. In section 4 we exploit this formula for recovering the location and the shape of the inclusion. A MUSIC (multiple signal classification) type of algorithm is developed for locating the inclusion. Numerical examples are given in section 5. A discussion section ends the paper.

Finally, we shall mention, in connection with our asymptotic expansion for the scattered wave, the nice work by Vogelius and Volkov [20].

2. Green's function for the unperturbed waveguide. This section is devoted to the derivation of an expression of the Green's function. We will separate the Green's function into three components: the guided component, the radiated component, and the evanescent component. We will also provide asymptotic results that show how the nonguided part of the Green's function decays along the core of the waveguide. Our approach for constructing the Green's function follows [12]. We note that one can also employ complex analysis for deriving an explicit representation of the Green's function, starting with the assumption of its separability in the variables x and y, and a representation in terms of a contour integral in the separation parameter; see [9].

For a function f, continuous of compact support, let u satisfy the Helmholtz equation

$$\left(\nabla \cdot \frac{1}{\mu_0}\nabla + \omega^2 \varepsilon_0\right) u = f \quad \text{in } \mathbb{R}^2_+ := \mathbb{R} \times]0, +\infty[\,, \tag{2}$$

with the boundary condition $u = 0$ on $y = 0$.

We introduce the following notation:

$$q(y) = \omega^2(\varepsilon_1\mu_1 - \varepsilon(y)\mu(y)),$$
$$d^2(\omega) = \omega^2(\varepsilon_1\mu_1 - \varepsilon_2\mu_2) \geq 0.$$

Let $g(y, \lambda)$ be defined by

$$\begin{cases} \partial_{yy} g(y,\lambda) + (\lambda - q(y)) g(y,\lambda) = 0 \quad \text{in }]0,h[\cup]h,+\infty[, \\ [g(.,\lambda)] = 0 \quad \text{on } y = h, \\ \left[\dfrac{1}{\mu}\partial_y g(.,\lambda)\right] = 0 \quad \text{on } y = h, \\ g(0,\lambda) = 0 \quad \text{and} \quad \partial_y g(0,\lambda) = \sqrt{\lambda}. \end{cases} \tag{3}$$

Setting $\phi(y, \lambda) = \sin(\sqrt{\lambda}y)$, we then write

$$g(y,\lambda) = \begin{cases} \phi(y,\lambda) \quad \text{if } y \in]0,h[, \\ \phi(h,\lambda)\cos\left[\sqrt{\lambda - d^2}(y-h)\right] + \dfrac{\mu_2}{\mu_1}\dfrac{\partial_y\phi(h,\lambda)}{\sqrt{\lambda - d^2}}\sin\left[\sqrt{\lambda - d^2}(y-h)\right] \\ \text{if } y \in]h,+\infty[. \end{cases}$$

For $\lambda \geq d^2$, $g(y, \lambda)$ is bounded. For $\lambda < d^2$, in view of the above expression of g, we impose the dispersion relation

$$\phi(h,\lambda) + \frac{\mu_2}{\mu_1}\frac{\partial_y\phi(h,\lambda)}{\sqrt{d^2 - \lambda}} = 0,$$

or equivalently,

$$\sqrt{d^2 - \lambda}\tan\sqrt{\lambda}h + \frac{\mu_2}{\mu_1}\sqrt{\lambda} = 0, \tag{4}$$

to make $g(y,\lambda)$ bounded in $\mathbb{R}^+$. It is straightforward to see that there are a finite number of roots $\lambda_l(\omega)$ of (4) with associated solutions: $g(y,\lambda_l)$ for $l=1,2,\dots,m$. Moreover, the set of eigenfunctions $g(y,\lambda), \lambda \in]0,+\infty[$ is complete in $L^2(\mathbb{R}_+)$. When the magnetic permeabilities μ_1 and μ_2 are equal ($\mu_1=\mu_2$), the completeness of the associated eigenvalue problem has been proved and an inverse transform formula has been rigorously derived in [12]. See also [21], [22], where the spectrum of the Pekeris operator is investigated. Here the following more general inverse transform formula from [5] will be needed. Let $f \in L^2(\mathbb{R}_+, \frac{dy}{\mu(y)})$. We have the inverse transform formula

$$
\begin{aligned}
f(x) = &\sum_{l=1}^{m} \frac{2\mu_1\sqrt{d^2-\lambda_l}\int_0^{+\infty} g(y,\lambda_l)f(y)\frac{dy}{\mu(y)}}{\frac{\mu_1}{\mu_2}\phi(h,\lambda_l)^2 + 2\sqrt{d^2-\lambda_l}\int_0^h \phi(y,\lambda_l)^2 dy} g(x,\lambda_l) \\
&+ \frac{1}{\pi}\int_{d^2}^{+\infty} \frac{\mu_2\sqrt{\lambda-d^2}\int_0^{+\infty} g(y,\lambda)f(y)\frac{dy}{\mu(y)}}{(\lambda-d^2)\phi(h,\lambda)^2 + (\frac{\mu_2}{\mu_1})^2\partial_y\phi(h,\lambda)^2} g(x,\lambda)d\lambda
\end{aligned} \tag{5}
$$

almost everywhere.

We now return to the Helmholtz equation (2). Let

$$
U(x,\lambda) = \int_0^{+\infty} u(x,y)g(y,\lambda)\,\frac{dy}{\mu(y)}.
$$

Multiplying (2) by $\frac{1}{\mu(y)}g(y,\lambda)$ and integrating with respect to the variable y over the interval $]0,+\infty[$, after some straightforward manipulations for $x \in \mathbb{R}$ we obtain

$$
\partial_{xx}U(x,\lambda) + (\omega^2\varepsilon_1\mu_1 - \lambda)U(x,\lambda) = \int_0^{+\infty} f(x,\eta)g(\eta,\lambda)\,\frac{d\eta}{\mu(\eta)}. \tag{6}
$$

The solution of (6), which is outgoing for $0 \le \lambda < \omega^2\varepsilon_1\mu_1$ and decays exponentially for $\lambda > \omega^2\varepsilon_1\mu_1$ as $|x| \to +\infty$, is readily given for $x \in \mathbb{R}$ by the following expression:

$$
U(x,\lambda) = \int_{-\infty}^{\infty} \frac{e^{i|x-\zeta|\sqrt{\omega^2\varepsilon_1\mu_1-\lambda}}}{2i\sqrt{\omega^2\varepsilon_1\mu_1-\lambda}} \int_0^{+\infty} f(\zeta,\eta)g(\eta,\lambda)\,\frac{d\eta}{\mu(\eta)}d\zeta. \tag{7}
$$

By the inversion formula (5), we have

$$
\begin{aligned}
u(x,y) = &\sum_{l=1}^{m} \frac{2\mu_1\sqrt{d^2-\lambda_l}U(x,\lambda_l)}{\frac{\mu_1}{\mu_2}\phi(h,\lambda_l)^2 + 2\sqrt{d^2-\lambda_l}\int_0^h \phi(y,\lambda_l)^2 dy} g(y,\lambda_l) \\
&+ \frac{1}{\pi}\int_{d^2}^{+\infty} \frac{\mu_2\sqrt{\lambda-d^2}U(x,\lambda)}{(\lambda-d^2)\phi(h,\lambda)^2 + (\frac{\mu_2}{\mu_1})^2\partial_y\phi(h,\lambda)^2} g(y,\lambda)d\lambda \quad \forall\,(x,y) \in \mathbb{R}^2_+;
\end{aligned}
$$

hence, by (7) and by interchanging the order of integration, we obtain that the solution u of (2) corresponding to the case where no energy is radiated from the far field ($x^2+y^2 \to +\infty, y>0$) can be represented by

$$
u(x,y) = \int_{\mathbb{R}^2_+} G(x,y,\zeta,\eta)f(\zeta,\eta)\,d\zeta d\eta,
$$

where the Green's function G is given by

$$
\begin{aligned}
&G(x,y,\zeta,\eta)\\
&:=\sum_{l=1}^{m}\frac{2\mu_1\sqrt{d^2-\lambda_l}}{\frac{\mu_1}{\mu_2}\phi(h,\lambda_l)^2+2\sqrt{d^2-\lambda_l}\int_0^h\phi(y,\lambda_l)^2dy}\left(\frac{e^{i|x-\zeta|\sqrt{\omega^2\varepsilon_1\mu_1-\lambda_l}}}{2i\sqrt{\omega^2\varepsilon_1\mu_1-\lambda_l}}\right)g(y,\lambda_l)g(\eta,\lambda_l)\\
&+\frac{1}{\pi}\int_{d^2}^{+\infty}\frac{\mu_2\sqrt{\lambda-d^2}}{(\lambda-d^2)\phi(h,\lambda)^2+(\frac{\mu_2}{\mu_1})^2\partial_y\phi(h,\lambda)^2}\left(\frac{e^{i|x-\zeta|\sqrt{\omega^2\varepsilon_1\mu_1-\lambda}}}{2i\sqrt{\omega^2\varepsilon_1\mu_1-\lambda}}\right)g(y,\lambda)g(\eta,\lambda)d\lambda.
\end{aligned}
$$

Note that the Green's function G has been constructed so that all the waves are outgoing.

Following [12], we now separate the Green's function G into three components $G=G^g+G^r+G^e$. The guided component

$$
\begin{aligned}
&G^g(x,y,\zeta,\eta)\\
&:=\sum_{l=1}^{m}\frac{2\mu_1\sqrt{d^2-\lambda_l}}{\frac{\mu_1}{\mu_2}\phi(h,\lambda_l)^2+2\sqrt{d^2-\lambda_l}\int_0^h\phi(y,\lambda_l)^2dy}\left(\frac{e^{i|x-\zeta|\sqrt{\omega^2\varepsilon_1\mu_1-\lambda_l}}}{2i\sqrt{\omega^2\varepsilon_1\mu_1-\lambda_l}}\right)g(y,\lambda_l)g(\eta,\lambda_l)
\end{aligned}
$$

corresponds to the solution that is concentrated near the core. The radiated component

$$
\begin{aligned}
&G^r(x,y,\zeta,\eta)\\
&:=\frac{1}{\pi}\int_{d^2}^{\omega^2\varepsilon_1\mu_1}\frac{\mu_2\sqrt{\lambda-d^2}}{(\lambda-d^2)\phi(h,\lambda)^2+(\frac{\mu_2}{\mu_1})^2\partial_y\phi(h,\lambda)^2}\left(\frac{e^{i|x-\zeta|\sqrt{\omega^2\varepsilon_1\mu_1-\lambda}}}{2i\sqrt{\omega^2\varepsilon_1\mu_1-\lambda}}\right)g(y,\lambda)g(\eta,\lambda)d\lambda
\end{aligned}
$$

and the evanescent component

$$
\begin{aligned}
&G^e(x,y,\zeta,\eta)\\
&:=\frac{1}{\pi}\int_{\omega^2\varepsilon_1\mu_1}^{+\infty}\frac{\mu_2\sqrt{\lambda-d^2}}{(\lambda-d^2)\phi(h,\lambda)^2+(\frac{\mu_2}{\mu_1})^2\partial_y\phi(h,\lambda)^2}\left(\frac{e^{i|x-\zeta|\sqrt{\omega^2\varepsilon_1\mu_1-\lambda}}}{2i\sqrt{\omega^2\varepsilon_1\mu_1-\lambda}}\right)g(y,\lambda)g(\eta,\lambda)d\lambda
\end{aligned}
$$

are radiated away from the source at (ζ,η).

We will need to estimate G^r and G^e for fixed y and η. Following [12] once again, we can apply Laplace's method [6], and obtain for $|x-\zeta|\to+\infty$ that

$$
G^e(x,y,\zeta,\eta)=O\left(\frac{1}{\omega|x-\zeta|}\right). \tag{8}
$$

Moreover, making use of the method of steepest descent [6], we can show that

$$
G^r(x,y,\zeta,\eta)=O\left(\frac{1}{\omega|x-\zeta|}\right)\quad\text{as } |x-\zeta|\to+\infty. \tag{9}
$$

We can therefore conclude that for a fixed y, as one looks down the core of the waveguide, the nonguided components of the waves die off like $O(1/\omega|x|)$.

Let $X=(x,y)$ and $Y=(\zeta,\eta)$. Observe that the Green's function for the problem,

$$
\begin{cases}
\nabla_X\cdot\dfrac{1}{\mu_0}\nabla_X G_0(X,Y)=\delta_Y \quad \text{in } \mathbb{R}\times]0,h[\cup\mathbb{R}\times]h,+\infty[,\\
G_0\big|_+=G_0\big|_-,\quad \dfrac{1}{\mu_2}\dfrac{\partial G_0}{\partial y}\bigg|_+=\dfrac{1}{\mu_1}\dfrac{\partial G_0}{\partial y}\bigg|_- \quad \text{on } y=h,\\
G_0=0 \quad \text{on } y=0,
\end{cases}
$$

is given by the following explicit formula. If $0 < \eta < h$, then

(10)

$$G_0(X,Y) = \mu_1 \begin{cases} \dfrac{2\mu_2}{\mu_1+\mu_2}\big[\Gamma(X-Y)-\Gamma(\overline{X}-Y)\big], & y > h, \\ \big[\Gamma(X-Y)-\Gamma(\overline{X}-Y)\big] + \dfrac{\mu_2-\mu_1}{\mu_1+\mu_2}\big[\Gamma(\overline{X}-Y+(0,2h))-\Gamma(X-Y-(0,2h))\big], \\ \qquad 0 < y < h. \end{cases}$$

Here $\Gamma(X) = (1/(2\pi)) \log|X|$ is the fundamental solution for the Laplacian and $\overline{X} = (x,-y)$. If $\eta > h$, the formula takes the form

(11)

$$G_0(X,Y) = \mu_2 \times \begin{cases} \big[\Gamma(X-Y)-\Gamma(\overline{X}-Y)\big] + \dfrac{\mu_1-\mu_2}{\mu_1+\mu_2}\big[\Gamma(\overline{X}-Y+(0,2h))-\Gamma(X-Y-(0,2h))\big], \\ \qquad y > h, \\ \dfrac{2\mu_1}{\mu_1+\mu_2}\big[\Gamma(X-Y)-\Gamma(\overline{X}-Y)\big], \quad 0 < y < h. \end{cases}$$

We will need the following lemma.

LEMMA 2.1. *For each M and a fixed but arbitrary (ζ,η) with $0 < \eta < h$,*

$$R(x,y,\zeta,\eta) := G(x,y,\zeta,\eta) - G_0(x,y,\zeta,\eta) \tag{12}$$

is $\mathcal{C}^1$ in (x,y) for $|x-\zeta| \le M$ and $0 \le y \le h$, and its $\mathcal{C}^1$-norm is bounded independently of (ζ,η).

Proof. Fix (ζ,η) and let $v(x,y) := G(x,y,\zeta,\eta)$ and $w(x,y) := G_0(x,y,\zeta,\eta)$. Choose $M > 0$ so that on the domain $\Omega_M :=]\zeta - M, \zeta + M[\times]0,h[$ the problem

$$\begin{cases} \left(\nabla \cdot \dfrac{1}{\mu_0}\nabla + \omega^2\varepsilon_0\right) u = 0 & \text{in } \Omega_M, \\ u = f & \text{on } \partial\Omega_M \end{cases}$$

is well posed. Since $(\nabla \cdot \frac{1}{\mu_0}\nabla + \omega^2\varepsilon_0)v = \delta_{(\zeta,\eta)}$ and $\nabla \cdot \frac{1}{\mu_0}\nabla w = \delta_{(\zeta,\eta)}$, the function R given by (12) satisfies

$$\left(\nabla \cdot \frac{1}{\mu_0}\nabla + \omega^2\varepsilon_0\right) R = -\omega^2\varepsilon_0 w \quad \text{in } \Omega_M.$$

Moreover, $R|_{\partial\Omega_M}$ is a piecewise $\mathcal{C}^1$-function, and $R(x,y) = 0$ if $y = 0$. Define

$$W(x,y) := -\omega^2\varepsilon_0 \int_{\Omega_M} G_0(x,y,\zeta,\eta) w(\zeta,\eta) dA.$$

Then one can easily see from the explicit forms (10) and (11) of G_0 that W is $\mathcal{C}^1$ on Ω_M and $\|W\|_{\mathcal{C}^1(\Omega_M)} \le C$ uniformly in (ζ,η). Observe that $R - W$ satisfies

$$\left(\nabla \cdot \frac{1}{\mu_0}\nabla + \omega^2\varepsilon_0\right)(R - W) = -\omega^2\varepsilon_0 W \quad \text{in } \Omega_M,$$

and hence by the standard regularity theorem for the elliptic equations we get

$$\|R - W\|_{\mathcal{C}^1(\Omega_{M/2})} \leq C$$

for some C uniformly in (ζ, η). This completes the proof. □

3. Asymptotic expansion of the scattered wave. In this section we derive an asymptotic formula for the perturbation $u_\alpha - u_0$ due to the presence of the inclusion $D = Z + \alpha B$ as α tends to 0.

For $k > 0$, let the fundamental solution Γ_k be defined by

$$\Gamma_k(X) = -\left(\frac{i}{4}\right) H_0^{(1)}(k|X|) \quad \text{for } X \neq 0,$$

where $H_0^{(1)}$ is the Hankel function of the first kind of order 0. For a bounded smooth domain D in $\mathbb{R}^2$, let $\mathcal{S}_D^k$ be the single layer potential defined by Γ_k; that is, for $\phi \in L^2(\partial D)$,

$$\mathcal{S}_D^k \phi(X) = \int_{\partial D} \Gamma_k(X - Y)\phi(Y) d\sigma(Y), \quad X \in \mathbb{R}^2.$$

Let $\widetilde{\mathcal{S}}_D$ be the single layer potential defined by G; that is, for $\psi \in L^2(\partial D)$,

$$\widetilde{\mathcal{S}}_D \psi(X) = \frac{1}{\mu_1} \int_{\partial D} G(X, Y)\psi(Y) d\sigma(Y), \quad X \in \mathbb{R}^2.$$

Suppose that the following assumption (H1) holds: the trivial solution is the unique solution to the Helmholtz equation

$$\left(\nabla \cdot \frac{1}{\mu_\alpha} \nabla + \omega^2 \varepsilon_\alpha\right) u = 0 \quad \text{in } \mathbb{R}^2_+,$$

with the boundary condition $u = 0$ on $y = 0$ and the decay estimates

$$\left| u_l(x) \mp i\beta_l \frac{du_l}{dx}(x) \right| = O\left(\frac{1}{|x|}\right) \quad \text{as } x \to \pm\infty,$$

for $l = 1, \ldots, m$, where $u_l(x) = \int_0^{+\infty} u(x, y) g(y, \lambda_l)\, dy$.

Following [2], an integral representation formula for the outgoing solution u_α of (1) can be proved.

LEMMA 3.1. *Suppose that $\omega\sqrt{\varepsilon_1 \mu_1}$ is not a Dirichlet eigenvalue of $-\Delta$ on D, and let $k_* := \omega\sqrt{\varepsilon_* \mu_*}$. The solution u_α of (1) can be represented by*

$$u_\alpha(X) = \begin{cases} u_0(X) + \widetilde{\mathcal{S}}_D \psi(X), & X \in \mathbb{R}^2_+ \setminus \overline{D}, \\ \mathcal{S}_D^{k_*} \phi(X), & X \in D, \end{cases} \tag{13}$$

where the pair $(\phi, \psi) \in L^2(\partial D) \times L^2(\partial D)$ is the unique solution to the system of integral equations

$$\begin{cases} \mathcal{S}_D^{k_*} \phi - \widetilde{\mathcal{S}}_D \psi = u_0 \quad \text{on } \partial D, \\ \dfrac{1}{\mu_*} \dfrac{\partial \mathcal{S}_D^{k_*} \phi}{\partial \nu}\bigg|_- - \dfrac{1}{\mu_1} \dfrac{\partial \widetilde{\mathcal{S}}_D \psi}{\partial \nu}\bigg|_+ = \dfrac{1}{\mu_1} \dfrac{\partial u_0}{\partial \nu} \quad \text{on } \partial D. \end{cases} \tag{14}$$

Here ν denotes the outward unit normal to ∂D; subscripts $+$ and $-$ indicate the limiting values as we approach ∂D from outside D and from inside D.

Proof. Define the operator $T : L^2(\partial D) \times L^2(\partial D) \to L^2(\partial D) \times L^2(\partial D)$ by

$$T(\phi,\psi) = \left(\mathcal{S}_D^{k_*}\phi - \widetilde{\mathcal{S}}_D\psi,\ \frac{1}{\mu_*}\frac{\partial \mathcal{S}_D^{k_*}\phi}{\partial \nu}\bigg|_- - \frac{1}{\mu_1}\frac{\partial \widetilde{\mathcal{S}}_D\psi}{\partial \nu}\bigg|_+ \right).$$

By (12), T is a Fredholm type of operator. Thus, in order to prove the existence and uniqueness of a solution to (14), it is enough to show that T is injective. Let the pair $(\phi,\psi) \in L^2(\partial D) \times L^2(\partial D)$ be a solution to the following homogeneous system of integral equations:

$$\begin{cases} \mathcal{S}_D^{k_*}\phi - \widetilde{\mathcal{S}}_D\psi = 0 \quad \text{on } \partial D, \\ \dfrac{1}{\mu_*}\dfrac{\partial \mathcal{S}_D^{k_*}\phi}{\partial \nu}\bigg|_- - \dfrac{1}{\mu_1}\dfrac{\partial \widetilde{\mathcal{S}}_D\psi}{\partial \nu}\bigg|_+ = 0 \quad \text{on } \partial D. \end{cases}$$

Introduce

$$v(X) = \begin{cases} \widetilde{\mathcal{S}}_D\psi(X), & X \in \mathbb{R}^2_+ \setminus \overline{D}, \\ \mathcal{S}_D^{k_*}\phi(X), & X \in D. \end{cases}$$

It is easy to see that v satisfies the equation $(\nabla \cdot (1/\mu_\alpha)\nabla + \omega^2\varepsilon_\alpha)\, v = 0$ in $\mathbb{R}^2_+$, with the boundary condition $v = 0$ on $y = 0$ together with the decay estimates

$$\left| v_l(x) \mp i\beta_l \frac{dv_l}{dx}(x) \right| = O\left(\frac{1}{|x|}\right) \quad \text{as } x \to \pm\infty,$$

for $l = 1, \dots, m$, where

$$v_l(x) = \int_0^{+\infty} v(x,y) g(y,\lambda_l)\, dy,$$

which hold because of the form of the Green's function G. Then, it immediately follows from (H1) that $v = 0$ in $\mathbb{R}^2_+$. Next, the unique continuation for the operator $(\Delta + \omega^2\varepsilon_*\mu_*)$ yields $\mathcal{S}_D^{k_*}\phi = 0$ in D. Since $(\Delta + \omega^2\varepsilon_1\mu_1)\widetilde{\mathcal{S}}_D\psi = 0$ in D and $\omega\sqrt{\varepsilon_1\mu_1}$ is not a Dirichlet eigenvalue of $-\Delta$ on D, then $\widetilde{\mathcal{S}}_D\psi = 0$ in $\mathbb{R}], \times]0, h[$, which leads to a contradiction because of the jump of the normal derivative of $\widetilde{\mathcal{S}}_D\psi$ on ∂D. ∎

The derivation of the asymptotic formula for $u_\alpha - u_0$ relies on the representation formula (13) and is parallel to that in [2]. However, there are some technical differences, and so we include the main steps for its derivation.

Let us introduce two more layer potentials. Define

$$\mathcal{S}_D\phi(X) = \int_{\partial D} \Gamma(X - Y)\phi(Y) d\sigma(Y), \quad X \in \mathbb{R}^2,$$

where $\Gamma(X)$ is the fundamental solution for the Laplacian Δ. We also define

$$\mathcal{S}_D^0\phi(X) = \frac{1}{\mu_1}\int_{\partial D} G_0(X,Y)\psi(Y) d\sigma(Y), \quad X \in \mathbb{R}^2.$$

Let

$$\widehat{\phi}(Y) := \phi(Z + \alpha Y), \quad \widehat{\psi}(Y) := \psi(Z + \alpha Y), \quad Y \in \partial B. \tag{15}$$

Because of (12), we have

$$\begin{aligned} G(Z+\alpha X, Z+\alpha Y) &= G_0(Z+\alpha X, Z+\alpha Y) + C + O(\alpha|X-Y|) \\ &= G_0(X,Y) + C + O(\alpha|X-Y|), \quad X,\ Y \in \partial B, \end{aligned}$$

for some constant C. Therefore,

$$\widetilde{\mathcal{S}}_D\phi(Z+\alpha X) = \alpha \mathcal{S}_B^0 \widehat{\phi}(X) + C + O(\alpha^2), \quad X \in \partial B,$$

where $O(\alpha^2) \leq C\alpha^2 \|\widehat{\phi}\|_{L^2(\partial B)}$. Here and in what follows, C denotes a constant which may be different at each occurrence. Since $\Gamma_{k_*}(X) - \Gamma(X)$ is $\mathcal{C}^1(\mathbb{R}^2)$, we also have

$$\mathcal{S}_D^{k_*}\phi(Z+\alpha X) = \alpha \mathcal{S}_B \widehat{\phi}(X) + C + O(\alpha^2), \quad X \in \partial B.$$

Since $u_0(Z+\alpha Y) = u_0(Z) + \alpha \nabla u_0(Z)\cdot Y + o(\alpha)$, the integral equation (14) takes the form

$$\begin{cases} \mathcal{S}_B\widehat{\phi} - \mathcal{S}_B^0\widehat{\psi} = C + \nabla u_0(Z)\cdot Y + O(\alpha) \quad \text{on } \partial B, \\ \dfrac{1}{\mu_*}\dfrac{\partial \mathcal{S}_B\widehat{\phi}}{\partial \nu}\bigg|_- - \dfrac{1}{\mu_1}\dfrac{\partial \mathcal{S}_B^0\widehat{\psi}}{\partial \nu}\bigg|_+ = \dfrac{1}{\mu_1}\nabla u_0(Z)\cdot\dfrac{\partial Y}{\partial \nu} + O(\alpha) \quad \text{on } \partial B. \end{cases} \tag{16}$$

Let (f,g) be the solution to

$$\begin{cases} \mathcal{S}_B f - \mathcal{S}_B^0 g = C + \nabla u_0(Z)\cdot Y \quad \text{on } \partial B, \\ \dfrac{1}{\mu_*}\dfrac{\partial \mathcal{S}_B f}{\partial \nu}\bigg|_- - \dfrac{1}{\mu_1}\dfrac{\partial \mathcal{S}_B^0 g}{\partial \nu}\bigg|_+ = \dfrac{1}{\mu_1}\nabla u_0(Z)\cdot\dfrac{\partial Y}{\partial \nu} \quad \text{on } \partial B. \end{cases} \tag{17}$$

Then

$$\widehat{\psi} = g + O(\alpha) \quad \text{on } \partial B. \tag{18}$$

Since $C + \nabla u_0(Z)\cdot Y$ is harmonic in B, the first equation in (17) yields

$$\mathcal{S}_B f(Y) - \mathcal{S}_B^0 g(Y) = C + \nabla u_0(Z)\cdot Y, \quad Y \in B,$$

and hence

$$\frac{\partial \mathcal{S}_B f}{\partial \nu}\bigg|_- - \frac{\partial \mathcal{S}_B^0 g}{\partial \nu}\bigg|_- = \nabla u_0(Z)\cdot\frac{\partial Y}{\partial \nu} \quad \text{on } \partial B.$$

Combining this with the second equation in (17), we get

$$\frac{\partial \mathcal{S}_B^0 g}{\partial \nu}\bigg|_+ - \frac{\mu_1}{\mu_*}\frac{\partial \mathcal{S}_B^0 g}{\partial \nu}\bigg|_- = \left(\frac{\mu_1}{\mu_*} - 1\right)\nabla u_0(Z)\cdot\frac{\partial Y}{\partial \nu} \quad \text{on } \partial B. \tag{19}$$

Observe that for each $h \in L^2(\partial B)$ with $\int_{\partial B} h\,d\sigma = 0$ there exists a unique solution $g \in L^2(\partial B)$ with $\int_{\partial B} g\,d\sigma = 0$ to the equation

$$\frac{\partial \mathcal{S}_B^0 g}{\partial \nu}\bigg|_+ - \frac{\mu_1}{\mu_*}\frac{\partial \mathcal{S}_B^0 g}{\partial \nu}\bigg|_- = h \quad \text{on } \partial B.$$

This fact can be proved using the method in Chapter 1 of [1], and so we omit its proof.

Let $Y = (y_1, y_2)$ and ψ_j, $j = 1, 2$, be the solution to

$$\left.\frac{\partial \mathcal{S}_B^0 \psi_j}{\partial \nu}\right|_+ - \frac{\mu_1}{\mu_*}\left.\frac{\partial \mathcal{S}_B^0 \psi_j}{\partial \nu}\right|_- = \left(\frac{\mu_1}{\mu_*} - 1\right)\frac{\partial y_j}{\partial \nu} \quad \text{on } \partial B. \tag{20}$$

It then follows from (18) and (19) that

$$\widehat{\psi} = \sum_{j=1}^{2} \frac{\partial u_0}{\partial x_j}(Z)\psi_j + O(\alpha) \quad \text{on } \partial B. \tag{21}$$

We are now ready to derive an asymptotic formula for $u_\alpha - u_0$. According to (13),

$$u_\alpha(X) = u_0(X) + \int_{\partial D} G(X, \Xi)\psi(\Xi)d\sigma(\Xi).$$

Making the change of variables $\Xi \to Z + \alpha Y$, $Y \in \partial B$, we get

$$u_\alpha(X) = u_0(X) + \alpha \int_{\partial B} G(X, Z + \alpha Y)\widehat{\psi}(Y)d\sigma(Y),$$

where

$$\widehat{\psi}(Y) := \psi(Z + \alpha Y), \quad Y \in \partial B.$$

Since

$$G(X, Z + \alpha Y) = G(X, Z) + \alpha \nabla_Y G(X, Z) \cdot Y + o(\alpha)$$

for X away from D, we get

$$u_\alpha(X) = u_0(X) + \alpha G(X, Z) \int_{\partial B} \widehat{\psi} d\sigma + \alpha^2 \nabla_Y G(X, Z) \cdot \int_{\partial B} Y\widehat{\psi}(Y)d\sigma(Y) + o(\alpha^2) \tag{22}$$

for X away from D.

By (14) we have

$$\psi = \frac{\mu_1}{\mu_*}\left.\frac{\partial \mathcal{S}_D^{k_*}\phi}{\partial \nu}\right|_- - \frac{\partial u_0}{\partial \nu} - \left.\frac{\partial \widetilde{\mathcal{S}}_D \psi}{\partial \nu}\right|_-,$$

and hence it follows that

$$\begin{aligned}
\alpha \int_{\partial B} \widehat{\psi} d\sigma &= \int_{\partial D} \psi d\sigma \\
&= \frac{\mu_1}{\mu_*} \int_D \Delta \mathcal{S}_D^{k_*}\phi - \int_D \Delta u_0 - \int_D \Delta \widetilde{\mathcal{S}}_D \psi \\
&= \frac{\mu_1}{\mu_*}\omega^2 \varepsilon_* \mu_* \int_D \mathcal{S}_D^{k_*}\phi - \omega^2 \varepsilon_1 \mu_1 \int_D u_0 - \omega^2 \varepsilon_1 \mu_1 \int_D \widetilde{\mathcal{S}}_D \psi \\
&= \omega^2 \mu_1 (\varepsilon_* - \varepsilon_1) \left[\int_D u_0 + \int_D \widetilde{\mathcal{S}}_D \psi\right],
\end{aligned}$$

where the last equality follows from (14). Note that

$$\int_D u_0 = \alpha^2 u_0(Z)|B| + O(\alpha^3).$$

We also have

$$\int_D \widetilde{\mathcal{S}}_D \psi = O(\alpha^3). \tag{23}$$

In fact, since $\int_{\partial B} \psi_j d\sigma = 0$, equation (21) yields $\int_{\partial B} \widehat{\psi} d\sigma = O(\alpha)$, and hence

$$\widetilde{\mathcal{S}}_D \psi(Z + \alpha X) = \alpha \mathcal{S}_B^0 \widehat{\psi} + O(\alpha) O(\alpha).$$

Thus we have (23). Therefore, we obtain

$$\alpha \int_{\partial B} \widehat{\psi} d\sigma = \alpha^2 \omega^2 \mu_1 (\varepsilon_* - \varepsilon_1) u_0(Z)|B| + O(\alpha^3). \tag{24}$$

On the other hand, it follows from (21) that

$$\int_{\partial B} Y \widehat{\psi}(Y) d\sigma(Y) = M \nabla u_0(Z) + O(\alpha), \tag{25}$$

where $M = (M_{ij})$ and

$$M_{ij} = \int_{\partial B} y_j \psi_i(Y) d\sigma(Y), \quad i, j = 1, 2. \tag{26}$$

By (22), (24), and (25), we finally arrive at the following theorem.

THEOREM 3.2. *Let u_α be the solution of* (1)*, and let M be the polarization tensor defined by* (26)*. Then, for $X = (x, y)$ bounded away from D, we have the pointwise expansion*

$$\begin{aligned} u_\alpha(X) = u_0(X) + \alpha^2 \Bigg[& \nabla_Y G(X, Z) \cdot M \nabla u_0(Z) \\ & + \omega^2 \mu_1 \varepsilon_1 \left(\frac{\varepsilon_*}{\varepsilon_1} - 1 \right) |B| G(X, Z) u_0(Z) \Bigg] + o(\alpha^2). \end{aligned} \tag{27}$$

A few words are in order on the matrix M defined by (26). It follows from the jump relation of the single layer potential and (20) that

$$\begin{aligned} \int_{\partial B} y_j \psi_i d\sigma &= \int_{\partial B} y_j \left[\frac{\partial \mathcal{S}_B^0 \psi_j}{\partial \nu} \bigg|_+ - \frac{\partial \mathcal{S}_B^0 \psi_j}{\partial \nu} \bigg|_- \right] d\sigma \\ &= \left(\frac{\mu_1}{\mu_*} - 1 \right) \int_{\partial B} y_j \frac{\partial \Phi_i}{\partial \nu} \bigg|_- d\sigma, \end{aligned}$$

where

$$\Phi_i(Y) = y_i + \mathcal{S}_B^0 \psi_i(Y), \quad Y \in \mathbb{R}^2.$$

Note that $\Phi_i(Y)$ for $Y = (y_1, y_2)$ is the solution to

$$\begin{cases} \Delta\Phi_i = 0 \quad \text{in } B \cup \left(\mathbb{R} \times \right]-\frac{z_y}{\alpha}, \frac{(h-z_y)}{\alpha}\left[\setminus \overline{B}\right) \cup \mathbb{R} \times \left]\frac{(h-z_y)}{\alpha}, +\infty\right[, \\ \Phi_i \text{ is continuous across } \partial B \text{ and } y_2 = \dfrac{(h-z_y)}{\alpha}, \\ \left.\dfrac{\partial \Phi_i}{\partial \nu}\right|_+ - \dfrac{\mu_1}{\mu_*}\left.\dfrac{\partial \Phi_i}{\partial \nu}\right|_- = 0 \quad \text{on} \quad \partial B, \\ \left.\dfrac{\partial \Phi_i}{\partial y}\right|_+ - \dfrac{\mu_2}{\mu_1}\left.\dfrac{\partial \Phi_i}{\partial y}\right|_- = 0 \quad \text{on} \quad y_2 = \dfrac{(h-z_y)}{\alpha}, \\ \Phi_i(Y) - y_i \to 0 \quad \text{as } |Y| \to \infty, \\ \Phi_i(Y) = 0 \quad \text{on } y_2 = -\dfrac{z_y}{\alpha}. \end{cases}$$

In its appearance M_{ij} may seem to be dependent on α. However, $M_{ij} = \text{constant} + O(\alpha)$. To see this, let us investigate three typical cases: (i) when $D = Z + \alpha B$ is away from the interface $y_2 = h$ and the boundary $y_2 = 0$, (ii) when D is close to the interface, (iii) when D is close to the boundary.

(i) Suppose that D is away from the interface and the boundary. In this case, after scaling, the distance from B to the interface $y_2 = (h - z_y)/\alpha$ and the boundary is of order $1/\alpha$. Thus one can see from (10) that

$$G_0(X, Y) = \mu_1 \Gamma(X - Y) + O(\alpha), \quad X, Y \in \partial B,$$

and hence (20) can written as

$$\left.\frac{\partial \mathcal{S}_B \psi_j}{\partial \nu}\right|_+ - \frac{\mu_1}{\mu_*}\left.\frac{\partial \mathcal{S}_B \psi_j}{\partial \nu}\right|_- = \left(\frac{\mu_1}{\mu_*} - 1\right)\frac{\partial y_j}{\partial \nu} + O(\alpha) \quad \text{on } \partial B. \tag{28}$$

Let g_j be the solution of (28) without $O(\alpha)$-term on the right-hand side. Then, $M(\frac{\mu_1}{\mu_*})$ defined by

$$M_{ij}\left(\frac{\mu_1}{\mu_*}\right) := \int_{\partial B} y_j g_i d\sigma$$

is the Pólya–Szegö polarization tensor whose properties were extensively studied in [1]. We get from (28) that

$$M = M\left(\frac{\mu_1}{\mu_*}\right) + O(\alpha),$$

and hence, in this case, the formula (27) holds with M replaced with $M(\frac{\mu_1}{\mu_*})$. Recall that if the inclusion B is a disk, then its polarization tensor $M(\frac{\mu_1}{\mu_*})$ takes the following explicit form:

$$M = \frac{2(\mu_1 - \mu_*)}{\mu_1 + \mu_*}|B|\, I_2, \tag{29}$$

where I_2 is the 2×2 identity matrix.

(ii) Suppose that D is close to the interface and that the distance between them is of order α. In this case, one can see from (10) that

$$G_0(X,Y) = \mu_1\left(\Gamma(X-Y) - \frac{\mu_2-\mu_1}{\mu_1+\mu_2}\Gamma(X-Y-(0,2h))\right) + O(\alpha), \quad X,Y \in \partial B.$$

By a similar argument one can show that

$$M_{ij} = \left(\frac{\mu_1}{\mu_*} - 1\right)\int_{\partial B_*} y_j \frac{\partial \widehat{\Phi}_i}{\partial \nu}\bigg|_- d\sigma + O(\alpha) := P_{ij}\left(\frac{\mu_*}{\mu_1}, \frac{\mu_1}{\mu_2}\right) + O(\alpha),$$

where $B_* = B - (0, (h-z_y)/\alpha)$ and $\widehat{\Phi}_i, i = 1,2$, is the solution to

$$\begin{cases} \Delta\widehat{\Phi}_i = 0 \quad \text{in } B_* \text{ and in } (\mathbb{R}\times\,]-\infty,0[\,\setminus\overline{B_*}) \cup \mathbb{R}\times\,]0,+\infty[, \\ \widehat{\Phi}_i \text{ is continuous across } \partial B_* \text{ and } y_2 = 0, \\ \dfrac{\partial\widehat{\Phi}_i}{\partial\nu}\bigg|_+ - \dfrac{\mu_1}{\mu_*}\dfrac{\partial\widehat{\Phi}_i}{\partial\nu}\bigg|_- = 0 \quad \text{on} \quad \partial B_*, \\ \dfrac{\partial\widehat{\Phi}_i}{\partial y}\bigg|_+ - \dfrac{\mu_2}{\mu_1}\dfrac{\partial\widehat{\Phi}_i}{\partial y}\bigg|_- = 0 \quad \text{on} \quad y_2 = 0, \\ \widehat{\Phi}_i(Y) - \widehat{Y}_i \to 0 \quad \text{as } |Y| \to \infty. \end{cases}$$

Here

$$\widehat{Y} = (\widehat{Y}_1, \widehat{Y}_2) = \begin{cases} (y_1, y_2) & \text{for } y_2 > 0, \\ \left(y_1, \dfrac{\mu_1}{\mu_2}y_2 + 1\right) & \text{for } y_2 < 0. \end{cases}$$

Thus in this case, we obtain that for $X = (x,y), 0 < y < h$, bounded away from D, the following pointwise expansion holds:

$$u_\alpha(X) = u_0(X) + \alpha^2\Bigg[\nabla_Y G(X,Z)\cdot P\left(\frac{\mu_*}{\mu_1}, \frac{\mu_1}{\mu_2}\right)\nabla u_0(Z) + \omega^2\mu_1\varepsilon_1\left(\frac{\varepsilon_*}{\varepsilon_1} - 1\right)|B|G(X,Z)u_0(Z)\Bigg] + o(\alpha^2).$$

The feature of the above formula is that it is expressed in terms of the new polarization tensor $P = (P_{ij})$.

The case when D is close to the boundary can be treated in a similar way, which we omit.

4. Reconstruction of the inclusion. In this section we develop a MUSIC type of algorithm for recovering the inclusion D from measurements of propagated modes excited by incident waves. MUSIC is generally used in signal processing problems as a method for estimating the individual frequencies of multiple-harmonic signals [19]. The MUSIC algorithm makes use of the eigenvalue structure of the so-called response matrix. A more detailed description of this algorithm can be found in [10], [7], and [3]; see also [17], [11], [14], and [15] for further background on closely related time-reversal methodologies and on MUSIC in this specific context.

Let $\beta_l = \sqrt{\omega^2\varepsilon_1\mu_1 - \lambda_l}$, and let

$$c_l = -i\frac{\mu_1\sqrt{d^2-\lambda_l}}{\frac{\mu_1}{\mu_2}\phi(h,\lambda_l)^2 + 2\sqrt{d^2-\lambda_l}\int_0^h \phi(y,\lambda_l)^2 dy}$$

for $1 \le l \le m$.

When the incident wave is a guided mode (of the unperturbed waveguide), then

$$u_0(x,y) = g(y,\lambda_{l_0})e^{-i\beta_{l_0}x}$$

for some $1 \le l_0 \le m$. Recall that $X = (x,y)$ and $Z = (z_x, z_y)$.

We compute

$$\nabla u_0(Z) = \begin{pmatrix} i\beta_{l_0} g(z_y,\lambda_{l_0}) \\ g'(z_y,\lambda_{l_0}) \end{pmatrix} e^{-i\beta_{l_0}z_x},$$

and, by making use of (8) and (9), we obtain that

$$\nabla G(X,Z) \approx \sum_{l=1}^m \frac{c_l}{\beta_l} e^{i\beta_l x} e^{-iz_x(\beta_{l_0}+\beta_l)} \begin{pmatrix} -i\beta_l g(z_y,\lambda_l) \\ g'(z_y,\lambda_l) \end{pmatrix} g(y,\lambda_l)$$

as $x \to +\infty$.

Suppose for the sake of simplicity that B is a disk; then, using (29), it follows that

$$\begin{aligned}(u_\alpha - u_0)(X) \approx |D| \sum_{l=1}^m \frac{c_l}{\beta_l} e^{i\beta_l x} g(y,\lambda_l) e^{-iz_x(\beta_{l_0}+\beta_l)} \\ \times \left[\frac{2(\mu_* - \mu_1)}{\mu_* + \mu_1} \big(\beta_{l_0}\beta_l g(z_y,\lambda_l) g(z_y,\lambda_{l_0}) + g'(z_y,\lambda_l) g'(z_y,\lambda_{l_0})\big) \right. \\ \left. + \omega^2 \mu_1 \varepsilon_1 \left(\frac{\varepsilon_*}{\varepsilon_1} - 1\right) g(z_y,\lambda_l) g(z_y,\lambda_{l_0}) \right]\end{aligned}$$

as $x \to +\infty$.

The coefficients of the scattered modes C_{ll_0}, which are excited by the incident wave u_0, are then approximated by

$$\begin{aligned}C_{ll_0} \approx |D| e^{-iz_x(\beta_{l_0}+\beta_l)} \left[\frac{2(\mu_* - \mu_1)}{\mu_* + \mu_1} \big(\beta_{l_0}\beta_l g(z_y,\lambda_l) g(z_y,\lambda_{l_0}) + g'(z_y,\lambda_l) g'(z_y,\lambda_{l_0})\big) \right. \\ \left. + \omega^2 \mu_1 \varepsilon_1 \left(\frac{\varepsilon_*}{\varepsilon_1} - 1\right) g(z_y,\lambda_l) g(z_y,\lambda_{l_0}) \right].\end{aligned}$$

Define the (response) matrix $C = (C_{ll_0})_{l,l_0=1,\dots,m}$. We now show how to apply the MUSIC algorithm for recovering the location Z and the volume $|D|$ of the inclusion from the above approximate formula for the matrix $C \in \mathbb{C}^{m\times m}$. We consider separately three cases in stating the following lemma.

LEMMA 4.1.

(a) *Suppose $\mu_* = \mu_1$. For $X = (x,y)$ in the core of the waveguide, define the vector $\mathbf{g}_{x,y} \in \mathbb{C}^m$ by*

$$\mathbf{g}_{x,y} = \big(g(y,\lambda_1)e^{-ix\beta_1}, \dots, g(y,\lambda_m)e^{-ix\beta_m}\big)^T,$$

where T denotes the transpose. Then

$$\mathbf{g}_{x,y} \in \mathrm{Range}(C) \quad \text{iff } x = z_x \text{ and } y = z_y. \tag{30}$$

(b) *Suppose* $\varepsilon_* = \varepsilon_1$. *For* $X = (x, y)$ *in the core of the waveguide, define the vector* $\mathbf{g}_{x,y} \in \mathbb{C}^{2m}$ *by*

(31)
$$\mathbf{g}_{x,y} = \big((\beta_1 g(y,\lambda_1), g'(y,\lambda_1))^T e^{-ix\beta_1}, \ldots, (\beta_m g(y,\lambda_m), g'(y,\lambda_m))^T e^{-ix\beta_m}\big)^T.$$

Then

$$\mathbf{g}_{x,y} \in \text{Range}(C) \quad \textit{iff } x = z_x \textit{ and } y = z_y.$$

(c) *Suppose* $\mu_* \neq \mu_1$ *and* $\varepsilon_* \neq \varepsilon_1$. *For* $X = (x, y)$ *in the core of the waveguide, define the vector* $\mathbf{g}_{x,y} \in \mathbb{C}^{3m}$ *by*

$$\mathbf{g}_{x,y} = \big((\beta_1 g(y,\lambda_1), g'(y,\lambda_1), g(y,\lambda_1))^T e^{-ix\beta_1}, \ldots, (\beta_m g(y,\lambda_m), g'(y,\lambda_m), g(y,\lambda_m)^T e^{-ix\beta_m}\big)^T.$$

Then

$$\mathbf{g}_{x,y} \in \text{Range}(C) \quad \textit{iff } x = z_x \textit{ and } y = z_y.$$

Proof. The idea of the proof of the characterization of the location of the inclusion in terms of the range of the matrix C is the same for the three cases above. Let us then for the sake of simplicity consider only the first case. For $X = (x, y)$ suppose that $\mathbf{g}_{x,y} \in \text{Range}(C)$ and $X \neq Z$. Then

(32) $\mathbf{g}_{x,y}$ is proportional to the vector $\big(g(z_y,\lambda_1)e^{-i\beta_1 z_x}, \ldots, g(z_y,\lambda_m)e^{-i\beta_m z_x}\big)^T$.

Consider now the Green's functions $G(\cdot, X)$ and $G(\cdot, Z)$. Identity (32) yields that the guided components of these Green's functions are proportional. This implies that the Green's functions $G(Y, X)$ and $G(Y, Z)$ are proportional for any Y in the core, $Y \notin \{X, Z\}$. The singularity of $G(\cdot, X)$ at the source X (see Lemma 2.1) then leads to a contradiction. ∎

The MUSIC algorithm is as follows. Denote by P the orthogonal projection onto the left null space (noise space) of C, which can be computed via a singular value decomposition (SVD) of the matrix C. We can form an image of the location Z by plotting, at each point $X = (x, y)$, the quantity

(a) $W := \|\mathbf{g}_{x,y}\|/\|P\mathbf{g}_{x,y}\|$ if $\mu_* = \mu_1$.
(b) $W_b := \|b \cdot \mathbf{g}_{x,y}\|/\|P(b \cdot \mathbf{g}_{x,y})\|$ for $b \in \mathbb{R}^2 \setminus \{0\}$ if $\varepsilon_* = \varepsilon_1$.
(c) $W_c := \|c \cdot \mathbf{g}_{x,y}\|/\|P(c \cdot \mathbf{g}_{x,y})\|$ for $c \in \mathbb{R}^3 \setminus \{0\}$ if $\mu_* \neq \mu$ and $\varepsilon_* \neq \varepsilon_1$.

In the case (a) (resp., (b), (c)), the matrix C has 1 (resp., 2, 3) significant singular values. The image space of C is of dimension 1 (resp., 2, 3).

This MUSIC type of algorithm can be used to recover the location of n well-separated electromagnetic inclusions, provided that $m > n$ (case (a)), $m > 2n$ (case(b)), and $m > 3n$ (case (c)).

5. Numerical results. Consider the function $f(\lambda)$ defined by

$$f(\lambda) = \sqrt{d^2 - \lambda}\tan\sqrt{\lambda}h + \frac{\mu_2}{\mu_1}\sqrt{\lambda}, \quad \lambda \in]0, d^2[. \tag{33}$$

From (4), the isolated eigenvalues λ_l, $l = 1, \ldots, m$, are defined by

$$f(\lambda_l) := \sqrt{\lambda_l - d^2}\ \text{tg}\ \sqrt{\lambda_l}h + \frac{\mu_2}{\mu_1}\sqrt{\lambda_l} = 0, \quad l = 1, \ldots, m.$$

We set $\mu_1 = 2, \varepsilon_1 = 2$ and $\mu_2 = 1, \varepsilon_2 = 1$, $\omega = 4$, and $h = 4$. For this set of parameters, the function $f(\lambda)$ has 8 zeros, $\lambda_l, l = 1, \ldots, 8$.

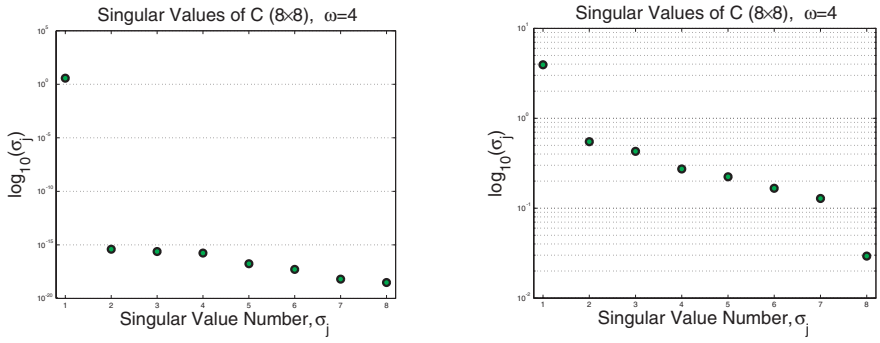


FIG. 1. *Case* (a) *(dielectric contrast only): distribution of the singular values of C (left) and in the case of noisy data* (12 *dB) (right).*

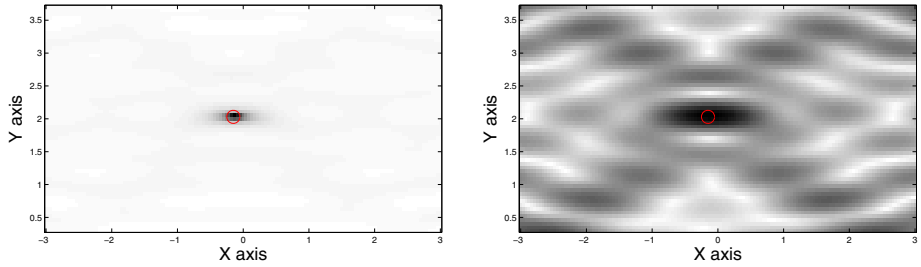


FIG. 2. *Case* (a) *(dielectric contrast only): gray-level (or color) maps of the amplitudes of W (left) and* $\mathcal{D}^{diel}$ *(right).*

5.1. Reconstruction of one inclusion. We consider a small homogeneous circular disk D of diameter $\alpha = 0.1$, centered at $Z = (-0.45, 2.03)$ within a rectangle search box prescribed as $\Omega = [-3, 3] \times [\delta, h - \delta] \subset \mathbb{R}^2$, where $\delta = 0.3$. The parameters of the inclusion are set to $\varepsilon_* = 5$ and/or $\mu_* = 5$. We need one more notation:

$$\mathbf{g}^1_{x,y} = \left(\beta_1 g(y, \lambda_1) e^{-ix\beta_1}, \ldots, \beta_m g(y, \lambda_m) e^{-ix\beta_m}\right)^T,$$

$$\mathbf{g}^2_{x,y} = \left(g'(y, \lambda_1) e^{-ix\beta_1}, \ldots, g'(y, \lambda_m) e^{-ix\beta_m}\right)^T,$$

$$\tau_1 = |D| \frac{2(\mu_* - \mu_1)}{\mu_* + \mu_1} \quad \text{and} \quad \tau_2 = |D| \omega^2 \varepsilon_1 \mu_1 \left(\frac{\varepsilon_*}{\varepsilon_1} - 1\right).$$

Using this notation, the vector $\mathbf{g}_{x,y}$, defined by (31), can be written as $\mathbf{g}_{x,y} = [\mathbf{g}^1_{x,y} \mathbf{g}^2_{x,y}]$.

5.1.1. Dielectric inclusion. We set $\varepsilon_* = 5$ and $\mu_* = \mu_1 = 2$. The singular values of the matrix $C \in \mathbb{C}^{8\times 8}$ and those in the case of noisy data (white Gaussian noise for both the amplitude and the phase of the scattered modes) with 12 dB signal-to-noise ratio are displayed in Figure 1. The maps of the amplitudes of W and the product $\mathcal{D}^{\text{diel}} = (v_1 v_1^*)\, \mathbf{g}_{x,y}$ in the case of noisy data are shown in Figure 2. Here v_1 denotes the first singular vector of the matrix C (i.e., the first eigenvector of the matrix CC^*).

The numerical results obtained here are easy to interpret. One singular value emerges from the seven others in the noise subspace. As for the singular vector,

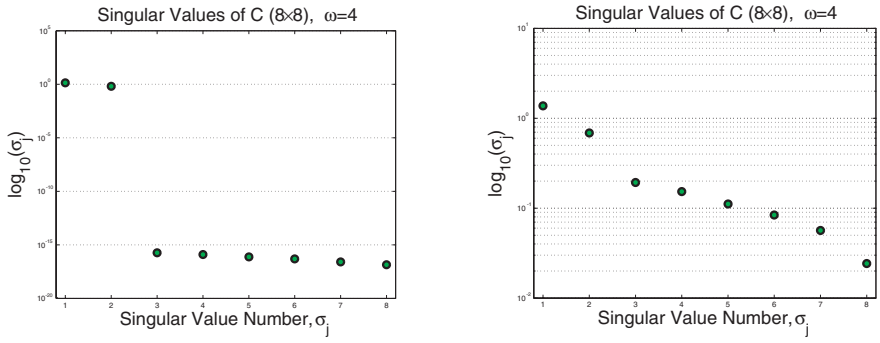


FIG. 3. *Case* (b) *(permeable contrast only): distribution of the singular values of C (left) and in the case of noisy data* (12 *dB) (right).*

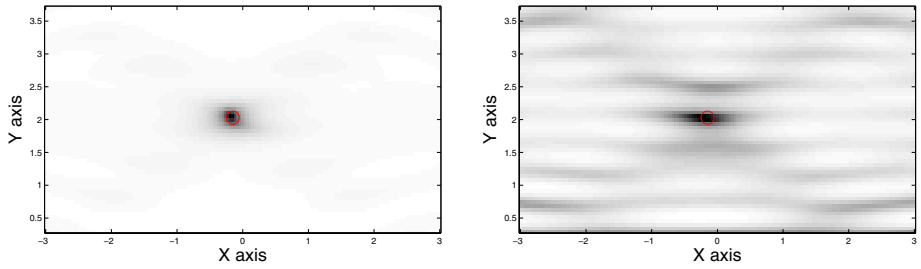


FIG. 4. *Case* (b) *(permeable contrast only): gray-level maps of the amplitudes of* W_b, $b = (1,0),(0,1)$ *(ordered from left to right).*

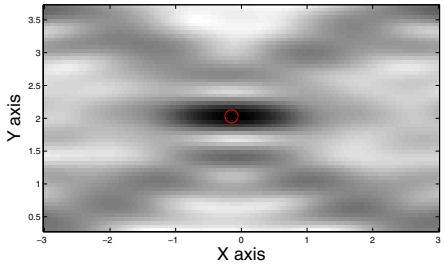


FIG. 5. *Case* (b) *(permeable contrast only): gray-level map of the amplitude of* $\mathcal{D}^{perm}$.

once operated upon by $\mathbf{g}_{x,y}$, it focuses onto the inclusion as expected. Note that, in this simple case, $v_1 = \mathbf{g}_{z_x,z_y}/\|\mathbf{g}_{z_x,z_y}\|$, with corresponding singular value $\sigma_1 = \|\mathbf{g}_{z_x,z_y}\|^2|D|\omega^2\varepsilon_1\mu_1(\varepsilon_*/\varepsilon_1 - 1)$.

5.1.2. Permeable inclusion. In this case we set $\mu_* = 5$ and $\varepsilon_* = \varepsilon_1 = 2$. The singular values of C and those in the case of noisy data with 12 dB signal-to-noise ratio are shown in Figure 3. The maps of the amplitudes of W_b, $b = (1,0)$ or $(0,1)$, in the case of noisy data are depicted in Figure 4. Here, the numerical results show that the first singular vector v_1 corresponds to $\mathbf{g}^1_{x,y}$. The map of the amplitude of the product $\mathcal{D}^{\text{perm}} = (v_1 v_1^*)\,\mathbf{g}^1_{x,y} + (v_2 v_2^*)\,\mathbf{g}^2_{x,y}$ is shown in Figure 5. The vectors v_1 and v_2 denote the first two eigenvectors of the matrix CC^*.

As previously, the results obtained in this case are easy to interpret. Only two

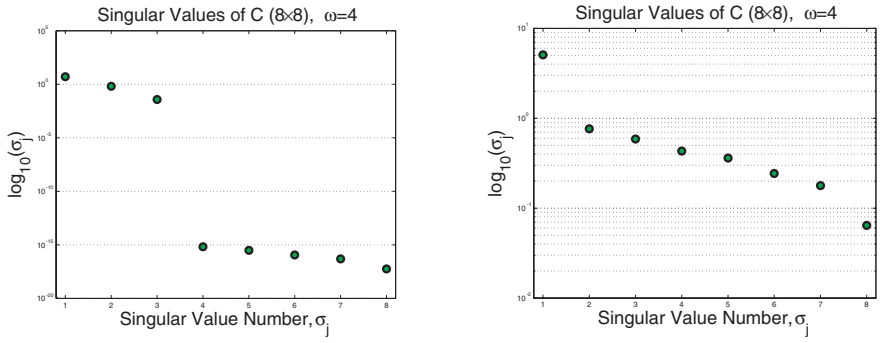


FIG. 6. *Case* (c) *(dielectric and permeable contrasts): distribution of the singular values of* C *(left) and in the case of noisy data (*12 *dB) (right).*

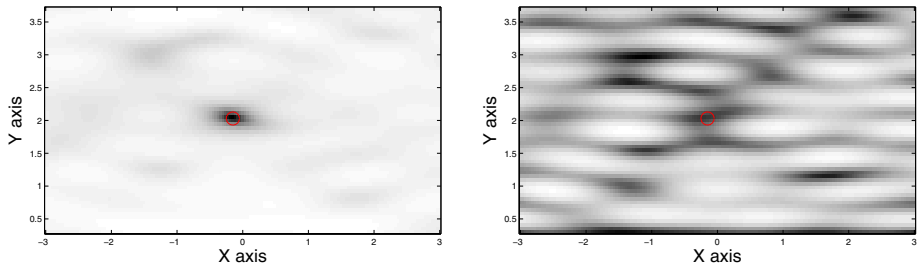


FIG. 7. *Case* (c) *(dielectric and permeable contrasts): gray-level maps of the amplitudes of* W_c, $c = (1, 0, 0)$ *(left),* $c = (0, 1, 0)$ *(right).*

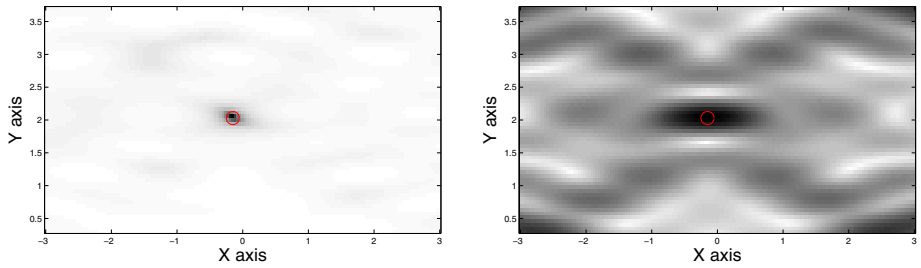


FIG. 8. *Case* (c) *(dielectric and permeable contrasts): gray-level maps of the amplitudes of* W_c, $c = (0, 0, 1)$ *(left), and* $\mathcal{D}$ *(right).*

singular values emerge from noise. The inclusion is clearly discriminated from the background, the visual aspect depending upon the choice of b. The focusing of the singular vectors is rather good.

5.1.3. Dielectric and permeable inclusion. We set $\mu_* = 5$ and $\varepsilon_* = 5$. The singular values of C and those in the case of noisy data with 12 dB signal-to-noise ratio are shown in Figure 6. The maps of the amplitudes of W_c, $c = (1, 0, 0), (0, 1, 0)$, in the case of noisy data are shown in Figure 7. The map of the amplitude of W_c, $c = (0, 0, 1)$, for noisy data is shown in Figure 8 (left). In this case, the numerical results show that all information about the location of the inclusion is contained in the first singular vector v_1. The map of the amplitude of the product $\mathcal{D} = (v_1 v_1^*)\, \mathbf{g}^1_{x,y} + (v_1 v_1^*)\, \mathbf{g}_{x,y}$ in

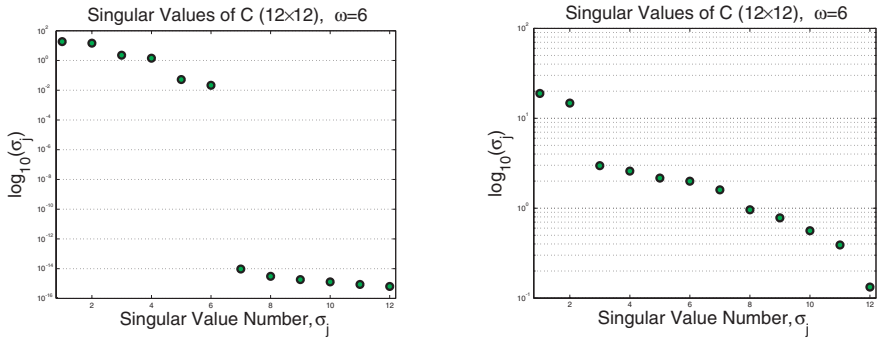


FIG. 9. *Case* (c) *(dielectric and permeable contrasts): distribution of the singular values of* C *(left) and those in the case of noisy data* (12 *dB) (right).*

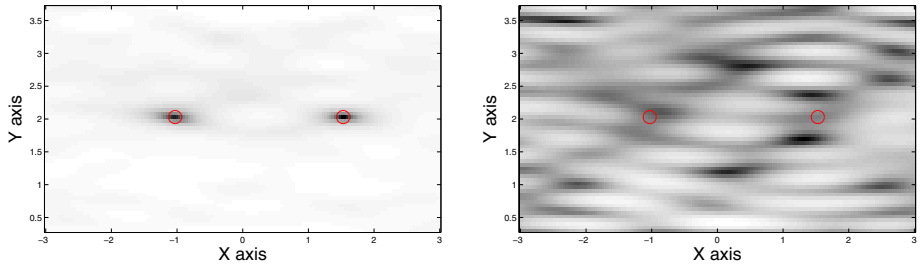


FIG. 10. *Case* (c) *(dielectric and permeable contrasts): gray-level maps of the amplitudes of* W_c, $c = (1, 0, 0)$ *(left),* $c = (0, 1, 0)$ *(right).*

the case of noisy data is shown in Figure 8 (right).

The results obtained in this case are less easy to interpret than before due to the more complicated character of the inclusion. However, the singular values of the signal subspace still emerge from the noise. The inclusion is clearly discriminated from the background (except for $c = (0, 1, 0)$), the visual aspect depending upon the choice of c.

5.2. Reconstruction of multiple inclusions. We set, as in the case of one inclusion, $\mu_1 = 2, \varepsilon_1 = 2$, $\mu_2 = 1, \varepsilon_2 = 1$, and $h = 4$, but now we take $\omega = 6$. For these parameters the function $f(\lambda)$, which is defined by (33), has more than twelve zeros.

We consider two small homogeneous circular disks D_1 and D_2 of diameter $\alpha = 0.1$, centered at $z_1 = (1.53, 2.03)$ and $z_2 = (-1.03, 2.03)$, within the rectangle search box $\Omega = [-3, 3] \times [\delta, h - \delta] \subset \mathbb{R}^2$, $\delta = 0.3$. The parameters of the inclusions are set to $\varepsilon_{*,1} = \varepsilon_{*,2} = 5$ and $\mu_{*,1} = \mu_{*,2} = 5$.

The singular values of C and those in the case of noisy data with 12 dB signal-to-noise ratio are shown in Figure 9. The maps of the amplitudes of W_c, $c = (1, 0, 0), (0, 1, 0)$, in the case of noisy data are shown in Figures 10 and 11 for $c = (0, 0, 1)$.

6. Conclusion. In this paper we have derived a new asymptotic formula for the scattered wave in an open waveguide in the presence of an inclusion of small

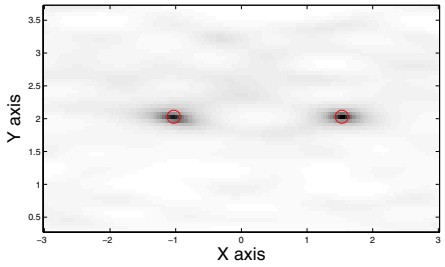


FIG. 11. *Case* (c) *(dielectric and permeable contrasts): gray-level maps of the amplitudes of* W_c, $c = (0, 0, 1)$.

diameter. We then successfully used this formula for the purpose of locating the inclusion from measurements of the propagated modes excited by incident waves, in the form of guided modes of the reference structure. In the case of multiple inclusions, improvements may include the use of a recursive procedure in which the function $W_\cdot(z)$ is changed after each inclusion is found; i.e., a new function $W_\cdot(z)$ is adjusted recursively by projecting the signal space away from the subspace spanned by the inclusions found [16]. Indeed, using more singular vectors than theoretically needed in the presence of noisy data seems to be useful [18].

A mathematical study of the properties of the eigenstructure of the response matrix C can be made following the arguments given in [4]. However, the analysis becomes more complicated because of the form of the Green's function G of the unperturbed waveguide.

REFERENCES

[1] H. AMMARI AND H. KANG, *Reconstruction of Small Inhomogeneities from Boundary Measurements*, Lecture Notes in Math. 1846, Springer-Verlag, Berlin, 2004.

[2] H. AMMARI AND H. KANG, *Boundary layer techniques for solving the Helmholtz equation in the presence of small inhomogeneities*, J. Math. Anal. Appl., 296 (2004), pp. 190–208.

[3] H. AMMARI, E. IAKOVLEVA, AND D. LESSELIER, *A MUSIC algorithm for locating small inclusions buried in a half-space from the scattering amplitude at a fixed frequency*, Multiscale Model. Simul., 3 (2005), pp. 597–628.

[4] H. AMMARI, E. IAKOVLEVA, AND D. LESSELIER, *Two numerical methods for recovering small inclusions from the scattering amplitude at a fixed frequency*, SIAM J. Sci. Comput., to appear.

[5] H. AMMARI AND F. TRIKI, *Resonances for microstrip transmission lines*, SIAM J. Appl. Math., 64 (2003), pp. 601–636.

[6] C. BENDER AND S. ORSZAG, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw–Hill, New York, 1978.

[7] M. CHENEY, *The linear sampling method and the MUSIC algorithm*, Inverse Problems, 17 (2001), pp. 591–595.

[8] J. COYLE AND P. MONK, *Scattering of time-harmonic electromagnetic waves by anisotropic inhomogeneous scatterers or impenetrable obstacles*, SIAM J. Numer. Anal., 37 (2000), pp. 1590–1617.

[9] J. A. DESANTO, *Scalar Wave Theory. Green's Functions and Applications*, Springer Series on Wave Phenomena, Springer-Verlag, Berlin, 1992.

[10] A. J. DEVANEY, *Super-resolution processing of multi-static data using time reversal and MUSIC*, J. Acoust. Soc. Amer., to appear.

[11] A. KIRSCH, *The MUSIC algorithm and the factorisation method in inverse scattering theory for inhomogeneous media*, Inverse Problems, 18 (2002), pp. 1025–1040.

[12] R. MAGNANINI AND F. SANTOSA, *Wave propagation in a* 2*-D optical waveguide*, SIAM J. Appl. Math., 61 (2000), pp. 1237–1252.

[13] R. Magnanini and F. Santosa, *Scattering in a 2D optical waveguide*, in Computational and Analytic Methods in Scattering and Applied Mathematics (a volume in memory of Ralph E Kleinman), F. Santosa and I. Stakgold, eds., Res. Notes Math. 417, CRC Press, Boca Raton, FL, 2000, pp. 195–208.

[14] T. D. Mast, A. I. Nachman, and R. C. Waag, *Focusing and imaging using eigenfunctions of the scattering operator*, J. Acoust. Soc. Amer., 102 (1997), pp. 715–725.

[15] N. Mordant, C. Prada, and M. Fink, *Highly resolved detection and selective focusing in a waveguide using the D.O.R.T. method*, J. Acoust. Soc. Amer., 105 (1999), pp. 2634–2642.

[16] J. C. Mosher and R. M. Leahy, *Source localization using recursively applied and projected (RAP) MUSIC*, IEEE Trans. Signal Processing, 47 (1999), pp. 332–340.

[17] C. Prada and J.-L. Thomas, *Experimental subwavelength localization of scatterers by decomposition of the time reversal operator interpreted as a covariance matrix*, J. Acoust. Soc. Amer., 114 (2003), pp. 235–243.

[18] C. Prada, J.-L. Thomas, and M. Fink, *The iterative time reversal process: Analysis of the convergence*, J. Acoust. Soc. Amer., 97 (1995), pp. 62–71.

[19] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1992.

[20] M. S. Vogelius and D. Volkov, *Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities*, Math. Model. Numer. Anal., 34 (2000), pp. 723–748.

[21] C. Wilcox, *Spectral analysis of the Pekeris operator in the theory of acoustic wave propagation*, Arch. Ration. Mech. Anal., 60 (1975), pp. 259–300.

[22] C. Wilcox, *Sound Propagation in Stratified Fluids*, Appl. Math. Sci. 50, Springer-Verlag, New York, 1984.

[23] B. Zhang and S. N. Chandler-Wilde, *Acoustic scattering by an inhomogeneous layer on a rigid plate*, SIAM J. Appl. Math., 58 (1998), pp. 1931–1950.

SIAM J. APPL. MATH.
Vol. 65, No. 6, pp. 2128–2140

TRANSPORT OF NUTRIENTS IN BONES*

GUILLERMO H. GOLDSZTEIN†

Abstract. Lacunar-canalicular systems are networks of pores (lacunae) interconnected by thin channels (canaliculi) that are embedded in bones. The efficient transport of nutrients within lacunar-canalicular systems is necessary to keep bones healthy. Several theories have been proposed to identify the physical phenomena responsible for this efficient transport. In this paper, we develop and study a mathematical model motivated by one of those theories.

Key words. bones, porous media, nutrient transport, solute transport, effective diffusion, mathematical modeling

AMS subject classifications. 76S05, 76R50, 92B05

DOI. 10.1137/040616632

1. Introduction. As illustrated in Figure 1.1, we consider long bones of our extremities such as the femur. Bones are porous media with complex microgeometry. The particular components of bones that we discuss here are osteons. These are cylindrical structures of about 120 μm radii that extend along the long axis of the bone (see Figure 1.1). An osteonal canal is located at the center of osteons. This canal contains blood vessels, a nerve, and bone fluid (see Figure 1.1). Pores, called lacunae, are distributed within the osteon. Thin channels, called canaliculi, and the lacunae form a connected system known as the lacunar-canalicular system, which is filled with fluid and is connected to the osteonal canal. A cartoon of the osteon microgeometry is given in Figure 1.1. More details on the structure of bones are given in [4, 12, 13, 2, 14] among numerous other articles.

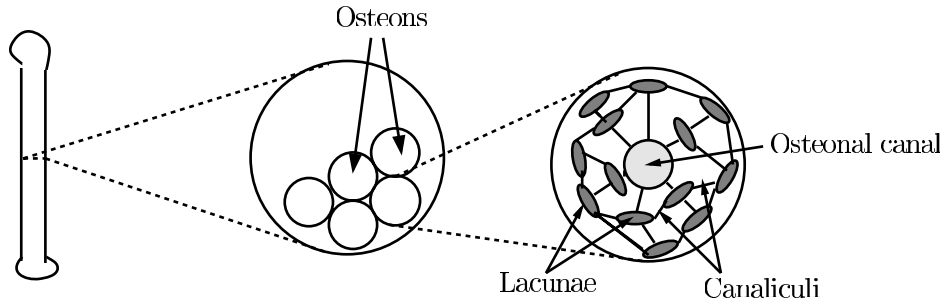


FIG. 1.1. *At left is the cartoon of a bone. A horizontal cross section (view from the top) is shown in the middle figure. The right-most image is a cartoon of the microgeometry of a section of an osteon (also viewed from the top); the white region is the solid part of the bone.*

Bones consume nutrients and produce waste products. It is believed that nutrients are transported from the osteonal canal into the rest of the osteon through the lacunar-canalicular system (see [17] and references therein). Waste products, on the other hand, are produced within the osteon and need to be transported to the osteonal

*Received by the editors October 8, 2004; accepted for publication (in revised form) March 23, 2005; published electronically August 9, 2005. This research was supported by the NSF.
http://www.siam.org/journals/siap/65-6/61663.html
†School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (ggold@math.gatech.edu).

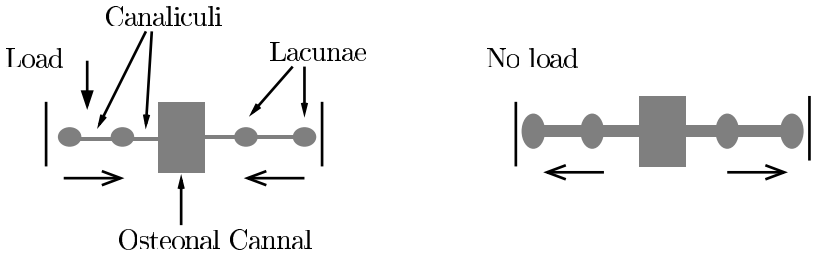


FIG. 1.2. *Cartoon of a vertical longitudinal section (parallel to the direction of the bone) of part of an osteon. The bone is supporting a load in the left-hand image. The load is removed in the right-hand image. The arrows at the bottom denote the direction of the flow.*

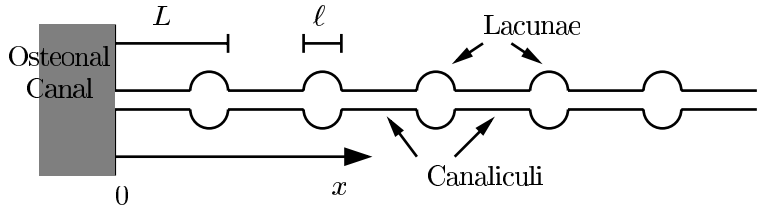


FIG. 1.3. *One-dimensional lacunar-canalicular system.*

canal so that they can be disposed of. Thus, efficient transport within the lacunar-canalicular system is necessary to maintain a healthy bone.

The solid part of the bone is an elastic material. Thus, activities that apply and remove loads to bones, such as walking, produce small deformations of the bone. As a consequence, the incompressible fluid that fills the lacunar-canalicular system is squeezed, producing fluid flow. This is illustrated in Figure 1.2. As shown in the left-hand image, when a bone is subjected to a load, fluid is squeezed out of the lacunar-canalicular system and into the osteonal canal. Once the load is removed, an equal volume of fluid flows back from the osteonal canal into the lacunar-canalicular system (see the right-hand image in Figure 1.2). After each cycle of a periodic load, there is no net volume of fluid transported from the osteonal canal into the lacunar-canalicular system simply because the volume of the bone does not change after each cycle.

In this paper we study the ideal one-dimensional lacunar-canalicular system illustrated in Figure 1.3. The length of each canaliculus is $L - \ell$, and the diameter of each lacuna is of the order ℓ (recall that the diameter of a set Ω is $\sup_{x,y\in\Omega} \|x - y\|$, where $\|x - y\|$ is the distance from x to y). As suggested by Figure 1.3, we assume that $L \gg \ell$.

We assume that there is a periodic exchange of fluid between the osteonal canal and the lacunar-canalicular system and denote that period by t_0. Motivated by the above discussion on flows in lacunar-canalicular systems induced by periodic applied loads, we assume that there is no net transport of any volume of fluid between the osteonal canal and the lacunar-canalicular system after each period. More precisely, we assume that there is $t^\star < t_0$ such that a volume V_F of fluid flows from the osteonal canal into the lacunar-canalicular system during the time interval $(0, t^\star)$ and the same volume of fluid flows back into the osteonal canal during the time interval $(t^\star, t_0)$.

We denote by V_ℓ the volume of each lacuna. Let D be the coefficient of diffusion of nutrients in the host liquid. Due to diffusion, the mixing of nutrients with the host

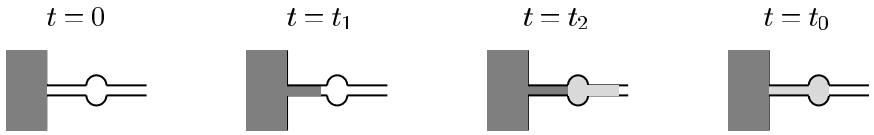


FIG. 1.4. *Nutrient concentration within the lacunar-canalicular system at four different times. The shades represent the concentration of nutrients. The darker the region, the higher the concentration of nutrients.*

fluid within each lacuna occurs in times of order ℓ^2/D. The time that an element of fluid stays in a lacuna is of the order $V_\ell t_0/V_F$. We restrict our attention to the parameter regime

$$\ell^2/D \ll \min\left\{1, \frac{V_\ell}{V_F}\right\} t_0. \tag{1.1}$$

Thus, since we are interested in events that occur in the time scale of t_0, we can and do assume that mixing is instantaneous within each lacuna.

Within the canaliculi, nutrients are transported along the direction of the x-axis by convection and an effect known as Taylor dispersion (that is the result of fluid motion and diffusion); see [16, 1]. Note that velocities within the canaliculi are of the order $V_F/(at_0)$, where a is the cross-sectional area of each canaliculus. Thus, in a period of time of the order t_0, the distances traveled by nutrients due to Taylor dispersion are of the order $\sqrt{(D + V_F^2/(48at_0^2D))t_0}$; see [16, 1]. We will assume that

$$\sqrt{Dt_0 + \frac{V_F^2}{48at_0D}} \ll L. \tag{1.2}$$

Since, as we will see later in the paper, we are interested in the parameter regime where fluid is convected from one end of a canaliculus to the other in times smaller than t_0, we can and do neglect diffusion and Taylor dispersion within the canaliculi in the direction of the x-axis.

The osteonal canal acts as a reservoir of nutrients, and thus the concentration of nutrients within the osteonal canal remains constant in time. Assume that the lacunar-canalicular system does not contain any nutrients initially. If $V_F < V_c$, where V_c is the volume of a canaliculus, the nutrients that enter the lacunar-canalicular system in the time interval $(0, t^\star)$ do not reach the first lacuna. Since there is no dispersion within the canaliculi and there is no net transport of volume of fluid between the osteonal canal and the lacunar-canalicular system after each period, all the nutrients that enter the system flow back to the osteonal canal during the time interval $(t^\star, t_0)$. Therefore, there is no transport of nutrients at the end of a cycle.

Assume now that $V_F > V_c$. In Figure 1.4 we show the osteonal canal and part of the lacunar-canalicular system (that includes only one lacuna) at four different times. The shades represent the concentration of nutrients: the darker the region, the higher the concentration of nutrients. At $t = 0$, the lacunar-canalicular system does not contain any nutrients. The times t_1 and t_2 satisfy $0 < t_1 < t_2 < t^\star$. The volume of fluid that enters the system in the time interval $(0, t_1)$ is smaller than V_c. As soon as some nutrient from the osteonal canal reaches the first lacuna, there is instantaneous mixing (of the nutrients with the host fluid) in that lacuna, and the fluid that flows from that lacuna into the second canaliculus carries nutrients (at some

lower concentration). This is illustrated in Figure 1.4; at $t = t_2$, the fluid that entered the system had reached the first lacuna. After a complete period, i.e., $t = t_0$, nutrient that was initially in the osteonal canal will be left in the lacunar-canalicular system (see Figure 1.4). In fact, if n is a positive integer such that $nV_c < V_F$, after a complete cycle there will be nutrients left in the first n lacunae and n canaliculi. Hence, there is a net transport of nutrients after each period. Moreover, in this paper we show that, in the parameter regime $V_c \ll V_F$, the system exhibits a diffusion-like macroscopic behavior with effective diffusion coefficient

$$D_{\text{eff}} = \left(\frac{V_\ell}{V_c + V_\ell}\right)^2 \left(\frac{V_F}{V_c + V_\ell}\right) \left(\frac{L^2}{t_0}\right). \tag{1.3}$$

This paper is motivated by the work in [17], where the authors make the key assumptions of instantaneous mixing within lacunae and negligible Taylor dispersion within canaliculi. They argue that these assumption are valid by showing that (1.1) and (1.2) are satisfied for typical parameter values (see also [18]). They consider a one-dimensional lacunar-canalicular system with five lacunae, where neighboring lacunae are connected by ten canaliculi. They propose a numerical algorithm and explore the system numerically. Our work is based on the same key physical assumptions. Our new contribution is a detailed mathematical analysis, from which a more explicit description of the behavior of the system is achieved; in particular, we obtain (1.3).

Identifying the phenomena responsible for nutrient transport in bones has been a subject of study for several years. It was first proposed in [15] that convection in the lacunar-canalicular system induced by loading and unloading the bone increases the transport of nutrients (see also [11]). This phenomenon was studied experimentally in [5]. However, this is a partial picture of the relevant phenomena, and there is no agreement among the scientific community on the complete set of physical mechanisms responsible for the transport of nutrients in bones. See [17, 3, 6, 7, 8, 11, 9, 10, 18] for some of the proposed theories and related experiments.

The content of the rest of this paper is the following. In section 2 we derive the governing equations. In section 3 and the appendix we obtain the asymptotic approximation to the governing equations in the limit of a thin canaliculus. In section 4 we consider some examples. The paper ends with conclusions in section 5.

2. Governing equations. The lacunar-canalicular system we consider, displayed in Figure 1.3, extends to infinity in one direction. The right wall of the osteonal canal is the origin of the coordinate system, $x = 0$; the location of the ith canaliculus is the segment $[(i-1)L, iL-\ell]$; and the location of the ith lacuna is $[iL-\ell, iL]$. The cross-sectional area of a canaliculus is a. We denote by $V_c = (L-\ell)a$ and V_ℓ the volumes of the canaliculus and lacuna, respectively. We assume a, V_c, and V_ℓ to be constants. An incompressible fluid, of constant density both in space and time, fills the lacunar-canalicular system. The concentration of nutrients in the osteonal canal remains at the constant value c_0 at all times t.

We denote by $c_i(t)$ the concentration of nutrients at time t in the ith lacuna, and thus $\rho_f c_i(t)$ is the density of nutrients at time t in the ith lacuna, where ρ_f is the fluid density. For x in the canaliculi, we denote by $c(x,t)$ the concentration of nutrients at x and time t. Since Taylor dispersion is neglected within the canaliculi, nutrients flow with the same velocity as the fluid $v = v(x,t)$ within the canaliculi (more precisely, $v(x,t)$ is the average fluid velocity in the cross-section of the canaliculi at x). Fluid incompressibility and mass conservation imply that v is independent of x;

i.e., $v = v(t)$. Thus, conservation of nutrients within the canaliculi reduces to

$$\frac{\partial c}{\partial t} + v\frac{\partial c}{\partial x} = 0 \tag{2.1}$$

for all x in canaliculi.

Whenever the velocity is positive, nutrients flow from the ith canaliculus into the ith lacuna at a rate $a\rho_f v(t)c(iL-\ell,t)$. Nutrients also flow out of that same lacuna into the $(i+1)$th canaliculus at a rate $a\rho_f v(t)c_i(t)$. Analogously, when the velocity is negative, nutrients flow from the $(i+1)$th canaliculus into the ith lacuna at a rate $-a\rho_f v(t)c(iL,t)$ and flow out of that same lacuna into the ith canaliculus at a rate $-a\rho_f v(t)c_i(t)$. This implies

$$V_\ell \frac{\mathrm{d}c_i}{\mathrm{d}t} = \begin{cases} av\,(c(iL-\ell,t) - c_i(t)) & \text{when } v(t) > 0, \\ av\,(c_i(t) - c(iL,t)) & \text{when } v(t) < 0 \end{cases} \tag{2.2}$$

for all positive integers i.

Whenever the velocity is positive, there is flow from each lacuna into the canaliculus located at its right, and thus the concentration of nutrients in the left end of a canaliculus is equal to the concentration of nutrients in the adjacent lacuna. Analogously, whenever the velocity is negative, the concentration of nutrients in the right end of a canaliculus is equal to the concentration of nutrients in the lacuna located at the right end of the canaliculus. Thus,

$$\begin{aligned} &c((i-1)L,t) = c_{i-1}(t) && \text{if } v(t) > 0, \\ &c((i+1)L-\ell,t) = c_{i+1}(t) && \text{if } v(t) < 0, \end{aligned} \tag{2.3}$$

the first of the above equations being valid for all integers $i \geq 2$ and the second for all integers $i \geq 0$. Similarly, whenever the velocity is positive, there is flow from the osteonal canal into the first canaliculus, and thus the concentration of nutrients in the left end of the first canaliculus is equal to c_0, the concentration of nutrients in the osteonal canal,

$$c(0,t) = c_0 \quad \text{if } v(t) > 0. \tag{2.4}$$

Consistent with our discussion in the introduction, we assume that the flow velocity in the canaliculi v is a known periodic function with period t_0 and zero time average

$$\int_0^{t_0} v(t)\,\mathrm{d}t = 0. \tag{2.5}$$

To simplify our analysis we assume that there exist $0 < t^\star < t_0$ such that $v(t) > 0$ if $t \in (0, t^\star)$ and $v(t) < 0$ if $t \in (t^\star, t_0)$. Thus, the volume of fluid that flows from the osteonal canal into the lacunar-canalicular system in the time interval $(0, t^\star)$ is

$$V_F = a\int_0^{t^\star} v(t)\mathrm{d}t. \tag{2.6}$$

Equations (2.1)–(2.4) can be solved once initial conditions and boundary conditions at ∞ are provided.

3. Solution to the governing equations in the thin canaliculus limit. Assume that the initial conditions on the concentration of nutrients is regular enough that there exists a smooth function $\rho_{\text{in}} = \rho_{\text{in}}(z)$ (except probably in isolated points) defined for all $z \geq 0$ such that

$$\rho_{\text{in}}(iL) = c_i(0) \quad \text{for all positive integers } i, \tag{3.1}$$

and the limit

$$c_\infty = \lim_{i\to\infty} c_i(0) \tag{3.2}$$

exists. (More precisely, we need that $\rho'_{\text{in}} = O(V_c/(V_F L))$ except in isolated points.)

Let $\rho = \rho(z,t)$ be the solution

$$\frac{\partial \rho}{\partial t} = D_{\text{eff}} \frac{\partial^2 \rho}{\partial z^2} \qquad \text{for } \ t > 0 \ \text{ and } \ z > 0, \tag{3.3}$$

where D_{eff} was defined in (1.3), subject to the initial conditions

$$\rho(z,0) = \rho_{\text{in}}(z) \quad \text{for } \ z > 0 \tag{3.4}$$

and boundary conditions

$$\rho(0,t) = c_0 \ \text{ and } \ \lim_{z\to+\infty} \rho(z,t) = c_\infty \ \text{ for } \ t \geq 0. \tag{3.5}$$

We extend the definition of ρ to $z < 0$ as follows,

$$\rho(z,t) = c_0 \ \text{ if } \ z \leq 0, \tag{3.6}$$

and let $z_i = z_i(t)$ be defined as

$$z_i(t) = iL - \frac{aL}{V_\ell + V_c} \int_0^t v(s)\,\mathrm{d}s. \tag{3.7}$$

In the appendix we show that ρ gives the asymptotic approximation of the concentrations; more precisely,

$$c_i(t) \simeq \rho\,(z_i(t), t) \quad \text{if } \ V_F \gg V_c. \tag{3.8}$$

From (3.3) and (1.3) and the fact that distance between z_i and z_{i+1} remains equal to L for all i and all t, it follows that D_{eff} given in (1.3) is the effective diffusion coefficient of nutrients in the lacunar-canalicular system.

4. Examples. As an example, we now assume that there are no nutrients within the lacunar-canalicular system initially. This corresponds to the initial condition

$$\rho(z,0) = \rho_{\text{in}}(z) = 0 \ \text{ for all } z > 0, \tag{4.1}$$

and the condition at ∞

$$\lim_{z\to+\infty} \rho(z,t) = 0 \ \text{ for } \ t \geq 0. \tag{4.2}$$

Given these conditions, ρ can be obtained explicitly; more precisely,

$$\rho(z,t) = c_0 - c_0 \frac{2}{\sqrt{\pi}} \int_0^{z/(2\sqrt{D_{\text{eff}}t})} \mathrm{e}^{-s^2}\,\mathrm{d}s. \tag{4.3}$$

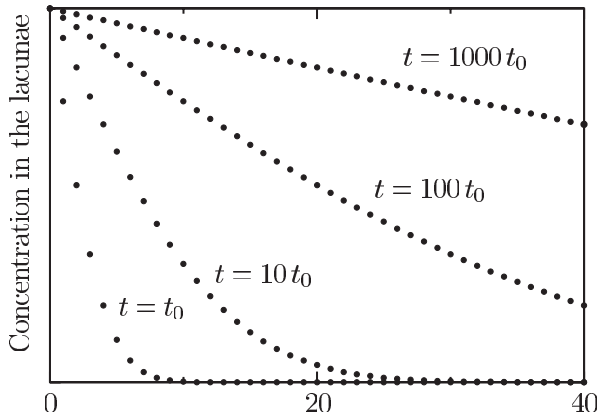


FIG. 4.1. *Plot of the concentration of nutrients in the lacunae* $c_i(t) \simeq \rho(z_i(t), t)$ *versus* i *for different fixed values of* t.

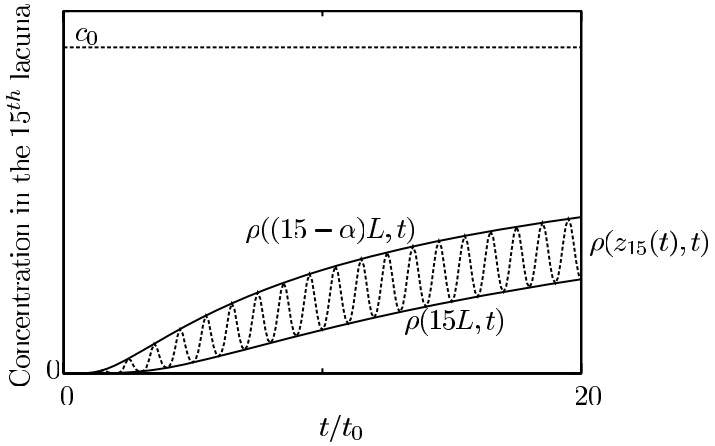


FIG. 4.2. *Plot of concentration of nutrients in the* 15*th lacuna,* $c_{15}(t) \simeq \rho(z_{15}(t), t)$ *(dashed line), and the envelopes* $\rho(15L, t)$ *(lower solid line) and* $\rho((15-\alpha)L, t)$ *(upper solid line) versus normalized time* t/t_0.

In the example presented in this section, we select the velocity

$$v(t) = \frac{\pi V_F}{a t_0} \sin\left(\frac{2\pi t}{t_0}\right), \tag{4.4}$$

and the parameters V_F, V_c, and V_ℓ satisfy $V_c = 0.01 V_\ell$ and $V_\ell = 0.2 V_F$.

In Figure 4.1, we show a plot of the concentration of nutrients in the lacunae, using the approximation $c_i(t) \simeq \rho(z_i(t), t)$, versus i for different fixed values of t.

We define the parameter α as follows:

$$\alpha = \frac{V_F}{V_c + V_\ell}. \tag{4.5}$$

Figure 4.2 shows the evolution of concentration in the fifteenth lacuna $c_{15}(t) \simeq \rho(z_{15}(t), t)$ plotted against normalized time t/t_0. The oscillations in concentration reflect the evolution in concentration in each cycle of the periodic velocity field v. We

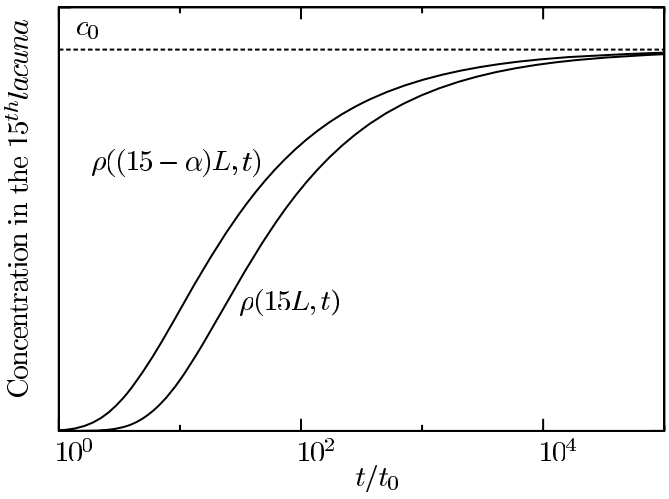


FIG. 4.3. *Plot of the envelopes of the concentration of nutrients in the* 15*th lacuna,* $\rho(15L,t)$ *and* $\rho((15-\alpha)L,t)$*, versus normalized time* t/t_0 *in log scale.*

have also plotted the envelopes that are $\rho(15L,t)$ (lower envelope) and $\rho((15-\alpha)L,t)$ (upper envelope) versus normalized time t/t_0. Figure 4.3 also shows the envelopes $\rho(15L,t)$ and $\rho((15-\alpha)L,t)$ in a longer time scale to illustrate the convergence to c_0 of the concentration in the fifteenth lacuna after a large number of cycles.

5. Conclusions. In this paper we studied the transport of nutrients in a one-dimensional model lacunar-canalicular system. Our motivation was a recently proposed explanation of how nutrients are transported within bones [17]. We have shown that the system exhibits a diffusion-like macroscopic behavior with effective diffusion coefficient, given in (1.3). Note that the effective diffusion coefficient is explicitly given in terms of the geometry of the system and the applied velocity field. Our analysis is the simplest possible that keeps the relevant physics. Nevertheless, our analysis can be extended to include effects neglected here, such as considering a finite and elastic system instead of a rigid and infinite one (as we do in this work). More experimental and theoretical studies are required for a better and more clear understanding of the processes responsible for the transport of nutrients in bones. We hope our work will prove to be an important step toward that goal.

Appendix. Asymptotic analysis of the thin canaliculus limit. Our analysis is valid for thin canaliculus; more precisely, we assume that V_c, the volume of each canaliculus, is much smaller than V_F, the volume of fluid that enters the lacunar-canalicular system during the part of the period where $v > 0$; i.e.,

$$\varepsilon = \frac{V_c}{V_F} \ll 1. \tag{A.1}$$

We first write the velocity v in the form

$$v(t) = \frac{V_F}{at_0} f\left(\frac{t}{t_0}\right). \tag{A.2}$$

Note that $f = f(s)$ is a periodic function of s with period 1 and that

$$\int_0^1 |f(s)|\, \mathrm{d}s = 2. \tag{A.3}$$

We define

$$\beta = t_0\varepsilon \left(\frac{1}{|f(t/t_0)|} + \varepsilon \frac{f'(t/t_0)}{2f^3(t/t_0)} + O(\varepsilon^2) \right). \tag{A.4}$$

In the subsection A.1 we show that

$$\begin{array}{ll} c(iL-\ell,t) = c_{i-1}(t-\beta) & \text{if } v(t) > 0, \\ c(iL,t) = c_{i+1}(t-\beta) & \text{if } v(t) < 0. \end{array} \tag{A.5}$$

We use (A.5) to transform (2.2) into

$$V_\ell \frac{\mathrm{d}c_i}{\mathrm{d}t} = \begin{cases} av\,(c_{i-1}(t-\beta) - c_i(t)) & \text{when } v(t) > 0, \\ av\,(c_i(t) - c_{i+1}(t-\beta)) & \text{when } v(t) < 0 \end{cases} \tag{A.6}$$

for all positive integers i, where β is again given by (A.4).

We now propose the ansatz

$$c_i(t) \simeq \rho\left(y = \varepsilon i, \tau = \frac{t}{t_0} \right), \tag{A.7}$$

where $\rho(y,\tau)$ is a smooth function. We define the parameter

$$\lambda = \frac{V_c}{V_\ell + V_c}. \tag{A.8}$$

In subsection A.2 we show that, given the ansatz (A.7), equations (A.6) reduce to the single PDE

$$\frac{\partial \rho}{\partial \tau} + \lambda f \frac{\partial \rho}{\partial y} \simeq \varepsilon \frac{\lambda |f|}{2} \left(\frac{\partial^2 \rho}{\partial y^2} + \frac{2}{f} \frac{\partial^2 \rho}{\partial y \partial \tau} + \frac{1}{f^2} \frac{\partial^2 \rho}{\partial \tau^2} - \frac{f'}{f^3} \frac{\partial \rho}{\partial \tau} \right), \tag{A.9}$$

where ρ and its derivatives are evaluated in (y,τ) and f and f' are evaluated in τ (terms of higher order in ε are neglected).

Our next step, which we carry out in subsection A.3, is to show that, neglecting corrections of order ε^2, equation (A.9) reduces to

$$\frac{\partial \rho}{\partial \tau} + \lambda f \frac{\partial \rho}{\partial y} \simeq \varepsilon \frac{\lambda |f|}{2} (1-\lambda)^2 \frac{\partial^2 \rho}{\partial y^2}. \tag{A.10}$$

Finally, the asymptotic approximation of section 3 results from the two-time-scale analysis of subsection A.4.

A.1. Derivation of (A.4) and (A.5). Let $X(s)$ be the solution of

$$X'(s) = v(s) \quad \text{and} \quad X(t) = iL - \ell, \tag{A.11}$$

where X' is the derivative of X. Fix t and let $\beta > 0$. If $(i-1)L \le X(s) \le iL - \ell$ for all $s \in [t-\beta, t]$, then (2.1) implies that c is constant along the characteristic paths;

i.e., $c(X(s),s)$ is independent of s for $s \in [t-\beta, t]$. Assume that $v(s)$ is positive for all $s \in [t-\beta, t]$ and β is implicitly given by the equation

$$X(t-\beta) = (i-1)L. \tag{A.12}$$

We have that $c(iL-\ell, t) = c(X(t), t) = c(X(t-\beta), t-\beta) = c((i-1)L, t-\beta)$. On the other hand, since $v(t-\beta) > 0$, equation (2.3) implies that $c((i-1)L, t-\beta) = c_{i-1}(t-\beta)$. Thus,

$$c(iL-\ell, t) = c_{i-1}(t-\beta). \tag{A.13}$$

To compute β, we first use the expression for v given in (A.2) and expand v in powers of $(s-t)$ to get

$$v(s) \simeq \frac{V_F}{at_0} f\left(\frac{t}{t_0}\right) + \frac{V_F}{at_0^2} f'\left(\frac{t}{t_0}\right)(s-t). \tag{A.14}$$

We then integrate this approximation of v and use the condition $X(t) = iL - \ell$ to get

$$X(s) \simeq iL - \ell + \frac{V_F}{at_0} f\left(\frac{t}{t_0}\right)(s-t) + \frac{V_F}{2at_0^2} f'\left(\frac{t}{t_0}\right)(s-t)^2. \tag{A.15}$$

Next use this approximation of $X(s)$ in (A.12) to get

$$(i-1)L \simeq iL - \ell - \frac{V_F}{at_0} f\left(\frac{t}{t_0}\right)\beta + \frac{V_F}{2at_0^2} f'\left(\frac{t}{t_0}\right)\beta^2. \tag{A.16}$$

We subtract $(i-1)L$ on both sides of the above equation, then multiply by a/V_F, note that $a(L-\ell) = V_c$, and recall that $\varepsilon = V_c/V_F$ to get

$$0 \simeq \varepsilon - f\left(\frac{t}{t_0}\right)\frac{\beta}{t_0} + \frac{1}{2} f'\left(\frac{t}{t_0}\right)\left(\frac{\beta}{t_0}\right)^2. \tag{A.17}$$

Once we note that we are considering the case $v(t) > 0$, and thus $f(t/t_0) > 0$, equation (A.17) and elementary calculations show the validity of (A.4) for t such that $v(t) > 0$. The case $v(t) < 0$ results from a similar analysis.

A.2. Derivation of (A.8). Use the ansatz (A.7) and the expression (A.2) for the velocity in (A.6), multiply that equation by t_0/V_F, recall that $\varepsilon = V_c/V_F$, make the change of variables $\tau = t/t_0$, and define $b = \beta/t_0$ to get

$$\frac{V_\ell}{V_F}\frac{\partial \rho}{\partial \tau}(i\varepsilon, \tau) = \begin{cases} f(\tau)\left(\rho((i-1)\varepsilon, \tau - b) - \rho(i\varepsilon, \tau)\right) & \text{if } f(\tau) > 0, \\ f(\tau)\left(\rho(i\varepsilon, \tau) - \rho((i+1)\varepsilon, \tau - b)\right) & \text{if } f(\tau) < 0, \end{cases} \tag{A.18}$$

where

$$b = \varepsilon \frac{1}{|f(\tau)|} + \varepsilon^2 \frac{f'(\tau)}{2f^3(\tau)} + O(\varepsilon^3). \tag{A.19}$$

Expanding in powers of ε, we have that, when $f(\tau) > 0$,

$$\rho((i-1)\varepsilon, \tau - b) \simeq \rho - \varepsilon\left(\frac{\partial \rho}{\partial y} + \frac{1}{f}\frac{\partial \rho}{\partial \tau}\right) + \frac{\varepsilon^2}{2}\left(\frac{\partial^2 \rho}{\partial y^2} + \frac{2}{f}\frac{\partial^2 \rho}{\partial y \partial \tau} + \frac{1}{f^2}\frac{\partial^2 \rho}{\partial \tau^2} - \frac{f'}{f^3}\frac{\partial \rho}{\partial \tau}\right).$$

Analogously, when $f(\tau) < 0$, we have

$$\rho((i+1)\varepsilon, \tau - b) \simeq \rho + \varepsilon\left(\frac{\partial \rho}{\partial y} + \frac{1}{f}\frac{\partial \rho}{\partial \tau}\right) + \frac{\varepsilon^2}{2}\left(\frac{\partial^2 \rho}{\partial y^2} + \frac{2}{f}\frac{\partial^2 \rho}{\partial y \partial \tau} + \frac{1}{f^2}\frac{\partial^2 \rho}{\partial \tau^2} - \frac{f'}{f^3}\frac{\partial \rho}{\partial \tau}\right).$$

In the last two equations ρ and its derivatives are evaluated in $(\varepsilon i, \tau)$, and f and f' are evaluated in τ. Once we plug these expressions into (A.18) and perform simple algebraic manipulations, we obtain (A.9).

A.3. Derivation of (A.10). From (A.9) we infer that

$$\frac{\partial \rho}{\partial \tau} = -\lambda f \frac{\partial \rho}{\partial y} + O(\varepsilon). \tag{A.20}$$

Taking derivatives with respect to y in (A.20), we get

$$\frac{\partial^2 \rho}{\partial y \partial \tau} = -\lambda f \frac{\partial^2 \rho}{\partial y^2} + O(\varepsilon). \tag{A.21}$$

On the other hand, taking derivatives with respect to τ in (A.20) and using (A.21), we have

$$\frac{\partial^2 \rho}{\partial \tau^2} = -\lambda f \frac{\partial^2 \rho}{\partial y \partial \tau} - \lambda f' \frac{\partial \rho}{\partial y} + O(\varepsilon) = \lambda^2 f^2 \frac{\partial^2 \rho}{\partial y^2} - \lambda f' \frac{\partial \rho}{\partial y} + O(\varepsilon). \tag{A.22}$$

Once we replace $\partial \rho / \partial \tau$, $\partial^2 \rho / \partial y \partial \tau$, and $\partial^2 \rho / \partial \tau^2$ in the right-hand side of (A.9) by the expressions obtained in the last three equations and neglect terms of order ε^2, we obtain (A.10).

A.4. Two-time-scale analysis on (A.10). We now follow the standard procedures in two-time-scale asymptotics. We introduce a second time scale

$$\theta = \varepsilon \lambda (1-\lambda)^2 \tau. \tag{A.23}$$

We need to replace $\partial \rho / \partial \tau$ by $\varepsilon \lambda (1-\lambda)^2 \partial \rho / \partial \theta + \partial \rho / \partial \tau$ in (A.10), treat τ and θ as independent variables, and assume that ρ depends on the three variables y, τ, and θ; i.e., $\rho = \rho(y, \tau, \theta)$. Equation (A.10) becomes

$$\varepsilon \lambda (1-\lambda)^2 \frac{\partial \rho}{\partial \theta} + \frac{\partial \rho}{\partial \tau} + \lambda f \frac{\partial \rho}{\partial y} = \varepsilon \frac{\lambda |f|}{2} (1-\lambda)^2 \frac{\partial^2 \rho}{\partial y^2}. \tag{A.24}$$

Next we expand ρ in powers of ε,

$$\rho = \rho_0 + \varepsilon \rho_1 + \varepsilon^2 \rho_2 + \cdots, \tag{A.25}$$

and require that $\rho_1 = \rho_1(y, \tau, \theta)$ be periodic (with period 1) in τ (this requirement makes the asymptotic approximation valid for long values of τ). We then plug this ansatz into (A.24) and collect powers of ε. At order 1 we get

$$\frac{\partial \rho_0}{\partial \tau} + \lambda f \frac{\partial \rho_0}{\partial y} = 0, \tag{A.26}$$

and at order ε

$$\lambda (1-\lambda)^2 \frac{\partial \rho_0}{\partial \theta} + \frac{\partial \rho_1}{\partial \tau} + \lambda f \frac{\partial \rho_1}{\partial y} = \frac{\lambda |f|}{2} (1-\lambda)^2 \frac{\partial^2 \rho_0}{\partial y^2}. \tag{A.27}$$

From (A.26), we obtain that the dependence of ρ_0 on y and τ is through the variable η defined as

$$\eta = y - \lambda \int_0^\tau f(s)\, \mathrm{d}s. \tag{A.28}$$

Thus, if we change independent variables from (y, τ, θ) to (η, τ, θ), we have

$$\rho = \rho(\eta, \theta) \quad \text{and} \quad \rho_1 = \rho_1(\eta, \tau, \theta). \tag{A.29}$$

In the new independent variables, (A.27) becomes

$$\lambda(1-\lambda)^2\frac{\partial \rho_0}{\partial \theta}+\frac{\partial \rho_1}{\partial \tau}=\frac{\lambda|f|}{2}(1-\lambda)^2\frac{\partial^2 \rho_0}{\partial \eta^2}. \tag{A.30}$$

Finally we take the average of the above equation with respect to τ, keeping η and θ fixed. Since ρ_1 is periodic in τ, we have

$$\int_0^1 \frac{\partial \rho_1}{\partial \tau}\,\mathrm{d}\tau = 0. \tag{A.31}$$

Recalling the definition of f and its properties (see (A.2) and (A.3)), we have $\int_0^1 |f|\,\mathrm{d}\tau = 2$. Thus, given that ρ_0 is independent of τ, we have that, after averaging with respect to τ, (A.30) becomes

$$\frac{\partial \rho_0}{\partial \theta}=\frac{\partial^2 \rho_0}{\partial \eta^2}. \tag{A.32}$$

We define the spatial variable

$$z=\frac{L}{\varepsilon}y-\frac{aL}{V_\ell+V_c}\int_0^t v(s)\,\mathrm{d}s. \tag{A.33}$$

From the different changes of variables made, it follows that

$$\eta=\frac{V_c}{V_F L}z \quad \text{and} \quad \theta=\frac{V_c^2 V_\ell^2}{V_F(V_\ell+V_c)^3}\frac{t}{t_0}. \tag{A.34}$$

The last three equations imply that ρ_0 satisfies (3.3). Thus, dropping the subindex 0 and observing the appropriate boundary and initial conditions, we obtain the asymptotic approximation of section 3.

Acknowledgments. The author thanks Professor Santamarina for introducing the author to this area of research and for stimulating discussions. This work was motivated by the work in [17].

REFERENCES

[1] R. Aris, *On the dispersion of solute matter in a fluid flowing through a tube*, Proc. Roy. Soc. London A, 235 (1956), pp. 67–77.

[2] R. A Chole and S. P. Tinling, *Incomplete coverage of mammalian bone cell matrix by lining cells*, Ann. Ontology Rhinology and Laryngology, 102 (1993), pp. 543–550.

[3] R. R. Cooper, J. W. Milgram, and R. A. Robinson, *Morphology of the osteon*, J. Bone Jt. Surg., 48 (1966), pp. 1239–1271.

[4] S. Cowin, *Bone poroelasticity*, J. Biomech., 32 (1999), pp. 217–238.

[5] M. L. Knothe Tate, U. Knothe, and P. Niederer, *Experimental elucidation of mechanical load-induced fluid flow and its potential role in bone metabolism and functional adaptation*, Am. J. Med. Sci., 316 (1998), pp. 189–195.

[6] M. L. Knothe Tate and U. Knothe, *An ex vivo model to study the transport processes and fluid flow in loaded bone*, J. Biomech., 33 (2000), pp. 247–254.

[7] M. L. Knothe Tate and P. Niederer, *A theoretical FE-based model developed to predict the relative contribution of convective and diffusive mechanisms for the maintenance of local equilibria within cortical bone*, in Advances in Heat and Mass Transfer in Biotechnology, S. Clegg, ed., American Society of Mechanical Engineers, New York, 1998, pp. 133–142.

[8] M. L. Knothe Tate, P. Niederer, and U. Knothe, *In vivo tracer transport through the lacunocanalicular system of rat bone in an environment devoid of mechanical loading*, Bone, 22 (1998), pp. 107–117.

[9] M. L. Knothe Tate, *Mixing mechanisms and net solute transport in bone*, Ann. Biomed. Eng., 29 (2001), pp. 810–811.
[10] M. L. Knothe Tate, *Interstitial fluid flow*, in Bone Biomechanics Handbook, S.C. Cowin, ed., CRC Press, New York, 2001, Vol. 23, pp. 1–29.
[11] R. H. Kufahl and S. Saha, *A theoretical model for stress-generated fluid flow in the canaliculi-lacunae network in bone tissue*, J. Biomech., 23 (1990), pp. 171–180.
[12] G. Li, J. T. Bronk, K. N. An, and P. J. Kelly, *Permeability of cortical bone of canine tibiae*, Microcirculation Res., 34 (1987), pp. 302–310.
[13] S. C. Miller and W. S. S. Jee, *Bone lining cells*, in Bone, Vol. 4: Bone Metabolism and Mineralization, K. V. Hall, ed., CRC Press, Boca Raton, 1992, pp. 1–19.
[14] W. F. Neuman and M. W. Neuman, *The chemical dynamics of bone*, University of Chicago Press, Chicago, 1958.
[15] K. Piekarski and M. Munro, *Transport mechanism operating between blood supply and osteocytes in long bones*, Nature, 269 (1977), pp. 80–82.
[16] G. I. Taylor, *Dispersion of soluble matter in solvent flowing slowly through a tube*, Proc. Roy. Soc. London A, 223 (1953), pp. 446–468.
[17] L. Wang, S. C. Cowin, S. Weinbaum, and S. P. Fritton, *Modeling tracer transport in an osteon under cyclic loading*, Ann. Biomed. Eng., 28 (2000), pp. 1200–1209.
[18] L. Wang, S. C. Cowin, S. Weinbaum, and S. P. Fritton, *In response to "Mixing mechanisms and net solute transport in bone" by M.L. Knothe Tate*, Ann. Biomed. Eng., 29 (2001), pp. 812–816.

SIAM J. APPL. MATH.
Vol. 65, No. 6, pp. 2141–2157

FIELD-INDUCED MOTION OF NEMATIC DISCLINATIONS*

PAOLO BISCARI† AND TIMOTHY J. SLUCKIN‡

Abstract. An individual defect in a nematic liquid crystal moves not only in response to its interaction with other defects but also in response to external fields. We analyze the motion of a wedge disclination in the presence of an applied field of strength H. We neglect backflow and seek steadily traveling patterns. The stationary picture yields a semi-infinite wall of strength π, bounded by the defect line. We find that the disclination advances into the region containing the wall at velocity $v(H)$, where v scales as $H/|\log H|$ as long as the magnetic coherence length is greater than the core radius. When the external field is applied in the presence of a pair of disclinations, their dynamics is strongly influenced. We compute the expected relative velocity of the disclinations as a function of distance and field. The natural tendency for the disclinations to annihilate each other can be overcome by a sufficiently strong field suitably directed.

Key words. nematic liquid crystals, dynamics, π-walls

AMS subject classifications. 76A15, 82D30

DOI. 10.1137/040618898

1. Introduction.

Singularities in liquid crystals, or *defects*, have played a fundamental role not only in the development of the understanding of the physics of liquid crystals but also in the later development of the topological theory of defects in condensed matter. In nematic liquid crystals, point, line, and wall defects can be found. Line defects were first classified by Sir Charles Frank, who noted that line defects came in classes with an integer or half-integer charge [1]. Later workers, using the topological theory of defects, realized that nematic liquid crystals sustain a topologically unique line defect [2, 3, 4]. However, Frank's naïve classification, which effectively supposes that the nematic order parameter is restricted to a plane, remains important in providing guidance and intuition for the physics of defects in nematic liquid crystals.

The topological total charge of a system is conserved during its evolution. Under many circumstances this remains true for Frank's definition of charge; this is a stronger condition. For instance, defects of opposite charges may annihilate, and defects of higher topological charges may decay to a bunch of defects of smaller topological charges. Topological dipoles may even be nucleated from a smooth field [5].

In this paper we focus on nematic disclinations, i.e., line defects. According to Frank's definition, the topological charge of a line defect is defined as the number of turns the director performs along a closed path surrounding the defect. This number may be half-integer because of the head-and-tail symmetry of nematic liquid crystal molecules. When the final director is rotated by an angle π with respect to the initial director, the physical state they describe does not exhibit any discontinuity.

The physical manifestation of the topological theory of defects in nematic liquid crystals concerns *escape into the third dimension*. Nematic disclinations of integer

*Received by the editors November 15, 2004; accepted for publication (in revised form) March 25, 2005; published electronically August 9, 2005. This research was supported by a British Council Anglo-Italian collaborative grant.

http://www.siam.org/journals/siap/65-6/61889.html

†Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy (paolo.biscari@polimi.it).

‡Faculty of Mathematical Studies, University of Southampton, Southampton SO17 1BJ, UK (t.j.sluckin@soton.ac.uk).

charge are not topologically stable. A suitable continuous transformation, involving the third previously neglected dimension, dissolves the singularity and leaves behind a regular field [6]. In a similar manner, all half-integral disclinations can be distorted into each other. However, free energy criteria often impede this process, and the Frank classification remains useful. In particular, annihilation of opposite-charge half-integral defects is favored, but of same-charge defects is hindered.

Defect dynamics has been widely studied either in the classical Oseen–Zocher–Frank (OZF) theory [7, 8, 1], or within the extended de Gennes–Ericksen theory [9, 10]. However, there are fundamental dilemmas in treating defect motion. The OZF theory is not a dynamical theory, and defect cores are regions in which, strictly speaking, the OZF theory does not apply. The theory can be extended to deal with moving defects, but only by introducing a phenomenological dissipation function using director rate of change. This minimal extension is incorrect; the full extension involves hydrodynamic terms which depend on the local director [11]. But the local director is not defined in the defect region, and so this extension is also inadequate to describe defect motion. The alternative de Gennes–Ericksen approach (now usually called the **Q**-tensor theory) is in principle up to the task, but now the defect regions are no longer special. In addition, it is now necessary to consider a whole set of new degrees of freedom. In fact, however, these new degrees of freedom are important only inside the core region.

Despite these problems, some theoretical progress has been made, in part because some authors have detected analogous topological structures in liquid crystals and cosmological models. A single disclination may move through an otherwise smooth field [12, 13], but then the problem is why the defect is moving in the first place. An implicit response to this question is to devote more attention to the interaction between two or more defects. In such cases the defects move, slowly, to reduce their elastic energy. For example, attraction between two point defects of opposite charges has been studied in both planar [14, 15, 16] and cylindrical [17, 18, 19] geometries. The attraction between two oppositely charged smectic disclinations has been studied in [20].

More precise quantitative descriptions of the defect evolution must necessarily take into account backflow effects [21, 22, 23], i.e., the interactions between director rotation and macroscopic molecular motion. The first analytical attempt to introduce backflow effects was performed in [24], where the macroscopic velocity field was coupled to the degree of orientation, though not to the director field. A series of recent numerical simulations [25, 26, 27, 28, 29] have determined the main effects of backflow coupling. The dynamical director patterns turn out to be strongly influenced in the final part of the annihilation process, but some effects may be noted even during all the defect evolution. In particular, the disclination speeds may be different [25], and recent numerical simulations suggest that positively charged disclinations can move almost twice as fast as those with negative charge [28].

In this paper we investigate the effect of an applied external field on the defect dynamics. We consider a simple geometry: a single $+\frac{1}{2}$, or a dipole of $\pm\frac{1}{2}$ disclinations, in planar symmetry. We compute the defect speed with the aid of the Leslie's dissipation balance [11]. It turns out that the external field exerts a profound effect on the defect dynamics and the defect interaction. By suitably adjusting the external field strength and direction, it is possible to drive a single disclination through the sample, as has been experimentally observed [30, 31]. More interestingly, the effect of the external field may go beyond a simple acceleration or deceleration of the annihilation process. The field can also stop the defects, or even reverse their veloc-

ities, thus transforming an attractive into a repulsive interaction. Depending on the external field direction, a critical defect distance may arise, which characterizes the defect interaction. They annihilate each other if they come closer than that distance; otherwise, they repel.

Throughout our presentation we will introduce and discuss some assumptions that simplify our analytical calculations. The 1-constant approximation in the elastic free energy, and a parabolic approximation in the magnetic energy, linearize the free energy derivatives. We remark that the parabolic approximation would have to be abandoned if we were to generalize the present study to highly charged defects. We also neglect backflow. This approximation allows us to derive an analytical condition which determines the defect velocity, and in particular the critical distance that reverses the defect interaction. A numerical analysis would correct these computed values, even if the described phenomena will certainly remain.

The plan of the paper is as follows. In section 2 we analyze the motion of a single disclination in an external field. Section 3 describes the defect interaction, and how it can be reversed with the aid of the applied field. In section 4 we summarize our results and compare them with the observed experimental data.

2. Single defect motion. We consider a $+\frac{1}{2}$ disclination embedded in an external magnetic field. This field will favor the director orientation of one side of the defect with respect to the other side. This asymmetry is sufficient, as we shall see, for energy considerations alone to determine that the defect will move and to determine the direction of its motion. Our task is to determine the magnitude of the velocity as a function of the field intensity. We adopt the 1-constant approximation and neglect backflow. This latter approximation implies that our estimated velocity will certainly be smaller than the actual velocity. In fact, backflow effects reduce the total dissipation, thus allowing faster director dynamics.

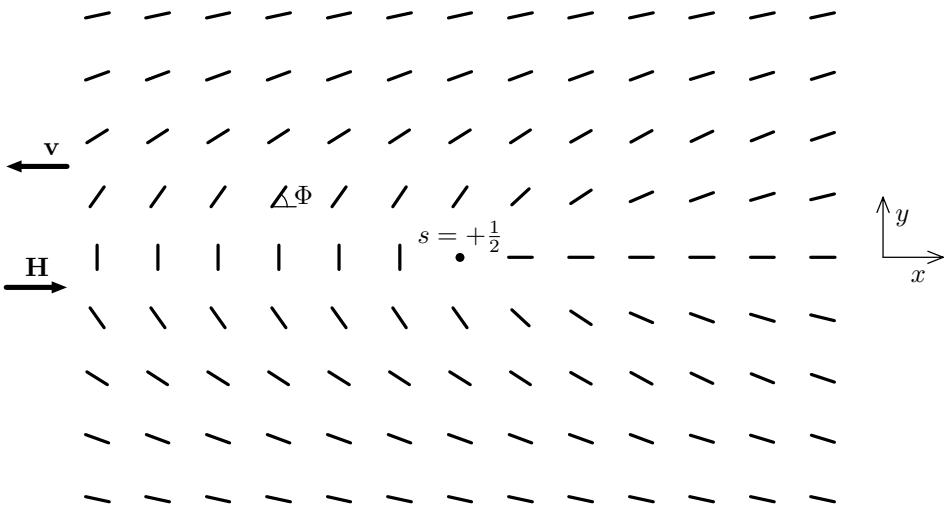


FIG. 1. *Geometric setting for the analysis of the single defect motion. The defect occupies the origin of the comoving reference frame and moves towards the π-wall. The angle Φ is determined by the director and the motion direction (parallel to the external field).*

Let us consider the reference frame illustrated in Figure 1. It moves with the disclination, which sits at the origin **O**. The x-axis is parallel to the external field.

Given any point **P** in the plane, let ϑ be the polar angle between the radial direction **OP** and the x-axis. Disclinations of half-integer charge do not benefit from escaping into the third dimension. Thus, we restrict our interest to the situation in which the director **n** is confined to the plane, and let $\Phi(\mathbf{P})$ be the angle between the nematic director at **P** and the x-axis.

In the absence of a field, the 1-constant approximation implies that the director angle varies linearly with the polar angle: $\Phi = \frac{1}{2}\,\vartheta$. In the presence of the field, the same linear dependence holds, roughly speaking, on length scales smaller than the magnetic coherence length ξ. So long as we confine ourselves to these length scales, the elastic energy overwhelms the external field. However, far from the defect, the magnetic energy forces the system toward its preferred value $\Phi = 0$. On the other hand, over a closed path around the defect, no matter how far that path may be from the defect, topology *forces* the angle Φ to rotate through π. This creates a dilemma, for the system must rotate through an angle π and yet remain at $\Phi = 0$.

These two constraints are reconciled by restricting the region over which Φ rotates. This region is a topologically irreducible *wall* between regions of space whose director is oppositely oriented. Associated with the wall is a well-defined surface free energy analogous to the surface tensions of fluid mechanics. This is a π-wall, since the director concentrates its π-rotation in it. In the presence of the field, the previously isolated defect line has been transformed into the trailing edge of a wall defect. The disclination then moves into the wall in order to reduce the wall area and consequently the free energy of the system.

2.1. Dissipation principle. We determine both the director field and the value of the disclination velocity by imposing the energy balance between the free-energy loss-rate and the dissipation [11]:

$$\dot{\mathcal{F}} + \mathcal{D} = 0. \tag{2.1}$$

In the 1-constant approximation, the free-energy functional is given by

$$\mathcal{F}[\Phi] = \int_{\mathbb{R}^2} \left(K \left| \nabla \Phi \right|^2 + \chi_{\mathrm{a}} H^2 \sin^2 \Phi \right) da, \tag{2.2}$$

where K is an average elastic constant, χ_{a} is the magnetic anisotropy, and H is the strength of an external magnetic field.

It is well known that a $+\frac{1}{2}$ disclination possesses an infinite core energy. The elastic free-energy density diverges at the defect, and the singularity is not integrable. There are two ways to avoid this divergence. The first consists of excluding from our integration region a small disk centered in the defect, the *core region* $\mathcal{B}_\circ$. The radius of the excised disk, the *core radius* $r_\circ$, is usually much smaller than all other characteristic lengths entering the problem, so that many studies have been performed in the limit of vanishing $r_\circ$ [32]. A more precise physical description of the defect requires an extension of the classical OZF theory, and the replacement of the nematic director with the nematic order tensor [33]. We will choose the first option and perform all the integrations in the pierced domain, which excludes $\mathcal{B}_\circ$. We further assume that the core radius is fixed. The basic physics of the phenomenon is well described using these approximations. However, in order to deal with external fields of any intensity, it would be interesting to apply the techniques of [34] to determine how and when the magnetic coherent length influences the core radius.

A second, though less worrying, free-energy divergence comes from the supposedly infinite size of the domain. The domain may extend indefinitely in the y-direction

without inducing any singularity, since both the elastic and the magnetic energy densities vanish away from the x-axis. On the other hand, there will be few cases in which a large-x cut will be needed to keep energies finite. In those cases we will assume that our domain is bounded by $|x| \leq L$. Whenever possible, we will perform the $L \to +\infty$ limit, and we will notice that the large scale length L will not finally enter in the defect velocity.

Our final approximation concerns the magnetic free energy. We will perform the *parabolic approximation* $\sin^2 \Phi \simeq \Phi^2$. This approximation allows us to obtain linear field equations in Φ. It can be used in the whole domain, provided that we define $\Phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, since in that case all the values attained by the director angle belong to the potential well of the equilibrium configuration $\Phi = 0$. We remark that this approximation would not be valid if we were interested in analyzing the equilibrium configurations of more complex defects.

When we neglect backflow, the dissipation assumes the simple expression

$$\mathcal{D} = \gamma_1 \int_{\mathcal{B}_\circ} \dot{\Phi}^2 \, dx \, dy, \tag{2.3}$$

where the pierced integration domain $\mathcal{B}_\circ$ comes into play since the dissipation density also diverges in the core region.

Let us perform an infinitesimal displacement $\delta\Phi$ of the director field. If we make use of the divergence theorem, the dissipation principle (2.1) requires

$$\int_{\partial\mathcal{B}_\circ} \delta\Phi \, \left(2K \, \nabla\Phi\right) \cdot \nu \, d\ell + \int_{\mathcal{B}_\circ} \delta\Phi \, \left[\gamma_1 \, \dot{\Phi} - 2K\Delta\Phi + 2\chi_{\mathrm{a}} H^2 \Phi\right] da = 0 \quad \forall \, \delta\Phi, \tag{2.4}$$

where ν is the outer normal along $\partial\mathcal{B}_\circ$. The arbitrariness of $\delta\Phi$ requires that the quantity in square brackets in the second integral must vanish identically:

$$\gamma_1 \, \dot{\Phi} = 2K\Delta\Phi - 2\chi_{\mathrm{a}} H^2 \, \Phi. \tag{2.5}$$

Equation (2.5) is the well-known time-dependent Ginzburg–Landau evolution equation of the system. It will supply us the director field. However, the Ginzburg–Landau equation alone is not able to guarantee the dissipation principle. We must check that the first integral in (2.4) also vanishes. The boundary $\partial\mathcal{B}_\circ$ is made of two parts: a small circle around the defect and a large boundary at infinity. The integral around the latter vanishes, since at infinity $\nabla\Phi$ fades everywhere except along the π-wall, where it is orthogonal to the outward normal. We are then left with the first integral in (2.4), performed along the boundary of the core region. This quantity has a simple physical meaning [35]: it is the power supplied by the core region to the rest of the domain. Thus, to require that all the free-energy loss be dissipated within the system is equivalent to requiring that no energy be supplied to it, either from the outside or from the core region.

2.2. Steadily moving defects. We will look for stationary solutions of equation (2.5). They aim at representing a defect moving at a constant speed v towards the π-wall. We will find that, for any positive value of v, it is possible to find a solution of (2.5) which satisfies the boundary conditions. However, there is just one value of the velocity which also cancels the first integral in (2.4) (or, equivalently, that satisfies the global dissipation condition). It is the only velocity at which the defect is able to move without any external boost.

In a steadily moving reference frame, traveling with the defect itself, the relative stationary differential equation follows by replacing the time derivative in (2.5) with $v\,\partial_x$. We thus obtain

$$\Delta\Phi - \frac{\lambda}{\xi}\,\frac{\partial\Phi}{\partial x} - \frac{\Phi}{\xi^2} = 0, \tag{2.6}$$

where $\xi := \sqrt{\frac{K}{\chi_a H^2}}$ is the magnetic coherence length, and $\lambda := \frac{\gamma_1}{2\sqrt{K\chi_a}}\,\frac{v}{|H|}$ is a dimensionless quantity. We look for solutions of the eigenvalue problem (2.6) that satisfy the symmetry requirement $\Phi(x,-y) = -\Phi(x,y)$ for all $y \neq 0$, and the boundary conditions

$$\Phi(x,0^+) = \begin{cases} 0 & \text{if } x > 0, \\ \dfrac{\pi}{2} & \text{if } x < 0, \end{cases} \qquad \text{and} \qquad \lim_{y\to\infty} \Phi(x,y) = 0 \quad \forall\, x \in \mathbb{R}, \tag{2.7}$$

where the latter condition is determined by the presence of the magnetic field. The boundary conditions (2.7) are singular only at the origin, where a disclination of topological charge $+\frac{1}{2}$ stands. Indeed, the discontinuity that the angle Φ suffers along the negative x-axis does not induce any physical singularity, since $\Phi = \pi/2$ and $\Phi = -\pi/2$ describe the same director orientation.

Among the solutions of the eigenvalue problem (2.6)–(2.7), the dissipation principle (2.1) will single out the only physical one.

2.3. Director field. We solve the linear partial differential equation (2.6) with a Fourier transform. To this end, it is useful to write the first of the boundary conditions (2.7) as

$$\Phi(x,0^+) = \frac{\pi}{4} - \frac{1}{4\mathrm{i}}\ \mathrm{PV}\!\int_{\mathbb{R}} \frac{\mathrm{e}^{\mathrm{i}qx}}{q}\,dq \qquad \forall\, x \neq 0, \tag{2.8}$$

where PV denotes the Cauchy principal value of an integral. We then look for solutions of (2.6) of the form

$$\Phi(x,y) = \frac{\pi}{4}\,g_1(y) - \frac{1}{4\mathrm{i}}\ \mathrm{PV}\!\int_{\mathbb{R}} \frac{\mathrm{e}^{\mathrm{i}qx}}{q}\,g_2(q,y)\,dq, \tag{2.9}$$

with

$$g_1(0) = g_2(q,0) = 1 \quad \text{and} \quad \lim_{y\to\infty} g_1(y) = \lim_{y\to\infty} g_2(q,y) = 0 \quad \forall\, q \in \mathbb{R}. \tag{2.10}$$

If we insert (2.9) into (2.6), we obtain

$$\frac{\pi}{4}\left(g_1'' - \frac{1}{\xi^2}\,g_1\right) - \frac{1}{4\mathrm{i}}\ \mathrm{PV}\!\int_{\mathbb{R}} \frac{\mathrm{e}^{\mathrm{i}qx}}{q}\left[\frac{\partial^2 g_2}{\partial y^2} - k^2(q)\,g_2\right] dq = 0, \tag{2.11}$$

where $k(q)$ will henceforth denote the positive-real-part square root of $k^2(q) = q^2 + \frac{\mathrm{i}q\lambda}{\xi} + \frac{1}{\xi^2}$. The solution of (2.11) and (2.10) in the upper half-plane $\{y \geq 0\}$ is

$$\Phi(x,y) = \frac{\pi}{4}\,\mathrm{e}^{-\frac{y}{\xi}} - \frac{1}{4\mathrm{i}}\ \mathrm{PV}\!\int_{\mathbb{R}} \frac{\mathrm{e}^{\mathrm{i}qx - k(q)y}}{q}\,dq, \tag{2.12}$$

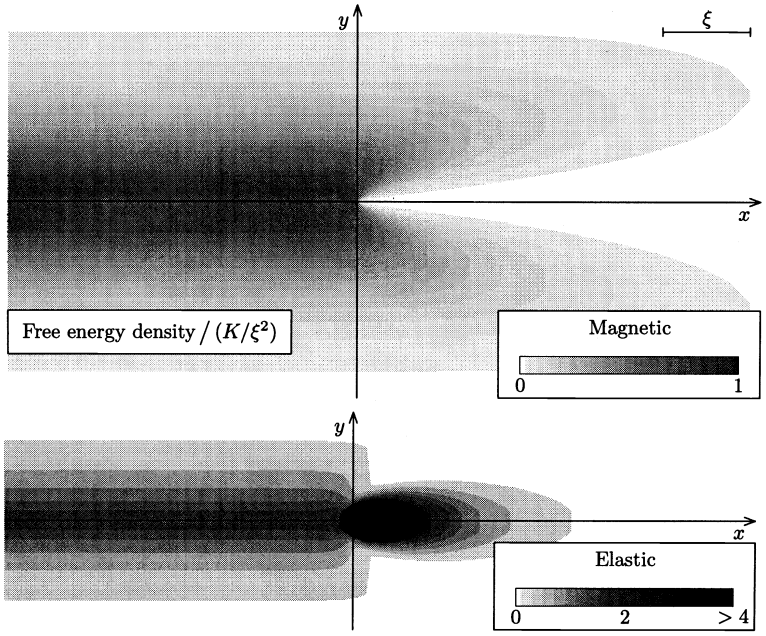


FIG. 2. *Density plots of the magnetic (upper graph) and elastic (lower graph) parts of the free-energy density, computed from the analytical solution* (2.13). *The elastic energy is localized close to the defect. In contrast, the magnetic energy is mostly packed around the π-wall. The defect core is not symmetric with respect to the defect position; it is slightly shifted behind it with respect to the direction of motion.*

whereas, by symmetry, the solution in the whole plane is given by

$$\Phi(x,y) = \varepsilon(y)\left[\frac{\pi}{4}\,\mathrm{e}^{-\frac{|y|}{\xi}} - \frac{1}{4\mathrm{i}}\ \mathrm{PV}\!\int_{\mathbb{R}} \frac{\mathrm{e}^{\mathrm{i}qx-k(q)|y|}}{q}\,dq\right], \tag{2.13}$$

with

$$\varepsilon(y) := \begin{cases} 1 & \text{if } y \geq 0, \\ -1 & \text{if } y < 0. \end{cases} \tag{2.14}$$

Figure 2 illustrates the solution (2.13). More precisely, in the upper plot it is possible to observe how the magnetic free-energy density is concentrated on the π-wall, whereas the lower plot shows that the elastic free-energy density is mostly localized on the defect. In both cases, it is possible to notice that the decay pattern from $\Phi = \pm\frac{\pi}{2}$ on the left x-axis to the equilibrium value $\Phi = 0$ becomes constant when we move some magnetic coherence lengths away from the defect.

We notice that no use of the core region has been made for the time being. In fact, the differential problem (2.6)–(2.7) admits a solution in $\mathbb{R}^2 \setminus (0,0)$, without any need to excise a finite region around the defect. We will see below that this will not be the case when we have to deal with the derivatives of the field, and in particular with the dissipation (2.3).

2.4. Disclination velocity. We have succeeded in finding a solution of the stationary Ginzburg–Landau equation for any value of v. We will now complete the

eigenvalue analysis and determine the correct disclination velocity by imposing the global dissipation condition (2.1).

We have already noticed that, in the moving reference frame, time derivatives transform into spatial derivatives. Thus,

(2.15)
$$\dot{\mathcal{F}} = v \int_{\mathcal{B}_\circ} \frac{\partial}{\partial x} \left(K \left| \nabla \Phi \right|^2 + \chi_{\mathrm{a}} H^2 \Phi^2 \right) dx\, dy = v \left[\int_{\mathbb{R}} \left(K \left| \nabla \Phi \right|^2 + \chi_{\mathrm{a}} H^2 \Phi^2 \right) dy \right]_{x=-\infty}^{x=+\infty} .$$

When $|x|$ is very large, the principal value of the q-integral in (2.13) is dominated by the low q values:

$$\lim_{x \to \pm\infty} \frac{1}{4\mathrm{i}} \ \mathrm{PV} \int_{\mathbb{R}} \frac{\mathrm{e}^{\mathrm{i}qx - k(q)y}}{q}\, dq = \frac{1}{4\mathrm{i}} \lim_{x \to \pm\infty} \ \mathrm{PV} \int_{\mathbb{R}} \frac{\mathrm{e}^{\mathrm{i}qx - \frac{y}{\xi}}}{q}\, dq = \pm \frac{\pi}{4}\, \mathrm{e}^{-\frac{y}{\xi}} . \tag{2.16}$$

Thus,

$$\dot{\mathcal{F}} = -\frac{\pi^2 K v}{\xi^2} \int_0^{+\infty} \mathrm{e}^{-\frac{2y}{\xi}}\, dy = -\frac{\pi^2 K v}{2\xi} . \tag{2.17}$$

The result (2.17) admits a simple physical interpretation, which already exhibits in (2.15). The quantity within square brackets in (2.15) is the free energy contained in an infinite vertical strip of unit width, centered at x. When $x \to +\infty$, both the elastic and magnetic energy densities relax to 0, as Figure 2 shows. As a consequence, the total energy stored in the right-side strip is negligible. The picture completely changes in the $x \to -\infty$ limit. A unit-width strip crossing the π-wall stores a finite amount of free energy, measured in (2.17). In fact, it is precisely the difference between the energies stored in those strips that keeps the defect moving. In unit time, the defect motion replaces a left-side strip of width v, with free energy given in (2.17), with a right-side strip of equal width, with no free energy.

While the free-energy variation depends only on the asymptotic structure of the director field, dissipation takes place in the whole domain. Indeed, it is so strong close to the moving defect that we will be forced to exclude the core region in order to avoid an infinite dissipation. We have

$$\begin{aligned}
\mathcal{D} &= \gamma_1 \int_{\mathbb{R}^2} \dot{\Phi}^2\, dx\, dy = \gamma_1 v^2 \int_{\mathbb{R}^2} \left(\frac{\partial \Phi}{\partial x} \right)^2 dx\, dy \\
&= \frac{\gamma_1 v^2}{16} \int_{\mathbb{R}^2} dx dy \ \mathrm{PV} \int_{\mathbb{R}} dq \ \mathrm{PV} \int_{\mathbb{R}} dq' \mathrm{e}^{\mathrm{i}(q+q')x - (k(q)+k(q'))|y|} \\
&= \frac{\gamma_1 v^2}{8} \ \mathrm{PV} \int_{\mathbb{R}} dq \ \mathrm{PV} \int_{\mathbb{R}} dq' \, \frac{2\pi \delta\big(q+q'\big)}{k(q)+k(q')} \\
&= \frac{\pi \gamma_1 v^2}{2} \int_0^{\frac{q_{\mathrm{M}}}{2\pi}} \frac{dq}{\sqrt{q^2 + \mathrm{i}q\lambda/\xi + 1/\xi^2} + \sqrt{q^2 - \mathrm{i}q\lambda/\xi + 1/\xi^2}} \\
&= -\frac{\mathrm{i}\pi K v}{2\xi} \int_0^{\frac{\xi}{r_\circ}} \left(\sqrt{1 + \frac{\mathrm{i}\lambda}{s} + \frac{1}{s^2}} - \sqrt{1 - \frac{\mathrm{i}\lambda}{s} + \frac{1}{s^2}} \right) ds,
\end{aligned} \tag{2.18}$$

where δ denotes the Dirac delta function. The high-q cutoff is needed in order to avoid the logarithmic divergence which the disclination induces both in the free energy (but

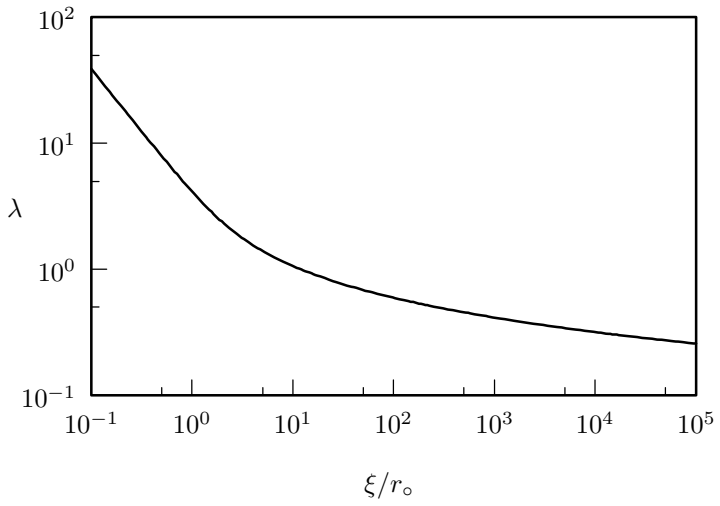


FIG. 3. *Disclination velocity as a function of the external applied field.*

not in its time-derivative) and in the dissipation. This is related to the inverse of the core radius: $q_{\mathrm{M}} = 2\pi/r_\circ$.

When we substitute (2.15) and (2.18) into the dissipation principle (2.1), we obtain the self-consistency equation that determines λ (i.e., v), as a function of the ratio $\xi/r_\circ$:

$$\int_0^{\frac{\xi}{r_\circ}} \left(\sqrt{1 + \frac{\mathrm{i}\lambda}{s} + \frac{1}{s^2}} - \sqrt{1 - \frac{\mathrm{i}\lambda}{s} + \frac{1}{s^2}} \right) ds = \mathrm{i}\pi. \tag{2.19}$$

The most interesting region in physical applications is $\xi \gg r_\circ$. The integral on the left-hand side is dominated by its logarithmic high-s divergence, and we obtain

$$\mathrm{i}\lambda \log \frac{\xi}{r_\circ} = \mathrm{i}\pi \quad \Longrightarrow \quad v = \frac{2\pi\sqrt{K\chi_{\mathrm{a}}}}{\gamma_1 \log\left(\xi/r_\circ\right)} \, |H| \qquad \text{when} \quad \xi \gg r_\circ. \tag{2.20}$$

However, it is interesting to push the analysis of (2.19) into the opposite regime, $\xi \ll r_\circ$. The low-s terms dominate the integral on the left-hand side, and we find

$$\mathrm{i}\lambda \, \frac{\xi}{r_\circ} = \mathrm{i}\pi \quad \Longrightarrow \quad v = \frac{2\pi\chi_{\mathrm{a}} r_\circ}{\gamma_1} \, H^2 \qquad \text{when} \quad \xi \ll r_\circ. \tag{2.21}$$

The quite intriguing quadratic behavior predicted by (2.21) must be handled carefully. When the external field becomes so intense that ξ becomes of the order of $r_\circ$, we have to question our assumption that $r_\circ$ is independent from ξ. A more complete theory, which can be derived following the steps of [34], would yield an $r_\circ(\xi)$, and thus a disclination velocity depending only on the strength of the applied field.

Figure 3 shows the numerical solution of (2.19). The plot highlights that the transition between the two asymptotic regimes derived above occurs for $\xi \gtrsim r_\circ$. Even though this is the limit up to which we can seriously trust our analytical result, Figure 3 suggests that the nonlinear effects increase the disclination velocity.

3. Defect interaction. Let us now consider two nematic defects of opposite topological charges $s = \pm\frac{1}{2}$ placed on the x-axis at a distance $\Delta = 2d$, but in the absence of any external field. After a certain period of time these defects amalgamate and thus annihilate each other. The two cores disappear at once, with a large and rapid reduction in free energy [14, 16, 28]. Imagine, however, that we apply a magnetic field that favors the director orientation of the molecules that lie within the defects. The defect speed will certainly be reduced. In this section we show that it can even be reversed. There exists a critical distance such that the defect interaction is attractive only if their mutual distance is smaller than the critical one. Otherwise, they repel.

Throughout this section, we will work out in detail the geometry illustrated in Figure 4. The applied field lies parallel both to the line connecting the defects and to the director orientation between them. In this geometry, the defects may only approach or separate, thus avoiding even more complicated motions such as mutual rotations.

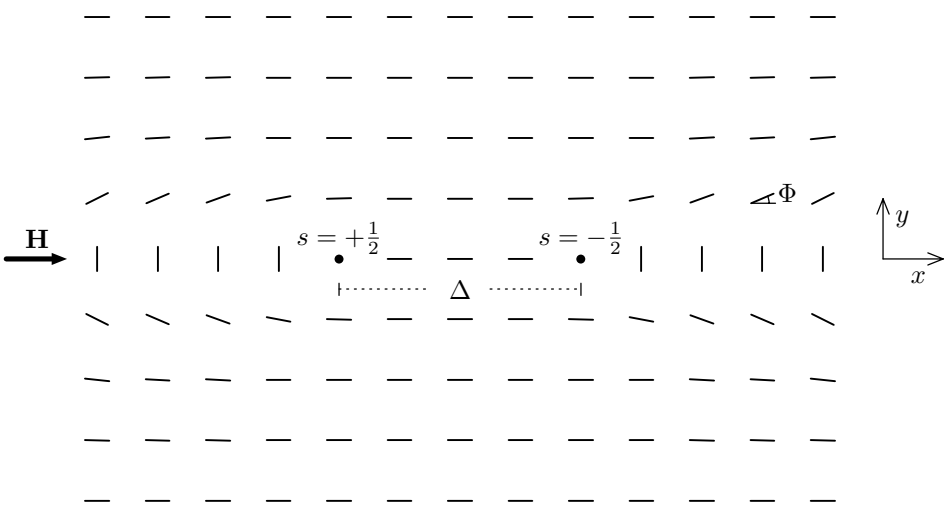


FIG. 4. *Two attracting defects placed at a distance Δ along the x-axis. The external field is parallel both to the line connecting the defects and to the director orientation between them.*

We compute the speed of two stationarily moving defects. We will show below that the stationarity assumption holds approximately even when the defect distance becomes of the order of, or smaller than, the magnetic coherence length. It must certainly be abandoned when one wants to describe the complete annihilation process, and in particular when the defect distance becomes of the order of the core radius.

In the final subsection we will briefly analyze the case when the external field determines a generic angle α with respect to the defect line. In this case the motion may be much more complicated. However, we are able to estimate how the nature of the defect interaction depends on α. More precisely, we will determine for which values of α the defect interaction may become repulsive, and how the critical distance depends on the external field direction.

The time-dependent Ginzburg–Landau equation (2.6) is linear, due to the parabolic approximation we used for the magnetic energy. We can thus obtain a solution describing two stationarily approaching (or separating) defects by simply superposing two functions of the type (2.13). More precisely, we add a solution describing a defect

placed at $x = +d$, traveling with velocity $-v$, and a solution describing a defect placed at $x = -d$, traveling with velocity $+v$. The velocity $v(d)$ will be determined later, again using the dissipation principle. The director field is given by

$$\Phi(x,y) = \varepsilon(y)\left[\frac{\pi}{2}\,\mathrm{e}^{-\frac{|y|}{\xi}} + \frac{1}{4\mathrm{i}}\ \mathrm{PV}\int_{\mathbb{R}} \frac{\mathrm{e}^{\mathrm{i}q(x-d)-k(q)|y|} - \mathrm{e}^{\mathrm{i}q(x+d)-k(-q)|y|}}{q}\,dq\right]$$

$$= \varepsilon(y)\left[\frac{\pi}{2}\,\mathrm{e}^{-\frac{|y|}{\xi}} - \int_0^{+\infty} \cos qx\ \sin\left(qd + \operatorname{Im} k(q)\,|y|\right)\,\mathrm{e}^{-\operatorname{Re} k(q)\,|y|}\,\frac{dq}{q}\right]. \tag{3.1}$$

Figure 5 illustrates the countervailing tendencies of the elastic and magnetic energies: the elastic contribution aims at annihilating the defects in order to relax the infinite core energy. In contrast, the magnetic field tries to broaden the intermediate region, where all the molecules are already correctly aligned.

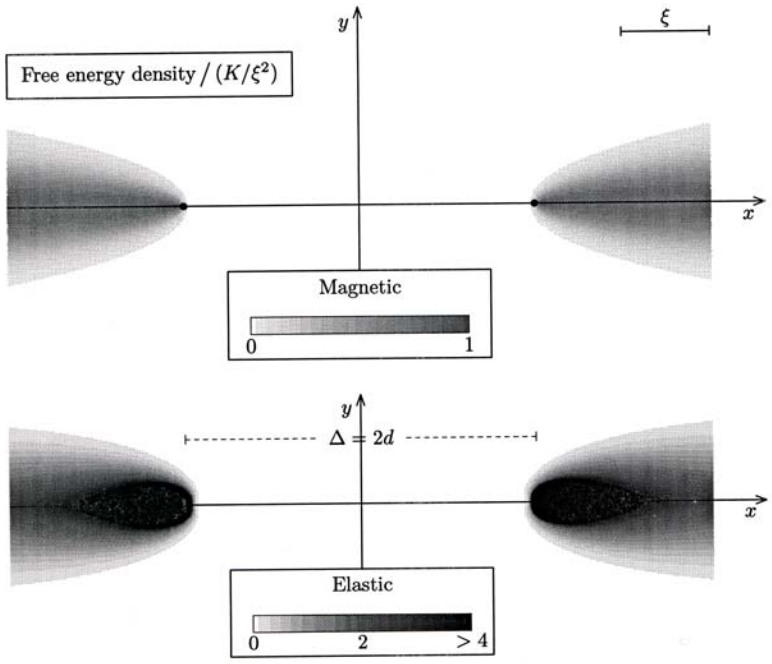


FIG. 5. *Density plot of the magnetic free-energy density (upper graph) and the elastic free-energy density (lower graph), computed from the analytical solution* (3.1). *For these particular values of the distance and the magnetic coherence length, it will turn out that the defects are separating (that is,* $v < 0$*). Both the magnetic and the elastic free-energy densities are mainly concentrated on the defect walls. Between the defects, the free-energy density is negligible.*

The free energy associated with the traveling configuration (3.1) is given by

$$\mathcal{F}(\Delta,\lambda) = K\pi\left[\frac{\pi L}{\xi} + \operatorname{arcsinh}\frac{\xi}{r_\circ} - K_0\left(\frac{\Delta}{\xi}\right) - \frac{\pi\Delta}{\xi} + 2\int_0^{\infty} \frac{\sin^2\frac{\Delta s}{2\xi}}{s^2\sqrt{1+s^2}}\,ds\right.$$

$$\left. + F_1(\Delta,\lambda) + F_2(\lambda)\right]. \tag{3.2}$$

The first term in (3.2) diverges in the $L \to +\infty$ limit. (L is the horizontal scale of the system: $x \in [-L, L]$.) This term corresponds to the energy of the two walls.

The second term is related to the infinite elastic energy stored in the defects. ($r_\circ$ is the core radius as in the preceding section.) The next three terms depend only on the defect distance, and are independent of their velocity. In particular, the term containing the modified Bessel function of the second kind K_0 is able to cancel the free-energy divergence when $\Delta \simeq r_\circ$. Finally, both F_1 and F_2 are velocity corrections, which vanish if λ (i.e., v) vanishes:

$$F_1(\Delta,\lambda) := \frac{1}{2}\int_0^\infty \mathrm{Re}\left[\left(\mu(\lambda,s)-\mu(0,s)+\frac{1}{\mu(\lambda,s)}-\frac{1}{\mu(0,s)}\right)\left(1-\mathrm{e}^{-\mathrm{i}s\Delta/\xi}\right)\right]\frac{ds}{s^2}$$
$$-\frac{1}{2}\int_0^\infty \mathrm{Re}\left[\left(\frac{1}{\mu(\lambda,s)}-\frac{1}{\mu(0,s)}\right)\mathrm{e}^{-\mathrm{i}s\Delta/\xi}\right]ds, \tag{3.3}$$

$$F_2(\lambda) := \int_0^\infty \left(\frac{1}{\mu(\lambda,s)+\mu(-\lambda,s)}-\frac{1}{2\mu(0,s)}\right)ds$$
$$-\frac{1}{4}\int_0^\infty \frac{\left(\mu(\lambda,s)-\mu(-\lambda,s)\right)^2}{\mu(\lambda,s)+\mu(-\lambda,s)}\left(1+\frac{1}{\mu(\lambda,s)\mu(-\lambda,s)}\right)\frac{ds}{s^2}, \tag{3.4}$$

where $\mu(\lambda,s)$ is the positive-real-part square root of $\mu^2(\lambda,s) = 1+\mathrm{i}s\lambda+s^2$, and $\lambda = \frac{\gamma_1}{2\sqrt{K\chi_\mathrm{a}}}\frac{v}{|H|}$, as in the preceding section.

In the absence of backflow, the dissipation stems only from the director rotation. When computing $\dot\Phi$, we assume that only the defect position d depends on time. In our stationary approximation, we thus neglect the time derivative of the velocity v. However, we remark that when we want to determine the nonstationary effects depending on $\dot v$, we are no longer allowed to use the solution (3.1). The dissipation function is given by

$$\mathcal{D} = \gamma_1\int_{\mathbb{R}^2}\dot\Phi^2\,dx\,dy = \frac{\pi\dot d^2}{2}\left[\mathrm{arcsinh}\,\frac{\xi}{r_\circ}+K_0\left(\frac{\Delta}{\xi}\right)+G_1(\Delta,\lambda)+G_2(\lambda)\right], \tag{3.5}$$

with

$$G_1(\Delta,\lambda) := \int_0^\infty \mathrm{Re}\left[\left(\frac{1}{\mu(\lambda,s)}-\frac{1}{\mu(0,s)}\right)\mathrm{e}^{-\mathrm{i}s\Delta/\xi}\right]ds \qquad \text{and} \tag{3.6}$$

$$G_2(\lambda) := 2\int_0^\infty \left(\frac{1}{\mu(\lambda,s)+\mu(-\lambda,s)}-\frac{1}{2\mu(0,s)}\right)ds. \tag{3.7}$$

The functions G_1 and G_2 vanish in the low-velocity limit $v\to 0$. The dissipation principle delivers the self-consistency equation that determines the defect velocity $v=-\dot d$:

$$\frac{\partial\mathcal{F}}{\partial d}\,\dot d+\mathcal{D}=0 \qquad \Longleftrightarrow$$
$$K_1\left(\frac{\Delta}{\xi}\right)+\frac{\Delta}{\xi}\int_0^\infty \frac{\sin s\;ds}{s\sqrt{\frac{\Delta^2}{\xi^2}+s^2}}+\xi\,\frac{\partial F_1}{\partial\Delta}$$
$$-\frac{\lambda}{2}\left[\mathrm{arcsinh}\,\frac{\xi}{r_\circ}+K_0\left(\frac{\Delta}{\xi}\right)+G_1(\Delta,\lambda)+G_2(\lambda)\right]=\pi. \tag{3.8}$$

Before analyzing the solutions $\lambda(\Delta)$ of (3.8), we want to stress the importance of the first two terms on its left-hand side. In fact, if we define

$$f(x) := K_1(x)+x\int_0^\infty \frac{\sin s}{s\sqrt{x^2+s^2}}\,ds, \tag{3.9}$$

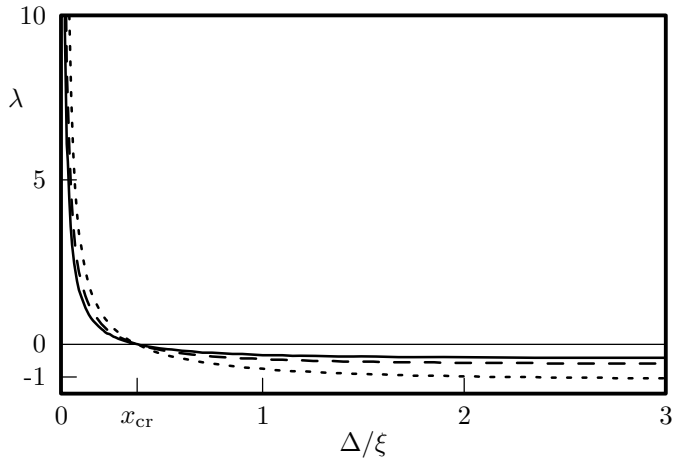


FIG. 6. *Defect velocity as a function of the external field for different values of the core radius* $r_\circ$: $\xi/r_\circ = 10$ *(dotted line),* 100 *(dashed),* 1000 *(full). For any value of the core radius, the defects attract only if their distance is smaller than* $\Delta_{\rm cr}$.

we have

$$\left.\frac{\partial \mathcal{F}}{\partial \Delta}\right|_{v=0} = \frac{K\pi}{\xi}\left[f\left(\frac{\Delta}{\xi}\right) - \pi\right]. \tag{3.10}$$

Thus, in general, the defects will approach or separate, depending on whether $f(\Delta/\xi)$ exceeds π or not. We postpone the analysis of the properties of f to the next subsection, when we will generalize (3.10) to the case of tilted applied fields. For now, we only remark that $f(x) = \pi$ when $x = x_{\rm cr} \doteq 0.377388$. The function f is greater than π (thus inducing defect attraction) when $\Delta < x_{\rm cr}\,\xi$. Defect repulsion is induced at distances greater than $x_{\rm cr}\,\xi$.

Figure 6 illustrates the numerical solutions of (3.8) for three different values of the ratio between the magnetic coherence length and the core radius. They exhibit the following properties:

- In the large distance limit, all Bessel functions decay exponentially with Δ/ξ. The functions G_1, G_2, and the derivative of F_1 vanish too. Furthermore,

$$\lim_{x\to\infty} x\int_0^\infty \frac{\sin s}{s\sqrt{x^2+s^2}}\,ds = \int_0^\infty \frac{\sin s}{s}\,ds = \frac{\pi}{2}\,. \tag{3.11}$$

  Thus, the large-distance limit of λ is given by

$$\lim_{\Delta\to\infty} \lambda(\Delta) = -\frac{\pi}{\operatorname{arcsinh}(\xi/r_\circ)} \quad\Longrightarrow\quad v \simeq -\frac{2\pi\sqrt{K\chi_{\rm a}}\,|H|}{\gamma_1\,\operatorname{arcsinh}(\xi/r_\circ)} \quad \text{when } \Delta \gg \xi. \tag{3.12}$$

  The defects repel and move at a constant speed. In fact, if we compare (3.12) with (2.20), we notice that in the large-distance limit the defects behave independently, each moving at the velocity computed in the 1-defect case, since $\operatorname{arcsinh}(\xi/r_\circ) \simeq \log(\xi/r_\circ)$ when $\xi \gg r_\circ$.
- The critical distance at which the defect interaction changes sign does not depend on the core radius $r_\circ$; all plots in Figure 6 cross the x-axis at $\Delta = x_{\rm cr}\,\xi$.
- In the small-distance limit, the stationary approximation we have used is not well justified. In this limit, the velocity diverges, and all terms depending

on $\dot{v}$ must be taken into account in the equation of motion. In any case, we remark that the Bessel function K_1 induces a divergence in the derivative of the free-energy that scales as the inverse of the defect distance. Figure 6 suggests that the fully nonlinear regime is limited to distances much smaller than the magnetic coherence length.

3.1. Tilted external fields. Let us now imagine that the external field is rotated by an angle α with respect to the direction of Figure 4. In this case, the nematic director is forced to relax to the external field direction everywhere but on the topologically irreducible π-wall. The motion will be much more complex—the defects will rotate, and the π-wall will not necessarily be straight at all times. However, it is possible to ascertain whether the defect interaction will lead to attraction or repulsion. To this end, we pin the defects at a distance Δ, and we look for the stationary director field in the presence of a tilted external field. Then, we compute the free energy of the stationary solution, and we check the sign of its derivative with respect to the distance. We stress that it is not possible to use this derivative in a dissipation principle to obtain a defect velocity. However, its sign will determine whether or not the defects, whatever their complex motion, will approach each other.

The stationary director field can be simply derived by running through the above steps again. In the presence of two defects placed at $x = \pm d$, and an external field tilted at an angle α with respect to the x-axis, the stationary configuration is given by

$$\Phi(x,y) = \varepsilon(y)\left[\alpha + \left(\frac{\pi}{2} - \alpha\right) \mathrm{e}^{-\frac{|y|}{\xi}} - \int_0^{+\infty} \cos qx \, \sin qd \, \mathrm{e}^{-k_0(q)\,|y|} \, \frac{dq}{q}\right], \tag{3.13}$$

where $k_0(q)$ coincides with the zero-velocity limit of $k(q)$ above: $k_0(q) := \sqrt{q^2 + 1/\xi^2}$.

If we compute the free energy associated with (3.13), and differentiate it with respect to Δ, we obtain the generalization of (3.10):

$$\left.\frac{\partial \mathcal{F}}{\partial \Delta}\right|_{v=0} = \frac{K\pi}{\xi}\left[f\left(\frac{\Delta}{\xi}\right) - \big(\pi - 2\alpha\big)\right], \tag{3.14}$$

with the same f defined in (3.9). The left panel of Figure 7 shows the plot of f. It enjoys the following properties:

- $f'(x) = \frac{1}{2}\big(K_0(x) - K_2(x)\big) < 0$ for all $x > 0$. Thus, the critical distance at which the defect interaction becomes repulsive increases when the external field is tilted. Furthermore, the free energy is a concave function of the distance between the defects.
- $\lim_{x\to 0} f(x) = +\infty$ and $\lim_{x\to\infty} f(x) = \frac{\pi}{2}$. More precisely,

$$\begin{aligned} f(x) &= \frac{1}{x} - \frac{x}{2}\log x + O(x) \text{ as } x \to 0^+ \quad \text{and} \\ f(x) &= \frac{\pi}{2} + o\left(\mathrm{e}^{-x}\right) \text{ as } x \to +\infty. \end{aligned} \tag{3.15}$$

Thus, the equation

$$\frac{\partial \mathcal{F}}{\partial \Delta} = 0 \quad \Longleftrightarrow \quad f(x) = \pi - 2\alpha, \tag{3.16}$$

which determines the equilibrium distances, possesses exactly one solution if and only if $\alpha < \frac{\pi}{4}$. This limiting value for the tilting angle could be predicted

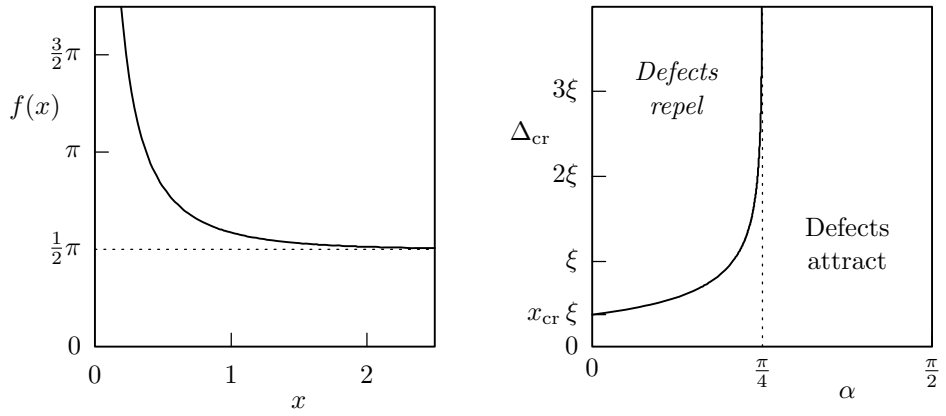


FIG. 7. *Left: Plot of the function f defined in the text. Right: Critical value of the defect distance as a function of the tilt angle α.*

easily. In fact, if the external field determines an angle greater than $\frac{\pi}{4}$ with the defect line, the director orientation between the defects costs *more* magnetic energy than the director orientation outside the defects. Thus, in this case, the external field strengthens the defect attraction at any distance.

In summary, we have the following:

- When $\alpha = 0$, the defects attract if $\Delta < \Delta_{\rm cr}\left(\frac{\pi}{2}\right) = 0.3774\xi$; they repel when $\Delta > \Delta_{\rm cr}\left(\frac{\pi}{2}\right)$.
- When $\alpha > 0$, the critical distance $\Delta_{\rm cr}(\beta)$ increases. It diverges when $\alpha \to \frac{\pi}{4}^{-}$.
- When $\alpha \geq \frac{\pi}{4}$, there is no critical distance: the defects always attract (the external field enhances the attraction).

4. Discussion. We have studied the motion of a single disclination, and a disclination dipole in an external field. Our results show that it is possible to drive the disclinations by suitably adjusting the external field direction and strength.

In the case of a single disclination, we have shown that the disclination velocity depends almost linearly on the field strength, since the coefficient depends on the logarithm of (2.20). Figure 3 shows that the linear scaling is abandoned when $\xi \simeq r_{\circ}$. Equation (2.21) shows that below this regime the disclination velocity is expected to scale as the square of the applied strength. However, this last prediction should be tested carefully, since it is surely influenced by our assumption that the core radius does not depend on the field strength.

A comparison between our analytical results and the experimental observations confirms our predictions and gives an estimate of the quantitative corrections that backflow effects require. If we again let v be the defect speed and H the external field intensity, the pioneering work of Geurst, Spruijt, and Gerritsma [30] reported a quadratic scaling $v \sim H^2$. However, their geometry is completely different from ours, since in this work the defect was confined in a very thin cell. Indeed, those authors observed that their measured velocity was consistent with the assumption that the presence of the external field did not influence the director field at all. The nematic cell was wider in the measurements reported by Cladis, van Saarloos, and Kortan [31], whose data fitted almost perfectly a linear relation similar to (2.20). This latter group realized that the defect speed is determined by a balance between the free energy gained by reducing the wall area and the dissipation stemming from the defect

core. Their estimate of the free energy gain coincides with ours (see (2.17)), but again they neglected the electric field effects when (over)estimating the dissipation. If we insert their numerical values into (2.20), we obtain a much closer value for the slope of their linear fit: their estimated defect speed was 0.43 times the actual measured speed, while, if we replace in (2.20) the same values for the same material parameters, the derived defect speed turns out to be 0.73 times the observed velocity.

We have also computed the velocity of two opposite disclinations in the presence of a field that promotes the director orientation between them. The velocity depends on the defect distance Δ, but, whatever the value of the ratio $\xi/r_\circ$, the defects approach if Δ is smaller than a critical distance $\Delta_{\rm cr}$; otherwise, they repel. We have finally generalized our calculations in order to deal with rotated external fields. Figure 7 (right) shows how the $\Delta_{\rm cr}$ depends on the angle α, which the external field determines, with the lie, which connects the defects.

The introduction of backflow may change the picture we have developed, in both the one- and two-disclination cases. In the one-disclination case the dissipation may sometimes be significantly reduced. The system can match the disclination motion with a flow that almost cancels the dissipation in the crucial core region. Whether this occurs seems likely to depend on the charge of the central disclination. We note that central physical and mathematical issues associated with backflow in nematic liquid crystal problems remain open even over a quarter of a century after its essential role was first realized [21].

Likewise in the two-disclination problem, our computed defect velocity is symmetrical in both disclinations. This is also an effect of the no-backflow simplification. A generalization of the present study should allow one to compute the different approaching or separating velocities. It could even happen that one defect moves towards the other, but that the other retreats faster still, allowing the defects eventually to separate. However, the existence of a critical distance that reverses the defect interaction cannot be erased by backflow effects. The critical distance stems from the balance between the elastic and magnetic free-energy gains. The magnetic gain does not depend on the defect distance, while the elastic gain vanishes when the defects move apart. Thus, at some intermediate distance one overwhelms the other.

Acknowledgments. P. B. acknowledges the hospitality of the University of Southampton, where this work was carried out.

REFERENCES

[1] F. C. Frank, *On the theory of liquid crystals*, Discuss. Faraday Soc., 25 (1958), pp. 19–28.
[2] N. D. Mermin, *The topological theory of defects in ordered media*, Rev. Mod. Phys., 51 (1979), pp. 591–648.
[3] M. Kléman, *Defects in liquid crystals*, Rep. Prog. Phys., 52 (1989), pp. 555–654.
[4] P. Biscari and G. Guidone Peroli, *A hierarchy of defects in biaxial nematics*, Commun. Math. Phys., 186 (1997), pp. 381–392.
[5] G. Guidone Peroli and E. G. Virga, *Nucleation of topological dipoles in nematic liquid crystals*, Commun. Math. Phys., 200 (1999), pp. 195–210.
[6] P. E. Cladis and M. Kléman, *Non-singular disclinations of strength $S = 1$ in nematics*, J. Phys. (Paris), 33 (1972), pp. 591–598.
[7] C. W. Oseen, *The theory of liquid crystals*, Trans. Faraday Soc., 29 (1933), pp. 883–900.
[8] H. Zocher, *The effect of a magnetic field on the nematic state*, Trans. Far. Soc., 29 (1933), pp. 945–957.
[9] P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals*, Clarendon Press, Oxford, UK, 1993.
[10] J. L. Ericksen, *Liquid crystals with variable degree of orientation*, Arch. Ration. Mech. Anal., 113 (1989), pp. 97–120.

[11] F.M. Leslie, *Continuum theory for nematic liquid crystals*, Contin. Mech. Thermodyn., 4 (1992), pp. 167–175.
[12] G. Ryskin and M. Kremenetsky, *Drag force on a line defect moving through an otherwise undisturbed field, disclination line in a nematic liquid crystal*, Phys. Rev. Lett., 67 (1991), pp. 1574–1577.
[13] E. I. Kats, V. V. Lebedev, and S. V. Malinin, *Disclination motion in liquid crystalline films*, J. Exp. Theor. Phys., 95 (2002), pp. 714–727.
[14] A. Pargellis, N. Turok, and B. Yurke, *Monopole-antimonopole annihilation in a nematic liquid crystal*, Phys. Rev. Lett., 67 (1991), pp. 1570–1573.
[15] L. M. Pismen and B. Y. Rubinstein, *Motion of interacting point defects in nematics*, Phys. Rev. Lett., 69 (1992), pp. 96–99.
[16] C. Denniston, *Disclination dynamics in nematic liquid crystals*, Phys. Rev. B, 54 (1996), pp. 6272–6275.
[17] G. Guidone Peroli and E. G. Virga, *Annihilation of point defects in nematic liquid crystals*, Phys. Rev. E, 54 (1996), pp. 5235–5241.
[18] G. Guidone Peroli and E. G. Virga, *Dynamics of point defects in nematic liquid crystals*, Phys. D, 111 (1998), pp. 356–372.
[19] P. Biscari, G. Guidone Peroli, and E. G. Virga, *A statistical study for evolving arrays of nematic point defects*, Liquid Crystals, 26 (1999), pp. 1825–1832.
[20] A. N. Pargellis, P. Finn, J. W. Goodby, P. Panizza, B. Yurke, and P. E. Cladis, *Defect dynamics and coarsening dynamics in smectic-C films*, Phys. Rev. A, 46 (1992), pp. 7765–7776.
[21] M. G. Clark and F. M. Leslie, *A calculation of orientational relaxation in nematic liquid crystals*, Proc. Roy. Soc. London A, 361 (1978), pp. 463–485.
[22] P. D. Olmsted and P. Goldbart, *Theory of the nonequilibrium phase transition for nematic liquid crystals under shear flow*, Phys. Rev. A, 41 (1990), pp. 4578–4581.
[23] P. D. Olmsted and P. Goldbart, *Isotropic-nematic transition in shear flow: State selection, coexistence, phase transitions, and critical behavior*, Phys. Rev. A, 46 (1992), pp. 4578–4993.
[24] G. Richardson, *Line disclination dynamics in uniaxial nematic liquid crystals*, Quart. J. Mech. Appl. Math., 53 (2000), pp. 49–71.
[25] E. J. Acosta, M. J. Towler, and H. G. Walton, *The role of surface tilt in the operation of pi-cell liquid crystal devices*, Liq. Cryst., 27 (2000), pp. 977–984.
[26] C. Denniston, E. Orlandini, and J. M. Yeomans, *Simulations of liquid crystal hydrodynamics in the isotropic and nematic phases*, Europhys. Lett., 52 (2000), pp. 481–487.
[27] C. Denniston, E. Orlandini, and J. M. Yeomans, *Lattice Boltzmann simulations of liquid crystal hydrodynamics*, Phys. Rev. E, 63 (2001), paper 056702.
[28] D. Svenšek and S. Žumer, *Hydrodynamics of pair-annihilating disclination lines in nematic liquid crystals*, Phys. Rev. E, 66 (2002), paper 021712.
[29] G. Tóth, C. Denniston, and J. M. Yeomans, *Hydrodynamics of topological defects in nematic liquid crystals*, Phys. Rev. Lett., 88 (2002), paper 105504.
[30] J. A. Geurst, A. M. J. Spruijt, and C. J. Gerritsma, *Dynamics of $s = \pm\frac{1}{2}$ disclinations in twisted nematics*, J. Physique, 36 (1975), pp. 653–664.
[31] P. E. Cladis, W. van Saarloos, and A. R. Kortan, *Dynamics of line defects in nematic liquid crystals*, Phys. Rev. Lett., 58 (1987), pp. 222–225.
[32] H. Brezis, J. M. Coron, and E. Lieb, *Harmonic maps with defects*, Comm. Math. Phys., 107 (1986), pp. 649–705.
[33] N. Schopohl and T. J. Sluckin, *Defect core structure in nematic liquid crystals*, Phys. Rev. Lett., 59 (1987), pp. 2582–2584.
[34] P. Biscari and T. J. Sluckin, *Expulsion of disclinations in nematic liquid crystals*, European J. Appl. Math., 14 (2003), pp. 39–59.
[35] P. Cermelli and E. Fried *The evolution equation for a disclination in a nematic liquid crystal*, Proc. Roy. Soc. London A, 458 (2002), pp. 1–20.

SIAM J. APPL. MATH.
Vol. 65, No. 6, pp. 2158–2177

SLOW PASSAGE THROUGH RESONANCE FOR A WEAKLY NONLINEAR DISPERSIVE WAVE*

SERGEI GLEBOV†, OLEG KISELEV‡, AND VLADIMIR LAZAREV†

Abstract. A solution of the nonlinear Klein–Gordon equation perturbed by a small external force is investigated. The frequency of the perturbation varies slowly and passes through a resonance. The resonance generates solitary packets of waves. The full asymptotic description of this process is presented.

Key words. nonlinear optics, resonance, solitons, nonlinear waves

AMS subject classifications. 35Q60, 37K40, 78M35

DOI. 10.1137/040618084

Introduction. This work is devoted to the problem of generation of a nearly monochromatic weakly nonlinear dispersive wave with small amplitude in a strong nonlinear media. It is well known that packets of nearly monochromatic waves propagate without changing their shape when the envelope function of the packet is a soliton of the nonlinear Schrödinger equation (NLSE). The solitary packets of waves would be more suitable for communication in optical fibers over a large distance if one could control the parameters of the envelope function for such packets. The wave packets have a soliton-like shape form for sufficiently large range of initial data. But the parameters of such self-generated solitons are difficult to predict in practice. This is explained by an instability of the parameters for solitons with respect to the initial data.

Here we propose a new approach for the generation of solitary packets of waves with given parameters. In our approach the wave packets appear due to a slow passage of the external driving force through the resonance. After the resonance, the envelope function of the wave packet is determined by the NLSE. In the most important cases the envelope function is a sequence of solitary waves which are called solitons. The wave packets, with the solitons as the envelope function, are propagated without dissipation. The parameters of the solitons are obviously defined by the value of the driving force on a resonance curve. We demonstrate this phenomenon for the perturbed nonlinear Klein–Gordon equation.

Here we give the mathematical basis for the proposed approach. This basis allows us to derive explicit formulas that define parameters for the solitary packets of waves with respect to the external driving force. Generation of the solitary packets of waves by the small driving force is described in detail. The formulas for the asymptotic solution before, after, and in the neighborhood of the resonant curve are obtained.

The proposed approach is based on a local resonance phenomenon. The local resonance in linear ordinary differential equations was investigated in papers [1, 2]. Later this phenomenon was studied in partial differential equations in the linear case

*Received by the editors November 2, 2004; accepted for publication (in revised form) March 10, 2005; published electronically August 22, 2005. This work was supported by grants RFBR 03-01-00716, Leading Scientific Schools 1446.2003.1, and INTAS 03-51-4286.
http://www.siam.org/journals/siap/65-6/61808.html
†Institute of Mathematics, Ufa State Petroleum Technical University, Ufa 450067, Russia (glebskie@rusoil.net, lazva@mail.ru).
‡Institute of Mathematics, USC RAS, Ufa 450077, Russia (ok@ufanet.ru).

[3] and in the weak nonlinear case [4, 5]. In these papers it was shown that the amplitude of the wave increases linearly when the wave passes through the local resonance. The increase of the amplitude is proportional to the width of the local resonance layer.

After the resonance a special proportion between the order of the solution and scales of independent variables appears. This magic proportion gives the NLSE for the envelope function. The deriving of the NLSE in such a case is a well-known result [6, 7, 8]. This result is justified in [9].

An important kind of solution of the NLSE is the soliton. The phenomenon of the generation of solitary waves for some nonlinear equations due to modulation instability is a well-known result [10]. For example, the detailed analytical description of the disappearance (generation) of the soliton due to modulation instability in the case of the Kadomtsev–Petviashvili equation was done in [11]. Some results about an appearance of solitons in the nonlinear Schrödinger equation due to instability were presented in [12]. The structural instability of the solitons for the Davey–Stewartson equation was shown in [13]. Such perturbations do not allow us to obtain solitons with the given parameters.

The generation of solitary waves by a small external resonant force was found by numerical simulation [14]. This simulation shows the possibility of generation of solitons by the external driving force. But it does not allow us to relate the parameters of the solitons and the perturbation. Therefore the problem about the generation of the soliton with the given parameters was still open.

The goal of this paper is to show that the process of generation of the solitary waves due to the local resonance is universal. This process allows us to control the parameters of the generated waves. This phenomenon previously was asymptotically investigated in the case of the nonlinear Schrödinger equation in [15]. In our work we consider the similar phenomenon in the nonlinear Klein–Gordon equation. Our approach demonstrates that solitary waves with the given parameters can be obtained for nonlinear systems.

This paper has the following structure. Section 1 contains the main result and an example of numerical simulations. Section 2 contains the asymptotic construction in the preresonant domain. In section 3 we construct the asymptotic solution in the neighborhood of the resonant curve. Section 4 is devoted to the construction of the postresonant asymptotics. All asymptotic approximations are matched. In the Summary (section 5) we outline the results and open problems.

1. Statement of the problem and result.

1.1. Statement of the problem. Let us consider the Klein–Gordon equation with a cubic nonlinearity

$$\partial_t^2 U - \partial_x^2 U + U + \gamma U^3 = \varepsilon^2 f(\varepsilon x) \exp\left\{ i \frac{S(\varepsilon^2 t, \varepsilon^2 x)}{\varepsilon^2} \right\} + \text{c.c.}, \quad 0 < \varepsilon \ll 1. \tag{1}$$

Here $\gamma =$ constant and c.c. stands for complex conjugate; $f(y)$ is smooth and rapidly vanishes as $y \to \pm\infty$. The phase function $S(y, z)$ of the driving force and all derivatives with respect to y, z are bounded. Here and below we use the following notation:

$$x_j = \varepsilon^j x, \quad t_j = \varepsilon^j t, \quad j = 1, 2;$$

$$l(t_2, x_2) \equiv (\partial_{t_2} S)^2 - (\partial_{x_2} S)^2 - 1.$$

We will construct a special asymptotic solution of (1) in a strip of finite width with respect to x_2, t_2. This domain covers the resonant curve $l = 0$. This asymptotic solution corresponds the forced oscillations as $l < 0$,

$$U \sim -\varepsilon^2 \frac{f}{l} \exp\left(\frac{iS(t_2, x_2)}{\varepsilon^2}\right) + c.c. \tag{2}$$

1.2. Main result. Let us formulate the main result of the work. If the solution of (1) has the form (2) when $l < 0$, then in the domain $l > 0$ this asymptotic solution is

$$U(x, t, \varepsilon) \sim \varepsilon \Psi(x_1, t_1, t_2) \exp\left\{\frac{i\varphi(x_2, t_2)}{\varepsilon^2}\right\} + c.c. \tag{3}$$

The phase function φ satisfies the eikonal equation

$$(\partial_{t_2}\varphi)^2 - (\partial_{x_2}\varphi)^2 - 1 = 0$$

with conditions

$$\varphi|_{l=0} = S|_{l=0}, \quad \partial_{t_2}\varphi|_{l=0} = \partial_{t_2} S|_{l=0}.$$

The envelope function of the leading-order term is a solution of the NLSE

$$2i\partial_{t_2}\varphi\partial_{t_2}\Psi + \partial_\xi^2\Psi + i[\partial_{t_2}^2\varphi - \partial_{x_2}^2\varphi]\Psi + \gamma|\Psi|^2\Psi = 0,$$

where the ξ is defined by

$$\frac{dx_1}{d\xi} = \partial_{t_2}\varphi, \quad \frac{dt_1}{d\xi} = \partial_{x_2}\varphi.$$

The initial condition for Ψ is

$$\Psi|_{l=0} = \int_{-\infty}^{\infty} d\sigma f(x_1) \exp\left(i\int_0^\sigma d\mu\lambda(x_1, t_1, \varepsilon)\right). \tag{4}$$

The integration in this integral is done in the characteristic direction related with (24), (25).

1.3. Higher-order terms and matching. The structure of the constructed asymptotic solution when $l < 0$ and $l > 0$ is sufficiently obvious. We concentrate on the description of the changing of the solution from the preresonant to postresonant form. This transition takes place in the thin layer near the curve $l = 0$. In this transition layer the amplitude of the solution increases due to the resonant pumping. The value of the amplitude is defined by the width of the resonant layer. We found the width of the layer by construction and analysis of the higher-order terms of the asymptotic solution in all domains. This analysis looks very complicated, but it is necessary to match the asymptotics of the solution in different domains and obtain formula (4). This formula defines the leading-order term of the solution after the slow passage through the resonance.

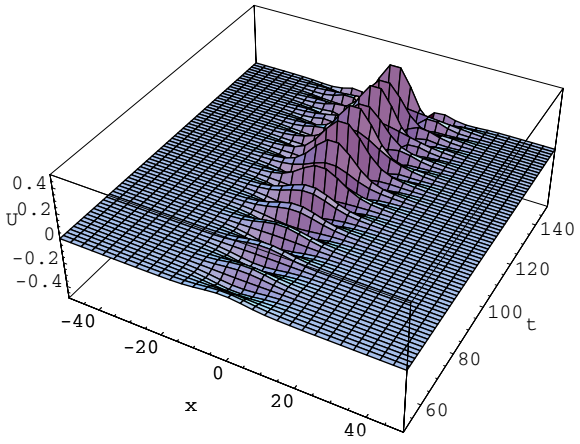


FIG. 1. *This picture shows the generation of the solitary packet of waves for* (1) *with special right-hand side* (5) *and at* $\varepsilon = 0.1$. *Initial conditions are* $U|_{t=0} = -\varepsilon^2 f \exp(iS/\varepsilon^2)|_{t=0}$, $\partial_t U|_{t=0} = -\varepsilon^2 \partial_t (f \exp(iS/\varepsilon^2))|_{t=0}$. *The resonant curve is* $t = 100$.

1.4. Numeric simulations. To illustrate the obtained result we consider (1) with $\gamma = 2$ and the simplest driving force, where

$$S = \frac{t_2^2}{2}, \quad f = \frac{2\sqrt{2}}{\sqrt{\pi}\cosh(2x_1)}. \tag{5}$$

In this case the curve of the local resonance is the line $t_2 = 1$. The preresonant solution has the form

$$U \sim \frac{-\varepsilon^2}{(t_2 - 1)} \frac{2\sqrt{2}}{\sqrt{\pi}\cosh(2x_1)} \cos\left(\frac{it_2^2}{\varepsilon^2}\right), \quad 0 < t_2 < 1.$$

In the domain $t_2 > 1$ the solution has the form

$$U \sim \varepsilon \exp\{i\varphi\left\{\frac{(x_2, t_2)}{\varepsilon^2}\right\} \Psi(x_1, t_1, t_2) + c.c.$$

Here $\varphi = t_2 - 1/2$. The function $\Psi(x_1, t_1, t_2)$ is the solution of the Cauchy problem for the NLSE:

$$2i\partial_{t_2}\Psi + \partial^2_{x_1 x_1}\Psi + 2|\Psi|^2\Psi = 0,$$

$$\Psi|_{t_2=1} = \frac{\sqrt{2}(1+i)}{\cosh(2x_1)}.$$

The solution of this Cauchy problem is the pure soliton. This soliton is the envelope function of the fast oscillating carrier in the postresonant solution of (1). The carrier with the soliton as the envelope function is the single solitary packet of waves. This packet is propagated in the nonlinear media without dispersion.

The numerical simulation at $\varepsilon = 0.1$ is given in Figure 1. The profile of the generation process for the solitary packet of waves can be seen in Figure 2.

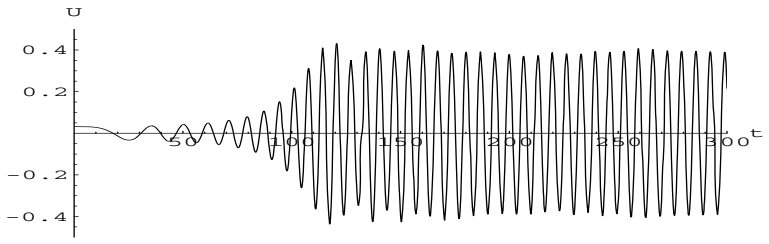


FIG. 2. *This picture shows a profile* $(U(x,t)|_{x=0})$ *of the generation process for the solitary packet of waves.*

1.5. Qualitative analysis. Here we explain the behavior of the solution for (1). All domains where we construct the solution are separated on three pairwise joint domains. The preresonant domain corresponds to the forced oscillations with amplitude of order ε^2. This oscillations down when the driving force becomes resonant.

The resonant layer is a thin domain near the resonant curve $l(x_2,t_2)=0$. In this layer the amplitude of the oscillations increases up to the order ε:

$$U(x,t,\varepsilon)\sim \varepsilon W_{1,1}(x_1,t_1,x_2,t_2)\exp\left\{\frac{iS}{\varepsilon^2}\right\}+c.c..$$

The coefficient $W_{1,1}(x_1,t_1,x_2,t_2)$ is defined by nonautonomous first-order partial differential equation

$$2i\partial_{t_2}S\partial_{t_1}W_{1,1}-2i\partial_{x_2}S\partial_{x_1}W_{1,1}-\lambda W_{1,1}=f,$$

with a given asymptotic behavior:

$$W_{1,1}\sim\frac{-f}{\lambda},\quad \lambda\to-\infty.$$

Here $\lambda=l/\varepsilon$.

The asymptotic behavior of $W_{1,1}$ as $\lambda\to\infty$ allows us to relate formulas (2) and (3).

The equation for $W_{1,1}$ may be written in the form of a first-order ordinary differential equation along the characteristic direction

$$\frac{d}{d\sigma}W_{1,1}+\lambda W_{1,1}=f.$$

Such an ordinary equation appears under studying of slow passage through resonance for a one-dimensional oscillator with slowly varying frequency [1]. The solution of equations of such type is defined by Fresnel integrals.

After the passage through this thin layer, the driving force becomes nonresonant. In this postresonant domain the amplitude of the solution has the order ε. In Figure 3 one can see the schematic position of the domains mentioned above.

Remark on WKB asymptotics. In this work we describe the special asymptotic solution of (1). This solution is defined by the driving force in the domain $l<0$. One can add any solution of WKB type [16] of the order ε^2 to this constructed solution, leading to an asymptotic solution for (1) in the form

$$\tilde{U}=U(t,x,\varepsilon)+\sum_{n\geq 2}^{N}\varepsilon^n\mathbf{U}_n(t,x,\varepsilon).$$

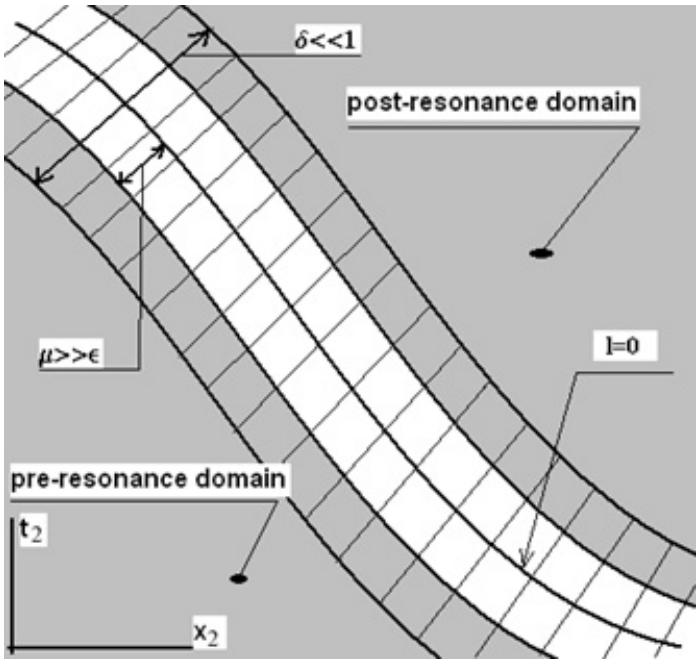


FIG. 3. *Schematic position of domains.*

The coefficients $\mathbf{U}_n(t,x,\varepsilon)$ of the asymptotics are calculated by standard methods of WKB-theory. This additional term leads to ponderous formulas and does not change the leading-order term of the postresonant asymptotics.

2. Preresonance expansion. In this section the formal asymptotic solution is constructed in the domain before the resonance. This domain is defined by the condition $l < 0$. The asymptotic expansion has a WKB-type form. The leading-order term of the asymptotics has the order of the driving force and oscillates with the frequency of the perturbation. The constructed asymptotics is valid when $-l \gg \varepsilon$. The result of this section is formulated below.

THEOREM 1. *In the domain $-l \gg \varepsilon$ the formal asymptotic solution of* (1) *modulo $O(\varepsilon^{N+1})$ has the form*

$$U = \sum_{n\geq 2}^{N} \varepsilon^n U_n(t,x,\varepsilon), \tag{6}$$

where

$$U_n = \sum_{k\in\Omega_n} U_{n,k}(t_2,x_2,\varepsilon x)\exp\left\{ik\frac{S(t_2,x_2)}{\varepsilon^2}\right\}.$$

The set Ω_n for the higher-order term is described by the formula

$$\Omega_n = \begin{cases} \{\pm 1\}, & n \leq 5, \\ \{\pm 1, \pm 3, \ldots, \pm(2l+3)\}, & l = \big[(n-6)/4\big], \quad n \geq 6. \end{cases}$$

The functions $U_{n,k}$ and $U_{n,-k}$ are complex conjugated.

The coefficients of the asymptotics $U_{n,k}$ are defined out of algebraic equations (7), (8), (9), *and* (11).

The proof of the theorem contains two steps. In the first step (subsection 2.1) we construct the coefficients of (6), and in the second step (subsection 2.2) we determine the domain of validity for (6).

2.1. Construction of preresonant asymptotics. Let us substitute (6) into (1) and collect the terms of the same order of ε. As a result we obtain a recurrent sequence of algebraic equations,

$$U_{2,1} = -\frac{f}{l}, \tag{7}$$

$$U_{3,1} = 2i\frac{\partial_{x_1} f \partial_{x_2} S}{l^2}, \tag{8}$$

$$U_{4,1} = \frac{2if[\partial_{t_2} S \partial_{t_2} l - \partial_{x_2} S \partial_{x_2} l] - 4(\partial_{x_2} S)^2 \partial_{x_1}^2 f}{l^3} - \frac{2i\partial_{t_2} f \partial_{t_2} S + \partial_{x_1}^2 f + i\partial_{t_2}^2 Sf}{l^2}, \tag{9}$$

where

$$l = (\partial_{t_2} S)^2 - (\partial_{x_2} S)^2 - 1.$$

The curve where the phase function S satisfies the eikonal equation is called the resonant curve,

$$l[S] = (\partial_{t_2} S)^2 - (\partial_{x_2} S)^2 - 1 = 0. \tag{10}$$

The amplitude $U_{n,1}$ has a singularity on this curve.

The formula for the nth-order term has the form

$$U_{n,k} = \frac{1}{l}\Bigg[\partial_{t_2 t_2}^2 U_{n-4,k} + 2ik\partial_{t_2} S \partial_{t_2} U_{n-2,k} + ikS_{t_2 t_2} U_{n-2,k} - 2ik\partial_{x_2} S \partial_{x_2} U_{n-2,k} - ik\partial_{x_2}^2 S U_{n-2,k} - \partial_{x_1 x_1}^2 U_{n-2,k} - 2\partial_{x_1 x_2}^2 U_{n-3,k} - \partial_{x_2 x_2}^2 U_{n-4,k} - 2ik\partial_{x_2} S \partial_{x_1} U_{n-1,k} + \gamma \sum_{\substack{n_1+n_2+n_3=n,\\ k_1+k_2+k_3=k \\ k\in\Omega_n}} U_{n_1,k_1} U_{n_2,k_2} U_{n_3,k_3}\Bigg]. \tag{11}$$

The first step of the proof of Theorem 1 is completed.

2.2. The asymptotic behavior of the coefficients. To realize the second step of the proof for Theorem 1 we need to determine the behavior of the coefficients of (6) as $l \to -0$.

LEMMA 1. *The coefficient $U_{n,k}$ has the following behavior:*

$$U_{n,k} = O(l^{-(n-k)}), \quad k > 0, l \to -0. \tag{12}$$

Proof. Let us prove this formula at $k = 1$. The validity of formula (12) for $n = 2, 3, 4$ obtains from (7), (8), (9). Now suppose that this formula is valid for the term $U_{n-1,1}$. The increase of the order of the singularity when $l \to 0$ occurs due to differentiation with respect to x_2, t_2 and the nonlinear term in formula (11). Differentiation of the terms in formula (11) leads to formula (12).

Let us consider $U_{n,k}$ for $k > 1$. The validity of formula (12) for small values of n and k obtains by direct calculations. Consider the nth-order term. It contains the terms with different values of k. The higher-order terms with $k = 3$ have the greatest order of singularity,

$$U_{n,3} = O(l^{-(n-3)}), \quad l \to -0. \tag{13}$$

It takes place because the right-hand side of (11) contains the term $U_{n-4,\pm1}U_{2,\pm1} \cdot U_{2,\pm1}$. The calculation of the order of singularity for this term leads to formula (13). The terms of the type of $U_{n_3,\pm3}U_{n_1,\mp k_1}U_{n_2,\pm k_1}$, $n_1 + n_2 + n_3 = n$, lead to weak singularities; for example, for $k_1 = 3$ we obtain that the order of singularity is equal to $n - 9$.

Consider the nonlinear term $U_{n_1,k_1}U_{n_2,k_2}U_{n_3,k_3}$ from right-hand side of (11) when the number of the higher-order term is equal to n. Calculate the order of the singularity for this term using the $(n-1)$th step of the induction. The indexes of the amplitudes are related by formulas

$$n_1 + n_2 + n_3 = n, \quad k_1 + k_2 + k_3 = k.$$

Using (12) for $n_1, n_2, n_3 < n$, we obtain, that the order of the singularity for this term is equal to $(n - k)$.

The right-hand side of (11) contains derivatives of previous terms with respect to x_2, t_2. It leads to the increase of the order of the singularity, but the leading order nevertheless we obtain from nonlinear terms. The lemma is proved. □

Domain of validity. The domain of validity as $l \to -0$ for the formal asymptotic solution in form (6) is defined by

$$\varepsilon \max_{x_2,t_2} |U_{n+1}| = o\left(\max_{x_2,t_2} |U_n|\right), \quad l < 0, \quad \varepsilon \to 0.$$

Using Lemma 1, we obtain

$$-l \gg \varepsilon.$$

Theorem 1 is proved. □

Remark. The constructed asymptotic solution (6) is also valid in the domain $l \gg \varepsilon$. However, we use the asymptotics for the preresonant domain (when $l < 0$) only.

Lemma 1 gives the asymptotic representation for (6), as $l \to -0$ has the form

$$U = \sum_{n=2}^{N} \varepsilon^n \sum_{k \in \Omega_n} \exp\left\{\frac{ikS}{\varepsilon^2}\right\} \sum_{j=-(n-k)}^{\infty} U_{n,k}^j \, l^j, \quad l \to -0. \tag{14}$$

3. Internal asymptotics. This part of the paper contains the asymptotic construction of the solution for (1) in the neighborhood of the curve $l = 0$. The domain of validity of this asymptotics intersects with the domain of validity of expansion (6). These expansions are matched.

THEOREM 2. *In the domain $|l| \ll 1$ the formal asymptotic solution for* (1) *modulo* $O(\varepsilon^{N+1})$ *has the form*

$$U = \sum_{n\geq 1}^{N} \varepsilon^n W_n(t_1, x_1, t_2, x_2, \varepsilon), \tag{15}$$

where

$$W_n = \sum_{k\in\Omega_n} W_{n,k}(x_2, t_2, x_1, t_1) \exp\left\{ ik\frac{S(t_2, x_2)}{\varepsilon^2}\right\}. \tag{16}$$

The function $W_{n,1}$ is a solution of the problem for (21), (23) *with zero condition as $\lambda \to -\infty$. When $k \neq 1$, $W_{n,k}$ is the solution of algebraic equation* (23). *The functions $W_{n,k}$ and $W_{n,-k}$ are complex conjugated.*

There is an essential difference between asymptotics (15) and external preresonance asymptotics (6). In the first place the leading-order term in (15) has order ε, while the leading order term in (6) has order ε^2. In the second place the coefficients of asymptotics (15) depend on fast variables $x_1 = x_2/\varepsilon$ and $t_1 = t_2/\varepsilon$.

The proof of theorem 2 consists of three steps. In the first step (subsection3.1) we derive equations for coefficients of the asymptotics. In the second step (subsection3.2) we solve the problems for the asymptotic coefficients. In the third step (subsection3.3) we determine the domain of the validity for expansion (15).

3.1. The equations for coefficients. Let us construct the internal asymptotic expansion in the domain $|l| \ll 1$. Define

$$\lambda(x_1, t_1, \varepsilon) = \frac{1}{\varepsilon} l(\varepsilon x_1, \varepsilon t_1). \tag{17}$$

In the domain $1 \ll -\lambda \ll \varepsilon^{-1}$ both asymptotics (6) and (15) are valid. This fact allows us to obtain the asymptotic representation for coefficients of the internal asymptotics. Substituting $l = \varepsilon\lambda$ into formula (14) and expanding the obtained expression in powers of ε, we find that

$$W_{n,k} = \sum_{j=(n-k+1)}^{\infty} \lambda^{-j} U^j_{n+1,k}(x_2, t_2, x_1), \quad k \in \Omega_n, \quad \lambda \to -\infty. \tag{18}$$

Let us obtain the differential equations for the coefficients of asymptotics (15). Substituting (15), (16) into (1) and collecting the terms with equal powers of small parameter and exponents, we find the equations for coefficients $W_{n,k}$. In particular, the terms of order ε^2 give us the equations for the leading-order terms of the asymptotics,

$$2i\partial_{t_2} S \partial_{t_1} W_{1,1} - 2i\partial_{x_2} S \partial_{x_1} W_{1,1} - \lambda W_{1,1} = f, \tag{19}$$

and the complex conjugated equation for $W_{1,-1}$.

The relation of order ε^3 in (1) gives us four equations. Two of them are complex conjugate differential equations for $W_{2,1}$ and $W_{2,-1}$:

$$\begin{aligned} &2i\partial_{t_2} S \partial_{t_1} W_{2,1} - 2i\partial_{x_2} S \partial_{x_1} W_{2,1} - \lambda W_{2,1} = \partial^2_{x_1} W_{1,1} - \partial^2_{t_1} W_{1,1} \\ &- i[\partial^2_{t_2} S - \partial^2_{x_2} S] W_{1,1} - 2i\partial_{t_2} S \partial_{t_2} W_{1,1} + 2i\partial_{x_2} S \partial_{x_2} W_{1,1} - 3\gamma |W_{1,1}|^2 W_{1,1}. \end{aligned} \tag{20}$$

Two another equations are algebraic. These last equations allow us to determine the function $W_{3,3}$

$$W_{3,3} = \frac{\gamma}{8}(W_{1,1})^3.$$

The higher-order terms are calculated in the same way. In particular, the term $W_{n,1}$ is determined by the differential equation

$$2i\partial_{t_2}S\partial_{t_1}W_{n,1} - 2i\partial_{x_2}S\partial_{x_1}W_{n,1} - \lambda W_{n,1} = F_{n,1}. \tag{21}$$

The right-hand side of (21) has the form

$$\begin{aligned} F_{n,1} = {} & -2i\partial_{t_2}S\partial_{t_2}W_{n-1,1} + 2i\partial_{x_2}S\partial_{x_2}W_{n-1,1} + (\partial_{t_2}S)^2W_{n-1,1} - (\partial_{x_2}S)^2W_{n-1,1} \\ & - \partial_{t_1}^2W_{n-1,1} + \partial_{x_1}^2W_{n-1,1} - \partial_{t_2}\partial_{t_1}W_{n-2,1} + \partial_{x_2}\partial_{x_1}W_{n-2,1} \\ & - \partial_{t_2}^2W_{n-3,1} + \partial_{x_2}^2W_{n-3,1} - \gamma \sum_{\substack{n_1+n_2+n_3=n+1,\\ k_1+k_2+k_3=1,\\ k_j\in\Omega_{n_j},\ j=1,2,3}} W_{n_1,k_1}W_{n_2,k_2}W_{n_3,k_3}. \end{aligned} \tag{22}$$

The term $W_{n,k}$, $k \neq 1$ is determined by algebraic equation

$$\begin{aligned} W_{n,k} = \frac{\gamma}{k^2-1}\Bigg(& -2i\partial_{t_2}S\partial_{t_2}W_{n-2,k} + 2i\partial_{x_2}S\partial_{x_2}W_{n-2,k} \\ & + (\partial_{t_2}S)^2W_{n-2,k} - (\partial_{x_2}S)^2W_{n-2,k} \\ & - \partial_{t_1}^2W_{n-2,k} + \partial_{x_1}^2W_{n-2,k} - \partial_{t_2}\partial_{t_1}W_{n-3,k} + \partial_{x_2}\partial_{x_1}W_{n-3,k} \\ & - \partial_{t_2}^2W_{n-4,k} + \partial_{x_2}^2W_{n-4,k} - \sum_{\substack{n_1+n_2+n_3=n+1,\\ k_1+k_2+k_3=k,\\ k_j\in\Omega_{n_j},\ j=1,2,3}} W_{n_1,k_1}W_{n_2,k_2}W_{n_3,k_3}\Bigg). \end{aligned} \tag{23}$$

Thus we complete step 1 of the proof for Theorem 2.

3.2. Solvability of equations for higher-order terms. In this subsection we present the explicit form for higher-order term $W_{n,1}$ and investigate the asymptotic behavior as $\lambda \to \pm\infty$.

3.2.1. Characteristic variables. The function $W_{n,1}$ satisfies (21). The solution is constructed by the method of characteristics. Define the characteristic variables σ, ξ. We choose a point (x_1^0, t_1^0) such that $\partial_{x_2}l|_{(x_1^0,t_1^0)} \neq 0$ is an origin, and denote by σ the variable along the characteristics for (21). We suppose $\sigma = 0$ on the curve $\lambda = 0$. The variable ξ measures the distance along the curve $\lambda = 0$ from the point (x_1^0, t_1^0). This point (x_1^0, t_1^0) corresponds to $\xi = 0$. The positive direction for parameter ξ coincides with the positive direction of x_2 in the neighborhood of (x_1^0, t_1^0).

The characteristic equations for (21) have a form

$$\frac{dt_1}{d\sigma} = 2\partial_{t_2}S(\varepsilon x_1, \varepsilon t_1), \quad \frac{dx_1}{d\sigma} = -2\partial_{x_2}S(\varepsilon x_1, \varepsilon t_1). \tag{24}$$

The initial conditions for the equations are

$$x_1|_{\sigma=0} = x_1^0, \quad t_1|_{\sigma=0} = t_1^0. \tag{25}$$

LEMMA 2. *The Cauchy problem for characteristics has a solution when* $|\sigma| < c_1\varepsilon^{-1}, c_1 = const. > 0$.

Proof. The Cauchy problem (24), (25) is equivalent to the system of integral equations

$$t_1 = t_1^0 + 2\int_0^\sigma \partial_{t_2} S(\varepsilon x_1, \varepsilon t_1)d\zeta, \qquad x_1 = x_1^0 - 2\int_0^\sigma \partial_{x_2} S(\varepsilon x_1, \varepsilon t_1)d\zeta. \tag{26}$$

Substituting $\tilde{t}_2 = (t_1 - t_1^0)\varepsilon,\ \tilde{x}_2 = (x_1 - x_1^0)\varepsilon$, we obtain

$$\tilde{t}_2 = 2\int_0^{\varepsilon\sigma} \partial_{t_2} S(\tilde{x}_2 - \varepsilon x_1^0, \tilde{t}_2 - \varepsilon t_1^0)d\zeta, \qquad \tilde{x}_2 = -2\int_0^{\varepsilon\sigma} \partial_{x_2} S(\tilde{x}_2 - \varepsilon x_1^0, \tilde{t}_2 - \varepsilon t_1^0)d\zeta.$$

The integrands are smooth and bounded functions on the plane x_2, t_2. There exists the constant $c_1 = \text{const.} > 0$ such that the integral operator is a contraction operator when $\varepsilon|\sigma| < c_1$. Lemma 2 is proved. □

It is convenient to use the following asymptotic formulas for the change of variables $(x_1, t_1) \to (\sigma, \xi)$.

LEMMA 3. *In the domain* $|\sigma| \ll \varepsilon^{-1}$ *the asymptotics as* $\varepsilon \to 0$ *of the solutions for Cauchy problem* (24), (25) *have the form*

$$x_1(\sigma, \xi, \varepsilon) - x_1^0(\xi) = -2\sigma\partial_{x_2} S + 2\sum_{n=1}^{N} \varepsilon^n \sigma^{n+1} g_n(\varepsilon x_1, \varepsilon t_1) + O(\varepsilon^{N+1}\sigma^{N+2}), \tag{27}$$

$$t_1(\sigma, \xi, \varepsilon) - t_1^0(\xi) = 2\sigma\partial_{t_2} S + 2\sum_{n=1}^{N} \varepsilon^n \sigma^{n+1} h_n(\varepsilon x_1, \varepsilon t_1) + O(\varepsilon^{N+1}\sigma^{N+2}), \tag{28}$$

where

$$g_n = -\frac{d^n}{d\sigma^n}(\partial_{x_2} S)\Big|_{\sigma=0}, \quad h_n = \frac{d^n}{d\sigma^n}(\partial_{t_2} S)\Big|_{\sigma=0}.$$

The lemma can be proved by integration by parts of (26). □

The next claim gives us the asymptotic formula which relates the variables σ and λ as $\sigma, \lambda \to \pm\infty$.

LEMMA 4. *Let be* $\sigma \ll \varepsilon^{-1}$*; then*

$$\lambda = \varphi(\xi)\sigma + O(\varepsilon\sigma^2), \qquad \varphi(\xi) = \frac{d\lambda}{d\sigma}\Big|_{\sigma=0} \qquad \text{as } \sigma \to \infty.$$

Proof. From formula (17) we obtain the representation

$$\lambda = \sum_{j=1}^{\infty} \lambda_j(x_1, t_1, \varepsilon)\sigma^j \varepsilon^{j-1},$$

where

$$\lambda_j(x_1, t_1, \varepsilon) = \frac{1}{j!}\frac{d^j}{d\sigma^j}\lambda(x_1, t_1, \varepsilon)|_{\sigma=0}.$$

It yields

$$\lambda = \frac{d\lambda}{d\sigma}\Big|_{\sigma=0}\sigma + O\left(\varepsilon\sigma^2 \frac{d^2\lambda}{d\sigma^2}\right).$$

Let

$$\left|\frac{d^2 l}{d\sigma^2}\right| \geq \text{const.} \qquad \xi \in R.$$

The function $d\lambda/d\sigma$ is not equal to zero:

$$\frac{d\lambda}{d\sigma} = \frac{1}{2}\left(-\partial_{x_2}\lambda\partial_{x_2}S + \partial_{t_2}\lambda\partial_{t_2}S\right) \neq 0.$$

Let us suppose $d\lambda/d\sigma > 0$. It yields

$$\lambda = \varphi(\xi)\sigma + O(\varepsilon\sigma^2), \qquad \varphi(\xi) = \left.\frac{d\lambda}{d\sigma}\right|_{\sigma=0}.$$

The lemma is proved. □

3.2.2. Solutions of the equations for higher-order terms. The higher-order terms $W_{n,\pm 1}$ are solutions of (21) with the given asymptotic behavior $\lambda \to -\infty$. Equation (21) can be written in characteristic variables as

$$i\frac{d}{d\sigma}W_{n,1} - \lambda W_{n,1} = F_{n,1}. \tag{29}$$

LEMMA 5. *The solution of* (21) *with the asymptotic behavior* (18) *as* $\lambda \to -\infty$ *has the form*

(30)
$$W_{n,1} = \exp\left(-i\int_0^\sigma d\zeta\lambda(x_1,t_1,\varepsilon)\right)\int_{-\infty}^{\sigma} d\zeta F_{n,1}(x_1,t_1,\varepsilon)\exp\left(-i\int_0^\zeta d\chi\lambda(x_1,t_1,\varepsilon)\right).$$

Proof. By direct substitution we see that expression (30) is the solution of (29). The asymptotics of this solution as $\lambda \to -\infty$ can be obtained by integration by parts and substitution:

$$\frac{d}{d\sigma} = 2\partial_{t_2}S\partial_{t_1} - 2\partial_{x_2}S\partial_{x_1}.$$

This yields

$$W_{n,1} = \sum_{j=0}^{\infty}\left(\frac{2\partial_{t_2}S\partial_{t_1} - 2\partial_{x_2}S\partial_{x_1}}{i\lambda}\right)^j\left[\frac{F_{n,1}}{i\lambda}\right], \quad \lambda \to -\infty. \tag{31}$$

From formula (22) we obtain that formulas (31) and (18) are equivalent. The lemma is proved. □

Thus we complete step 2 of the proof for Theorem 2.

3.3. Asymptotics as $\lambda \to \infty$ and domain of validity of the internal asymptotics. The domain of validity of the internal expansion is determined by the asymptotic behavior of higher-order terms. In this section we show that the nth-order term of the asymptotic solution increases as λ^{n-1} when $\lambda \to \infty$. This increase of higher-order terms allows us to determine the domain of validity for internal asymptotics (15) as $\lambda \to \infty$.

3.3.1. Asymptotics of higher-order terms. This section contains two propositions concerning asymptotic behavior as $\lambda \to \infty$ for higher-order terms in (15). The first lemma describes the asymptotic behavior of higher-order terms as $\lambda \to \infty$, and the second one contains a result about asymptotics of the phase function.

LEMMA 6. *The asymptotic behavior of $W_{n,1}$ when $1 \ll \lambda \ll \varepsilon^{-1}$ has the form*

$$W_{n,1} = \sum_{j=0}^{n-1}\sum_{k=0}^{j-1}\left(\lambda^j \ln^k|\lambda| W_{n,1}^{(j,k)}(\xi)\right)\exp\left(-i\int_0^\sigma d\zeta\lambda(x_1,t_1,\varepsilon)\right) + \sum_{j=0}^{\infty}\left(\frac{2\partial_{t_2}S\partial_{t_1} - 2\partial_{x_2}S\partial_{x_1}}{i\lambda}\right)^j\left[\frac{F_{n,1}}{i\lambda}\right]. \tag{32}$$

Proof. The asymptotic behavior of the coefficients $W_{n,1}$ is calculated recurrently. Let us calculate the asymptotic behavior of the leading-order term

$$\begin{aligned} W_{1,1} &= \exp\left(-i\int_0^\sigma d\zeta\lambda(x_1,t_1,\varepsilon)\right)\int_{-\infty}^{\zeta} d\zeta f(x_1)\exp\left(i\int_0^\sigma d\chi\lambda(x_1,t_1,\varepsilon)\right) \\ &= \exp\left(-i\int_0^\sigma d\zeta\lambda(x_1,t_1,\varepsilon)\right)\int_{-\infty}^{\infty} d\zeta f(x_1)\exp\left(i\int_0^\zeta d\chi\lambda(x_1,t_1,\varepsilon\right) \\ &\quad - \exp\left(-i\int_0^\sigma d\zeta\lambda(x_1,t_1,\varepsilon)\right)\int_{-\sigma}^{\infty} d\zeta f(x_1)\exp\left(i\int_0^\zeta d\chi\lambda(x_1,t_1,\varepsilon)\right). \end{aligned}$$

Further, by integration by parts of the last term, we obtain formula (32) at $n = 1$, where

$$W_{1,1}^{(0,0)}(\xi) = \int_{-\infty}^{\infty} d\sigma f(x_1)\exp\left(i\int_0^\sigma d\chi\lambda(x_1,t_1,\varepsilon)\right),$$
$$F_{1,1} = f(x_1).$$

To calculate the asymptotics of $W_{n,1}$ in formula (30) we use the asymptotics with respect to σ of the previous correction terms. In this case the integral (30) contains the increasing terms with respect to σ. We eliminate this growing part from the integral explicitly. The residual integral converges as $\sigma \to \infty$. It can be calculated in the same manner as it was calculated for $W_{1,1}$ and yields formulas (32) for any n,

$$W_{n,1}^{(k,0)}(\xi) = \frac{1}{k}W_{n-1,1}^{(k-1,0)}(\xi) + \frac{\gamma}{k}\sum W_{m_1,\chi_1}^{(\kappa_1,0)}(\xi)W_{m_2,\chi_2}^{(\kappa_2,0)}(\xi)W_{m_3,\chi_3}^{(\kappa_3,0)}(\xi),$$

where $m_1+m_2+m_3 = n+1,\ \kappa_1+\kappa_2+\kappa_3 = k-1,\ \chi_1+\chi_2+\chi_3 = 1$, and

$$W_{n,1}^{(j,k)}(\xi) = \frac{1}{j}W_{n-1,1}^{(j-1,k)}(\xi) + \frac{1}{k}W_{n-1,1}^{(j-1,k-1)}(\xi) + \Sigma_1 + \Sigma_2.$$

We define

$$\Sigma_1 = \frac{\gamma}{j}\sum W_{m_1,\chi_1}^{(\kappa_1,\nu_1)}(\xi)W_{m_2,\chi_2}^{(\kappa_2,\nu_2)}(\xi)W_{m_3,\chi_3}^{(\kappa_3,\nu_3)}(\xi),$$

where $m_1+m_2+m_3 = n-1,\ \kappa_1+\kappa_2+\kappa_3 = j-1,\ \chi_1+\chi_2+\chi_3 = 1, \nu_1+\nu_2+\nu_3 = k$, and

$$\Sigma_2 = \frac{\gamma}{k}\sum W_{m_1,\chi_1}^{(\kappa_1,\nu_1)}(\xi)W_{m_2,\chi_2}^{(\kappa_2,\nu_2)}(\xi)W_{m_3,\chi_3}^{(\kappa_3,\nu_3)}(\xi),$$

where $m_1+m_2+m_3 = n-1,\ \kappa_1+\kappa_2+\kappa_3 = j-1,\ \chi_1+\chi_2+\chi_3 = 1, \nu_1+\nu_2+\nu_3 = k-1$.

The lemma is proved. □

To complete the proof of Theorem 2 we need to obtain the domain of validity of asymptotics (15). The formal series (15) is asymptotic when

$$\varepsilon \max_{x_1,x_2,t_1,t_2} |W_{n+1}| = o\left(\max_{x_1,x_2,t_1,t_2} |W_n|\right), \quad \varepsilon \to 0.$$

Lemma 6 gives $\lambda \ll \varepsilon^{-1}$. After substitution $\lambda = \varepsilon l$, we obtain $l \ll 1$. Theorem 2 is proved. □

3.4. Asymptotics of the phase function as $\lambda \to \infty$. To obtain the asymptotics as $\lambda \to \infty$ we need to derive the asymptotics of the phase function in formula (32).

LEMMA 7. *As* $\lambda \to \infty$,

$$\int_0^\sigma d\xi\lambda = \frac{S}{\varepsilon^2} + \frac{1}{\varepsilon}(\partial_{x_2} S(x_1 - x_1^0) + \partial_{t_2} S(t_1 - t_1^0)) + O(\varepsilon\lambda^3). \tag{33}$$

Proof. Substitute the asymptotics of λ from Lemma 6. Calculate the asymptotics of the integral in formula (33)

$$\int_0^\sigma d\zeta\lambda(x_1, t_1, \varepsilon) = \int_0^\sigma \frac{d\zeta}{2}\left[(-\partial_{x_2} l\partial_{x_2} S + \partial_{t_2} l\partial_{t_2} S)\zeta + O(\varepsilon\zeta^2)\right]$$
$$= (-\partial_{x_2} l\partial_{x_2} S + \partial_{t_2} l\partial_{t_2} S)\frac{\sigma^2}{4} + O(\varepsilon\sigma^3).$$

The asymptotics of the phase function $S(x_2, t_2)$ in the neighborhood of the curve $l = 0$ is represented by a segment of the Taylor series. It yields

$$\frac{S}{\varepsilon^2} = \frac{1}{\varepsilon}(\partial_{x_2} S(x_1 - x_1^0) + \partial_{t_2} S(t_1 - t_1^0))$$
$$+ \frac{1}{2}(S_{x_2x_2}(x_1 - x_1^0)^2 + 2S_{x_2t_2}(x_1 - x_1^0)(t_1 - t_1^0) + S_{t_2t_2}(t_1 - t_1^0)^2)$$
$$+ O(\varepsilon(|t_1 - t_1^0| + |t_1 - t_1^0|)^3).$$

Substitute instead of $(x_1 - x_1^0)$ and $(t_1 - t_1^0)$ their asymptotic behavior with respect to ε from Lemma 3. This substitution and the result of Lemma 4 complete the proof of Lemma 7. □

The asymptotics as $\lambda \to -\infty$ contains fast oscillating terms with phase functions $kS, k \in Z$. The leading-order term of the asymptotics as $\lambda \to \infty$ contains the oscillations with an additional phase function. We obtain this result from Lemma 6. Denote this new phase function by $\varphi(x_2, t_2)/\varepsilon^2$. The asymptotics of this function is obtained in Lemma 7. The nonlinearity and additional phase function lead to more complicated structure of the phase set for higher-order terms of the asymptotics as $\lambda \to \infty$.

LEMMA 8. *The phase set K_n for the nth-order term of the asymptotics as $\lambda \to \infty$ is determined by formula*

$$K_1 = \pm\varphi, \quad K_2 = \pm\varphi, \pm S, \quad K_n = \cup_{j_1+j_2+j_3=n}\chi_{j_1} + \chi_{j_2} + \chi_{j_3}, \ \chi_{j_k} \in K_{j_k}.$$

Proof. The proof of this lemma follows from the asymptotic formula for the nth-order term. Representation (15), formula (32), and Lemma 6 allow us to construct

the asymptotics as $\lambda \to \infty$ of the internal expansion in an explicit form:

$$U = \sum_{n=1}^{N} \varepsilon^n \left(\sum_{j=0}^{n-1} \sum_{k=0}^{n-2} \lambda^j \ln^k |\lambda| W_{n,1}^{(j,k)}(\xi) \right)$$
$$\times \exp\left[-i\left(\frac{1}{\varepsilon}(\partial_{x_2} S(x_1 - x_1^0) + \partial_{t_2} S(t_1 - t_1^0)) + O(\varepsilon\lambda^3) \right) \right]$$
$$+ \sum_{n=1}^{N} \varepsilon^n \left(\sum_{j=0}^{\infty} \left(\frac{2\partial_{t_2} S \partial_{t_1} - 2\partial_{x_2} S \partial_{x_1}}{i\lambda} \right)^j \left[\frac{F_{n,1}}{i\lambda} \right] \right) \exp\left\{ i \frac{S(t_2, x_2)}{\varepsilon^2} \right\}$$
$$+ \sum_{n=2}^{N} \varepsilon^n \left(\sum_{k \in \Omega, k \neq \pm 1} W_{n,k} \exp\left\{ ik \frac{S(t_2, x_2)}{\varepsilon^2} \right\} \right) + \text{ c.c.} \tag{34}$$

This representation and formula (23) complete the proof of the lemma. □

4. Postresonant expansion. This section contains the construction of the asymptotics of the solution for (1) after the passage through the resonance. The constructed solution has the order ε and oscillates. The envelope function of these oscillations satisfies the nonlinear Schrödinger equation. This section consists of two parts. The first part contains the construction of the formal asymptotic solution. We obtain the equations for the higher-order terms of the asymptotics. The asymptotic behavior for the higher-order terms as $l \to +0$ follows from section 3.4. In the second part of this section we determine the domain of validity for this external asymptotics near the resonant curve $l(x_2, t_2) = 0$. The matching method gives us the initial conditions for higher-order terms of the asymptotics.

The main result of this section is formulated in the following theorem.

THEOREM 3. *In the domain $l \gg \varepsilon$ the formal asymptotic solution of* (1) *modulo $O(\varepsilon^{N+1})$ has the form*

$$U(x,t,\varepsilon) = \sum_{1}^{N} \varepsilon^n \sum_{k=0}^{n-2} \ln^k(\varepsilon) \left(\sum_{\pm\varphi} \exp\left\{ \pm \frac{i\varphi(x_2,t_2)}{\varepsilon^2} \right\} \Psi_{n,k,\pm\varphi}(x_1,t_1,t_2) \right.$$
$$\left. + \sum_{\chi \in K'_{n,k}} \exp\left\{ \frac{i\chi(x_2,t_2)}{\varepsilon^2} \right\} \Psi_{n,k,\chi}(x_1,t_1,t_2) \right). \tag{35}$$

Here the function $\varphi(x_2, t_2)$ satisfies the eikonal equation

$$(\partial_{t_2}\varphi)^2 - (\partial_{x_2}\varphi)^2 - 1 = 0 \tag{36}$$

and initial condition on the curve $l = 0$:

$$\varphi|_{l=0} = S|_{l=0}, \quad \partial_{t_2}\varphi_{l=0} = \partial_{t_2} S|_{l=0}.$$

The leading-order term of the asymptotics is a solution of the Cauchy problem for the nonlinear Schrödinger equation

$$2i\partial_{t_2}\varphi\partial_{t_2}\Psi_{1,0,\varphi} + \partial_\xi^2 \Psi_{1,0,\varphi} + i[\partial_{t_2}^2\varphi - \partial_{x_2}^2\varphi]\Psi_{1,0,\varphi} + \gamma|\Psi_{1,0,\varphi}|^2\Psi_{1,0,\varphi} = 0,$$

$$\Psi_{1,0,\varphi}|_{l=0} = \int_{-\infty}^{\infty} d\sigma f(x_1) \exp\left(i \int_0^{\sigma} d\chi \lambda(x_1, t_1, \varepsilon) \right),$$

where ξ is defined by

$$\frac{dx_1}{d\xi} = \partial_{t_2}\varphi, \quad \frac{dt_1}{d\xi} = \partial_{x_2}\varphi.$$

The coefficients $\Psi_{n,k,\pm\varphi}$ are determined from Cauchy problems for linearized Schrödinger equation (42). *The coefficients $\Psi_{n,k,\chi}$, $\chi \in K'_{n,k}$, are determined from algebraic equations* (43). *The set $K'_{n,k} = K_{n,k}\backslash\{\pm\varphi\}$.*

The proof of this theorem contains two steps. In the first step (subsection 4.1) we derive the recurrent system of the problems for the coefficients of the expansion (35). In the second step (subsection 4.1) we define the domain of validity for (35).

4.1. Structure of the second external asymptotics. Let us construct the formal asymptotic solution from Theorem 3.

LEMMA 9. *The coefficients of formal asymptotic solution* (35) *satisfy recurrent system of equations* (36), (41), (42), *and* (43).

Proof. Substitute (35) into (1) and collect the terms of the same order with respect to ε. This yields $N+1$ equations and a residual of the order ε^{N+1}. After collecting the terms with the same phase functions, we obtain the recurrent system of the equations for the coefficients of (35).

The terms with the phase function φ/ε^2 and of the order ε^1 give us (36) for the phase function of eigenoscillations. The initial data is determined by the matching condition and represented by the value of the driven phase S on the resonance curve $l = 0$,

$$\varphi|_{l=0} = S|_{l=0}, \quad \partial_{t_2}\varphi|_{l=0} = \partial_{t_2}S|_{l=0}.$$

The terms of the order ε^2,

$$2i\left(\partial_{t_2}\varphi\partial_{t_1}\Psi_{1,0,\varphi} - \partial_{x_2}\varphi\partial_{x_1}\Psi_{1,0,\varphi}\right) = 0,$$

give us the homogeneous transport equation

$$\partial_{t_2}\varphi\partial_{t_1}\Psi_{1,0,\varphi} - \partial_{x_2}\varphi\partial_{x_1}\Psi_{1,0,\varphi} = 0. \tag{37}$$

This equation allows us to determine the dependence with respect to the characteristic variable ζ of the leading-order term. Equation (37) along the characteristics

$$\frac{dx_1}{d\zeta} = -\partial_{x_2}\varphi, \quad \frac{dt_1}{d\zeta} = \partial_{t_2}\varphi \tag{38}$$

can be written in the form of an ordinary differential equation:

$$\frac{d\Psi_{1,0,\varphi}}{d\zeta} = 0. \tag{39}$$

This gives us that $\Psi_{1,0,\varphi}$ depends on ξ. The variable ξ is defined by

$$\frac{dx_1}{d\xi} = \partial_{t_2}\varphi, \quad \frac{dt_1}{d\xi} = \partial_{x_2}\varphi.$$

The terms of the order ε^3 which oscillate as $\exp(i\varphi/\varepsilon^2)$ are

$$\begin{aligned} &2i\left(\partial_{t_2}\varphi\partial_{t_1}\Psi_{2,0,\varphi} - \partial_{x_2}\varphi\partial_{x_1}\Psi_{2,0,\varphi}\right) \\ &+ 2i\partial_{t_2}\varphi\partial_{t_2}\Psi_{1,0,\varphi} + [(\partial_{t_1}\xi)^2 - (\partial_{x_1}\xi)^2]\partial^2_{\xi\xi}\Psi_{1,0,\varphi} \\ &+ i[\partial^2_{t_2}\varphi - \partial^2_{x_2}\varphi]\Psi_{1,0,\varphi} + \gamma|\Psi_{1,0,\varphi}|^2\Psi_{1,0,\varphi} = 0. \end{aligned}$$

It is convenient to write this equation in the form of an ordinary differential equation in terms of the characteristic variables:

$$\frac{d\Psi_{2,0,\varphi}}{d\zeta} = -2i\partial_{t_2}\varphi\partial_{t_2}\Psi_{1,0,\varphi} - [(\partial_{t_1}\xi)^2 - (\partial_{x_1}\xi)^2]\partial^2_{\xi\xi}\Psi_{1,0,\varphi}$$
$$- i[\partial^2_{t_2}\varphi - \partial^2_{x_2}\varphi]\Psi_{1,0,\varphi} - \gamma|\Psi_{1,0,\varphi}|^2\Psi_{1,0,\varphi}. \tag{40}$$

Equation (39) shows that the right-hand-side of (40) does not depend on ζ. To avoid the secular terms in the asymptotics we demand that the right-hand side of the equation be equal to zero. This allows us to determine the dependence of the leading-order term on the slow variable t_2,

$$2i\partial_{t_2}\varphi\partial_{t_2}\Psi_{1,0,\varphi} + [(\partial_{t_1}\xi)^2 - (\partial_{x_1}\xi)^2]\partial^2_{\xi\xi}\Psi_{1,0,\varphi}$$
$$+ i[\partial^2_{t_2}\varphi - \partial^2_{x_2}\varphi]\Psi_{1,0,\varphi} + \gamma|\Psi_{1,0,\varphi}|^2\Psi_{1,0,\varphi} = 0. \tag{41}$$

The equations for the higher-order terms are obtained in the same manner,

$$2i\left(\partial_{t_2}\varphi\partial_{t_1}\Psi_{n+1,k,\varphi} - \partial_{x_2}\varphi\partial_{x_1}\Psi_{n+1,k,\varphi}\right) = 2i\partial_{t_2}\varphi\partial_{t_2}\Psi_{n,k,\varphi} - \partial^2_{\xi\xi}\Psi_{n,k,\varphi}$$
$$- i[\partial^2_{t_2}\varphi - \partial^2_{x_2}\varphi]\Psi_{n,k,\varphi} + \partial_{t_1}\xi\partial^2_{\xi t_2}\Psi_{n-1,k,\varphi}$$
$$- \gamma \sum_{k_1,k_2,l_1,l_2,m_1,m_2,\alpha,\beta,\delta} \Psi_{k_1,k_2,\alpha}\Psi_{l_1,l_2,\beta}\Psi_{m_1,m_2,\delta},$$

where $k_1 + l_1 + m_1 = n + 2,\ k_2 + l_2 + m_2 = k,\ \alpha + \beta + \delta = \varphi,\ \alpha \in K_{k_1,k_2},\ \beta \in K_{l_1,l_2},\ \delta \in K_{m_1,m_2}$.

To construct the uniform asymptotic expansion with respect to ζ we obtain the linearized Schrödinger equation for higher-order term

$$2i\partial_{t_2}\varphi\partial_{t_2}\Psi_{n,k,\varphi} + \partial^2_{\xi\xi}\Psi_{n,k,\varphi} + i[\partial^2_{t_2}\varphi - \partial^2_{x_2}\varphi]\Psi_{n,k,\varphi}$$
$$= -\partial_{t_1}\xi\partial^2_{\xi t_2}\Psi_{n-1,k,\varphi} - \gamma \sum_{k_1,k_2,l_1,l_2,m_1,m_2,\alpha,\beta,\delta} \Psi_{k_1,k_2,\alpha}\Psi_{l_1,l_2,\beta}\Psi_{m_1,m_2,\delta}, \tag{42}$$

where $k_1 + l_1 + m_1 = n + 2,\ k_2 + l_2 + m_2 = k,\ \alpha + \beta + \delta = \varphi,\ \alpha \in K_{k_1,k_2},\ \beta \in K_{l_1,l_2},\ \delta \in K_{m_1,m_2}$.

The amplitudes $\Psi_{n,\chi}$ at $\chi \neq \pm\varphi$ are determined by the algebraic equations

$$\left[-(\partial_{t_2}\chi)^2 + (\partial_{x_2}\chi)^2 + 1\right]\Psi_{n,k,\chi} = F_{n,k,\chi}, \quad \chi \neq \pm\varphi. \tag{43}$$

Here the right hand-side of the equation depends on the previous terms and their derivatives

$$F_{n,k,\chi} = -2i\chi_{t_2}\partial_{t_1}\Psi_{n-1,k,\chi} + 2i\chi_{x_2}\partial_{x_1}\Psi_{n-1,k,\chi} - 2i\chi_{t_2}\partial_{t_2}\Psi_{n-2,k,\chi}$$
$$- i\left[\chi_{t_2t_2} - \chi_{x_2x_2}\right]\Psi_{n-2,k,\chi} - \partial^2_{t_1t_2}\Psi_{n-3,k,\chi} - \partial^2_{t_2t_2}\Psi_{n-4,k,\chi}$$
$$- \gamma \sum_{k_1,k_2,l_1,l_2,m_1,m_2,\alpha,\beta,\delta} \Psi_{k_1,k_2,\alpha}\Psi_{l_1,l_2,\beta}\Psi_{m_1,m_2,\delta},$$

where $k_1 + l_1 + m_1 = n - 4,\ k_2 + l_2 + m_2 = k,\ \alpha + \beta + \delta = \chi,\ \alpha \in K_{k_1,k_2},\ \beta \in K_{l_1,l_2},\ \delta \in K_{m_1,m_2}$.

These equations are similar to the equations for the amplitudes from the pre-resonance section. Lemma 9 is proved. □

The right-hand side of (42) has a singularity as $l \to 0$. The singularity appears due to $\Psi_{n,k,\chi}$ at $\chi \neq \pm\varphi$. The analysis of the right-hand side of the equation allows us to calculate the order of singularity as $l \to 0$. It is equal to $O(l^{-(n-1)})$. Below we prove the solvability of (42) with the given asymptotics as $l \to 0$.

LEMMA 10. *The asymptotics as $l \to 0$ of the solution of* (42) *has the form*

$$\Psi_{n,k,\varphi}(x_1,t_1,t_2) = \sum_{j=-(n-2)}^{1} \sum_{m=0}^{j-1} \Psi^{j,m}_{n,k,\varphi}(x_1,t_1)\; l^j (\ln l)^m + O(1), \quad l \to 0. \tag{44}$$

Proof. Determine the order of the singularity of the right-hand side of the equation as $l \to 0$. Consider (42) for $n = 3, k = 0$. The solution of this equation gives us the coefficient $\Psi_{3,0,\varphi}$. The nonlinearity contains the term $|\Psi_{2,0,S}|^2\Psi_{1,0,\varphi}$. The function $\Psi_{2,0,S}$ has the singularity of the order l^{-1} as $l \to 0$. It determines the order of the singularity for the right-hand side l^{-2}. We construct the asymptotics of $\Psi_{3,0,\chi}$ in the form

$$\Psi_{3,0,\varphi} = \Psi^{-1,0}_{3,0,\varphi} l^{-1} + \Psi^{0,1}_{3,0,\varphi} \ln(l) + \Psi^{1,1}_{3,0,\varphi} l \ln(l) + \widehat{\Psi}_{3,0,\varphi}. \tag{45}$$

Substitute (45) into (42) for $n = 3$. It leads to a recurrent system of equations for coefficients $\Psi^{(j,k)}_{3,0,\varphi}$:

$$-2i\partial_{t_2}\varphi\partial_{t_2} l\Psi^{(-1,0)}_{3,0,\varphi} = -\Psi_{1,0,\varphi}|\Psi_{2,0,S}|^2 l^2,$$

$$2i\partial_{t_2}\varphi\partial_{t_2} l\Psi^{(0,1)}_{3,0,\varphi} = L[\Psi^{(-1,0)}_{3,0,\varphi}],$$

$$2i\partial_{t_2}\varphi\partial_{t_2} l\Psi^{(1,1)}_{3,0,\varphi} = L[\Psi^{(0,1)}_{3,0,\varphi}].$$

Here we denote the linear operator by

$$L[\Psi] = 2i\partial_{t_2}\varphi\partial_{t_2}\Psi + \partial^2_\xi\Psi + i[\partial^2_{t_2}\varphi - \partial^2_{x_2}\varphi]\Psi + \gamma\big(2|\Psi_{1,0,\varphi}|^2\Psi + (\Psi_{1,0,\varphi})^2\Psi^*\big).$$

The regular part $\widehat{\Psi}_{3,0,\varphi}$ of the asymptotics satisfies the nonhomogeneous linear Schrödinger equation. The right-hand side of the equation is smooth,

$$L[\widehat{\Psi}_{3,0,\varphi}] = -l\ln|l| L[\Psi^{(1,1)}_{3,0,\varphi}] - 2i\partial_{t_2}\varphi\partial_{t_2} l\Psi^{(1,1)}_{3,0,\varphi}.$$

The initial condition for the regular part of the asymptotics is determined below by matching with the internal asymptotic expansion.

The structure of the terms $\Psi_{n,k,\pm\varphi}$ for $n > 3$ has a similar form. The right-hand side of (42) depends on junior terms. These singularities can be eliminated:

$$F_{n,k,\varphi} = \sum_{j=0}^{-(n-2)} \sum_{m=0}^{-j+1} l^j \ln^m|l| f^{(j,m)}_{n,k,\varphi} + \widehat{F}_{n,k,\varphi}.$$

The coefficients $f^{(j,m)}_{n,k,\varphi}$ do not contain singularities as $l \to 0$. These coefficients are easy calculated.

The direct substituting of (44) into (42) and collecting the terms with the same order of l completes the proof of Lemma 10. □

Thus we complete step 1 of the proof of Theorem 3.

4.2. The domain of validity of the second external asymptotics and matching procedure. The domain of validity of the second external asymptotics is determined by

$$\varepsilon \max_{\xi,t_2,x_2} |V_{n+1}| = o\left(\max_{\xi,t_2,x_2} |V_n| \right), \quad \varepsilon \to 0.$$

Formulas (35) and (44) give the condition

$$l \ll \varepsilon.$$

The domain $|l| \ll 1$ of validity of the internal asymptotics and domain of validity of the second external asymptotics intersect. This fact allows us to complete the construction of the second external asymptotics by a matching method [17]. The structure of singular parts of the internal asymptotics as $\lambda \to +\infty$ and external asymptotics as $l \to 0$ are equivalent. The coefficients are coincident due to our construction. The matching of regular parts of these asymptotics takes place due to

$$\Psi_{n,0,\varphi}|_{l=0} = W_n^{(0,0)}(\xi).$$

The function $W_n^{(0,0)}(\xi)$ is determined in Lemma 6.

In particular, the initial condition for the leading-order term has the form

$$\Psi_{1,0,\varphi}|_{l=0} = \int_{-\infty}^{\infty} d\sigma f(x_1) \exp\left(i \int_0^{\sigma} d\chi \lambda(x_1, t_1, \varepsilon) \right).$$

The soliton theory for the nonlinear Schrödinger equation leads us to the fact that the function $\Psi_{1,0,\varphi}$ contains the solitary waves when $f(x_1)$ is sufficiently large.

Theorem 3 is now proved. □

5. Summary. In this work we found the formula connecting the form of the resonant pumping and the shape of the solution after the slow passage through the resonance. In particular, it gives the mathematical basis for a solution of the important problem of nonlinear optics about the generation of the solitary packets of nearly monochromatic weakly nonlinear dispersive waves. We present the obvious form for the perturbation which generates such packets with a soliton as an envelope function.

The proposed approach opens ways for solving similar problems in pure and applied mathematics. The pure mathematical problem is a justification of the constructed asymptotic formulas obtained in this work. The applied problems are the generation of the solitary packet of waves with few carrier frequencies. One more applied problem is a pumping of the amplitudes of nonlinear waves up to the order of 1. Such resonant pumping would be useful for magnetics, Josephson junctions, and other applications in physics.

Acknowledgments. We are grateful to I. V. Barashenkov, L. A. Kalyakin, and B. I. Suleimanov for helpful comments and for help in improving the mathematical presentation of the results.

REFERENCES

[1] J. Kevorkian, *Passage through resonance for a one-dimensional oscillator with slowly varying frequency*, SIAM J. Appl. Math., 20 (1971), pp. 364–373.

[2] L. Rubenfeld, *The passage of weakly coupled nonlinear oscillators through internal resonance*, Stud. Appl. Math., 57 (1977), pp. 77–92.
[3] J. C. Neu, *Resonantly interacting waves*, SIAM J. Appl. Math., 43 (1983), pp. 141–156.
[4] L. A. Kalyakin, *Lokal'nyi rezonans v slabonelineinykh zadachakh*, Mat. Zametki, 44 (1988), pp. 697–699 (in Russian).
[5] S. G. Glebov, *O slabonelineinoi zadache s lokal'nym rezonansom*, Differ. Uravn., 31 (1995), pp. 1402–1408 (in Russian).
[6] P. L. Kelley, *Self-focusing of optical beams*, Phys. Rev. Lett., 15 (1965), pp. 1005–1008.
[7] V. I. Talanov, *O Samofokusirovke Malykh Puchkov v Nelineinykh Sredakh*, Pis'ma v ZhETF, 1965, n2, pp. 218–222.
[8] V. E. Zaharov, *Ustoichivost' periodicheskikh voln s konechnoi amplitudoi na poverkhnosti glubokoi zhidkosti*, Zh. Prikladnoi Mekh. Tekh. Fiz. 1968, n2, pp. 86–94.
[9] L. A. Kalyakin, *Dlinnovolnovye asymptotiki.Integriruemye uravneniya kak asymptoticheskii predel nelineinykh sistem*, Uspekhi Mat. Nauk, 44 (1989), pp. 5–34.
[10] B. B. Kadomtsev and V. I. Petviashvili, *Ob ustoichivosti uedinennykh voln v slabodispergiruyuschikh sredakh*, DAN SSSR, 194 (1970), pp. 753–756.
[11] S. P. Novikov, S. V. Manakov, L. P. Pitaevskij, and V. E. Zakharov, *Theory of Solitons*, Consultants Bureau, Culver City, CA, 1984.
[12] N. V. Alexeeva, I. V. Barashenkov, and D. E. Pelinovsky, *Dynamics of the parametrically driven NLS solitons beyond the onset of the oscillatory instability*, Nonlinearity, 12 (1999), pp. 103–140.
[13] R. R. Gadylshin and O. M. Kiselev, *On solitonless structure of the perturbed soliton solution for the Davey–Stewartson equation*, Theoret. and Math. Phys., 106 (1996), pp. 167–173.
[14] L. Friedland and A. G. Shagalov, *Excitation of solitons by adiabatic multiresonant forcing*, Phys. Rev. Lett., 8 (1998), pp. 4357–4360.
[15] S. G. Glebov, O. M. Kiselev, and V. A. Lazarev, *Soliton generation by local resonance*, Proc. Steklov Inst. Math., Suppl., 1 (2003), pp. S84–S90.
[16] A. Jeffrey and T. Kawahara, *Asymptotic Methods in Nonlinear Wave Theory*, Pitman, Boston, 1982.
[17] A. M. Il'in, *Matching of Asymptotic Expansions of Solutions of Boundary Value Problem*, AMS, Providence, RI, 1992.

SIAM J. APPL. MATH.
Vol. 65, No. 6, pp. 2178–2198

INSTABILITY OF THE IONOSPHERIC PLASMA: MODELING AND ANALYSIS*

CHRISTOPHE BESSE†, JEAN CLAUDEL‡, PIERRE DEGOND†, FABRICE DELUZET†, GÉRARD GALLICE‡, AND CHRISTIAN TESSIERAS‡

Abstract. This paper is concerned with the theory and modeling of plasma instabilities in the ionosphere. We first consider the so-called striation model, which consists of balance equations for the density and momenta of the plasma species, coupled with an elliptic equation for the potential. The linearized instability of this model is analyzed in the framework of Fourier theory, both for smooth and discontinuous steady states. Then, we show that the dissipation mechanisms at work in the more refined "dynamo model" allow us to stabilize high wave-number perturbations. We also analyze turbulence as a possible source of additional dissipation (in a similar way as in fluid mechanics). To this aim, we use the statistical approach to turbulence and derive a so-called turbulent striation model, of which we analyze the stability properties. Numerical experiments are used to support our investigations.

Key words. Euler–Maxwell system, dynamo model, striation model, ionospheric plasma, striations, turbulence, statistical approach, linearized stability analysis

AMS subject classifications. 82D10, 76N99, 76X05, 76W06, 78M35

DOI. 10.1137/040606582

1. Introduction.

This paper is concerned with the modeling and analysis of plasma instabilities in the ionosphere, at altitudes ranging between a few hundred and a thousand kilometers (the F region). The plasma may be created, either by the natural ionization of the atmosphere or by possible artificial causes (such as, e.g., thermonuclear explosions [31], [43], [19]). The ionospheric plasma is strongly structured by the earth's magnetic field. Indeed, the mobility of the ionized species (i.e., their velocity in response to an external electric field) is strongly anisotropic: while field-aligned mobilities (i.e., mobilities in the direction of the magnetic field) are large, transverse mobilities (also called Pedersen mobilities) are quite small. Additionally, a component of the plasma velocity orthogonal to both the electric and magnetic fields appears as a result of the Hall effect. This component is the major actor in the so-called $E \times B$ drift instability, which we are going to discuss in the present paper.

At lower altitudes, the density of the neutral atmosphere is large, and the plasma is dragged by the motion of the neutral molecules (or neutral wind). As a result, a net electrical current flows across the magnetic field lines and generates an induced electric field. This is the so-called ionospheric dynamo effect [1]. The reader can refer to [37], [21], [9], and [2] for reviews about ionospheric physics. In the presence of a gradient in the plasma density, the neutral wind can trigger the $E \times B$ drift instability,

*Received by the editors April 11, 2004; accepted for publication (in revised form) March 29, 2005; published electronically September 8, 2005. This work was supported by the "Centre d'Etudes Scientifiques d'Aquitaine" of the "Commisariat à l'Energie Atomique" and by the European network HYKE, funded by the EC as contract HPRN-CT-2002-00282.

http://www.siam.org/journals/siap/65-6/60658.html

†Mathématiques pour l'Industrie et la Physique, CNRS UMR 5640, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse Cedex 4, France (besse@mip.ups-tlse.fr, degond@mip.ups-tlse.fr, deluzet@mip.ups-tlse.fr).

‡Commissariat à l'Energie Atomique, Centre d'Etudes Scientifiques et Techniques d'Aquitaine, BP2, 33114 Le Barp, France (Jean.Claudel@cea.fr, Gerard.Gallice@cea.fr, Christian.Tessieras@cea.fr).

which bears strong similarities to the Rayleigh–Taylor instability in fluid mechanics [10]. This instability produces strong inhomogeneities (the ionospheric striations), which soon propagate over hundreds of kilometers along the magnetic field lines. The generation of plasma irregularities is reviewed in [13], [14], [36].

Our goal is to discuss some aspects of the mathematical and numerical modeling of this instability. Striations as well as related instability phenomena of the ionospheric plasma have been the subject of a wide literature (see, e.g., discussions of the "Spread F" in [44], [26], [34]; of the equatorial electrojet in [8], [41], [38], [39]; and of Barium releases experiments in [12]). The well-accepted mathematical model for these phenomena is the "dynamo" model [44], [12], which consists of mass and momentum balance equations for the plasma species. A simpler model, the "striation" model, is obtained when the field-aligned mobilities are supposed infinite. The derivation of these models and their interrelations are reviewed in [3] and will be briefly recalled in section 2.

The $E\times B$ drift instability is well described in the framework of the striation model (see section 3). A linear stability analysis indeed shows that exponential density profiles are unstable (see the review in [14] and section 3.2). Exponential density profiles are the only nonconstant smooth stationary states which allow analytical computations (via Fourier analysis). However, they are quite unrealistic, and a better theory should consider discontinuous density profiles. We consider this problem in section 3.3 and show that the striation model is also unstable in this case for certain configurations of the neutral wind. In a companion work [6], we show that smooth density profiles which are linearly unstable are nonlinearly unstable. However, the proof of [6] does not extend to discontinuous solutions. Similarly, we do not know, even for smooth density profiles, whether linear stability implies nonlinear stability.

In practice, the instability saturates and cascades towards smaller scales by nonlinearity [13], [39], [41], [25], [33], until it is ultimately damped by dissipation. In the striation model, however, all dissipation mechanisms have been removed. In section 4, we reintroduce dissipation effects by considering the dynamo model, where both finite temperature and finite conductivity effects are retained. A linearized stability analysis shows that high-wave-number perturbations are damped. However, in practice, the magnitude of the dissipation is too small, and we must consider other sources of dissipation.

In this paper, we investigate the possible influence of fluid turbulence. In fluid mechanics, it is a well-known fact that turbulence may enhance dissipation (see [32] and references therein). The statistical approach to fluid turbulence considers averages of the Navier–Stokes equations over various approximate realizations of the same solution. The chains of resulting statistical equations are closed by various types of phenomenological assumptions, which are still mathematically unjustified except in very simple cases, such as that considered in [27]. The models obtained (such as the K-ϵ model) involve additional terms compared with the standard Navier–Stokes equations, which describe the enhancement of diffusion by turbulence.

In section 5, we develop a similar statistical framework to model turbulence within the striation model (see also [28] for an application to MHD theory). We first derive an averaged striation model, for which we propose a closure ansatz inspired by [27]. This leads to a diffusive version of the striation model, the "turbulent striation model." To find the value of the turbulent diffusion constant, a stability analysis of the model is performed. It allows us to relate the threshold wave-number for instability (i.e., the typical size of the finest persisting structures in the plasma, which can be

experimentally observed) with the value of this constant.

Section 6 is devoted to two-dimensional numerical simulations. Their goal is to provide numerical and quantitative evidence of the features predicted in section 5, namely to show that the turbulent striation model produces persisting structures whose typical sizes are related to the magnitude of the diffusion. Three-dimensional simulations of the striation model are shown in [4]. A review collecting material from [3], [4], [6] as well as from the present paper is presented in [5].

Turbulence modeling in ionospheric plasmas has been widely investigated in the literature. Most of the approaches rely on nonlinear Fourier analysis [39], [22], [23] and bear similarities with the spectral approach to turbulence in fluid mechanics [29] (see also [15] for applications of these ideas to other plasma physics contexts). In using the statistical approach, we have chosen a slightly different route.

2. The "dynamo" and "striation" models of the ionospheric plasma. We consider two different species of particles: electrons and one-ion species. They are assumed so dilute that they have no influence on the dynamics of the neutrals, the velocity of which $u_n(x,t)$ (also called the neutral wind) is supposed known. In [3], a hierarchy of models for the ionospheric plasma has been derived. Of particular interest in the present study are the "dynamo" and "striation" models. The dynamo model is written as follows:

$$\partial_t n + \nabla \cdot (n u_i) = 0, \tag{2.1}$$

$$-\nabla\phi + u_{e,i} \times B = \kappa q_{e,i} \left[\nu_{e,i}(u_{e,i} - u_n) + \eta \nabla \log n \right], \tag{2.2}$$

$$\nabla \cdot j = 0, \quad \kappa j = n(u_i - u_e), \tag{2.3}$$

where we denote by $n(x,t)$ the density of the plasma; $u_e(x,t)$, $u_i(x,t)$ the electron and ion velocities; $j(x,t)$ the plasma current; $\phi(x,t)$ the electric potential; $B(x)$ the earth magnetic field; and $\nu_e(x)$, $\nu_i(x)$ the electron-neutral and ion-neutral collision frequencies, respectively. These quantities depend on the three-dimensional position coordinate x and on the time $t \geq 0$. The parameters η and κ are dimensionless and defined below. Equation (2.2) actually consists of two equations, one for the electrons (with the index "e" chosen everywhere) and one for the ions (with the index "i"). We let $q_i = 1$, $q_e = -1$. We suppose that the geomagnetic field $B(x)$ is unperturbed by the presence of the plasma and is known. Similarly, the collision frequencies $\nu_e(x)$, $\nu_i(x)$, which primarily depend on the neutral density, are supposed known. The plasma is supposed quasi-neutral; i.e., the electron and ion densities coincide with n. Despite the quasi-neutrality, the electron and ion velocities can be different, giving rise to a nonzero plasma current j. We have supposed that the electron and ion gases are isothermal with the same uniform temperature, which seems a valid physical hypothesis for the earth ionosphere [7]. The ionization-recombination terms which should appear on the right-hand side (r.h.s.) of (2.1) have been omitted as well. Typical ionization-recombination times are of the order of several hours, which is about the typical growth time of the instability. Therefore, these terms would make only a small correction to the analysis below and have been omitted for the sake of simplicity.

System (2.1)–(2.3) is written in dimensionless units. The scaling units and their typical values in the situations of interest are summarized in Table 2.1 below. The parameters η and κ are given by

$$\eta = \frac{k_B \bar{T}}{m_i \bar{u}^2} \frac{1}{\bar{\nu}_i \bar{t}} \qquad \kappa = \frac{m_e \bar{\nu}_e}{e\bar{B}} = \frac{m_i \bar{\nu}_i}{e\bar{B}}$$

TABLE 2.1
Scaling units.

Quantity	Scaling unit	Value
Time	$\bar{t}$	10^3 s
Length	$\bar{x}$	10^5 m
Velocity	$\bar{u} = \bar{x}/\bar{t}$	10^2 ms^{-1}
Density	$\bar{n}$	10^{12} m^{-3}
Temperature	$\bar{T}$	10^3 K
Magnetic field	$\bar{B}$	10^{-5} T
Electric potential	$\bar{\phi} = \bar{u}\bar{B}\bar{x}$	10^2 V
e-n collision frequency	$\bar{\nu}_e$	10^2 s^{-1}
i-n collision frequency	$\bar{\nu}_i = \frac{m_e}{m_i}\bar{\nu}_e$	10^{-2} s^{-1}

(where k_B is the Boltzmann constant) and respectively measure the ratio of the thermal energy to the ion drift energy and the typical number of electron-neutral or ion-neutral collisions per rotation period in the geomagnetic field. These two parameters have typical values (according to Table 2.1) $\eta \sim 10^1$, $\kappa \sim 10^{-4}$. Since κ is small, it is meaningful to investigate the limit of the dynamo model when $\kappa \to 0$. This leads to the so-called striation model that we give in more detail below.

Before doing so, we rewrite the dynamo model in a more appropriate form. In a local reference frame in which the last basis vector is aligned with the magnetic field, the ion and electron mobility matrices $\mathbb{M}_e$ and $\mathbb{M}_i$ are given by

$$\mathbb{M}_e = \begin{pmatrix} \mu_e^P & -\mu_e^H & 0 \\ \mu_e^H & \mu_e^P & 0 \\ 0 & 0 & \mu_e^\parallel \end{pmatrix}, \quad \mathbb{M}_i = \begin{pmatrix} \mu_i^P & \mu_i^H & 0 \\ -\mu_i^H & \mu_i^P & 0 \\ 0 & 0 & \mu_i^\parallel \end{pmatrix},$$

where the electron and ion Pedersen, Hall, and field-aligned mobilities are respectively defined by

$$\mu_{e,i}^P = \frac{\kappa\nu_{e,i}}{(\kappa\nu_{e,i})^2 + |B|^2}, \quad \mu_{e,i}^H = \frac{|B|}{(\kappa\nu_{e,i})^2 + |B|^2}, \quad \mu_{e,i}^\parallel = \frac{1}{\kappa\nu_{e,i}}.$$

In the situation $\kappa \to 0$, the electron or ion field-aligned mobilities tend to infinity.

Thanks to the mobility matrices, (2.2) and (2.3) may be rewritten as

$$u_{e,i} = \mathbb{M}_{e,i}\left(-q_{e,i}\nabla\phi + \kappa(\nu_{e,i}u_n - \eta\nabla\log n)\right), \tag{2.4}$$

$$-\nabla\cdot(n(\mathbb{M}_i + \mathbb{M}_e)\nabla\phi) = -\kappa\nabla\cdot(n[\mathbb{M}_i(\nu_i u_n - \eta\nabla\log n) - \mathbb{M}_e(\nu_e u_n - \eta\nabla\log n)]). \tag{2.5}$$

It is clear that the conductivity matrix $n(\mathbb{M}_i + \mathbb{M}_e)$ is positive definite (provided that ν_i or ν_e is positive and finite). Therefore, (2.5) is a three-dimensional elliptic equation for ϕ.

Now, we assume that the magnetic field is constant and uniform (see Figure 2.1). An extension to the nonuniform B case is given in [3] and [4]. Let us denote by $(\hat{x}_1, \hat{x}_2, \hat{x}_3)$ the orthonormal coordinate basis, with $\hat{x}_3$ aligned with B. We can choose the scaling units such that $|B| = 1$, so that $B = \hat{x}_3$. We denote by $\underline{x} = (x_1, x_2)$ the position vector in the two-dimensional plane orthogonal to B and by $\underline{\nabla} = (\partial_{x_1}, \partial_{x_2})$ the two-dimensional gradient. For any three-dimensional vector $a = (a_1, a_2, a_3)$, we define $\underline{a} = (a_1, a_2)$ as its projection onto this plane.

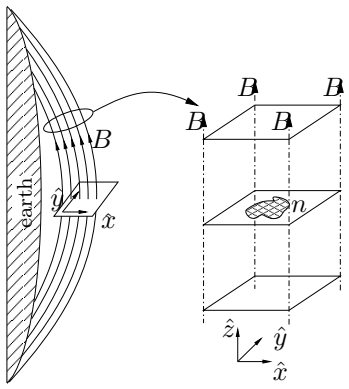


FIG. 2.1. *Geometry of the earth environment and reduction to a Cartesian geometry.*

When $\kappa \to 0$, the dynamo model reduces to the so-called striation model [3]:

$$\partial_t n + \nabla \cdot (nu) = 0, \tag{2.6}$$

$$u = -\underline{\nabla}\phi \times B + ((u_n - \eta\nu^{-1}\nabla \log n) \cdot B)B, \tag{2.7}$$

$$\underline{\nabla} \cdot (-\sigma(\underline{x})\underline{\nabla}\phi + (U_n - 2\eta\underline{\nabla}N) \times B) = 0, \tag{2.8}$$

with $\phi = \phi(\underline{x})$, $\sigma(\underline{x}) = \int n\nu dx_3$, $U_n = \int n\nu u_n dx_3$, $N = \int n dx_3$, $\nu = \nu_i + \nu_e$, and $u_i = u_e = u$. The striation model couples a three-dimensional convection-diffusion equation (2.6), (2.7) for the density n with a two-dimensional elliptic equation (2.8) for the electric potential ϕ. The coefficients of the elliptic equation (2.8) involve integrals of n over x_3, i.e., along the magnetic field lines. The infinite conductivity of the plasma along the magnetic field lines constrains the electric potential to be constant along these lines, i.e., to depend only on the two-dimensional coordinate $\underline{x}$.

If we additionally suppose that u_n is orthogonal to B and that all data and unknowns are independent of x_3, the striation model reduces to the following monolayer striation model:

$$\partial_t n + \nabla \cdot (nu) = 0, \quad u = -\nabla\phi \times B, \tag{2.9}$$

$$\nabla \cdot (nh) = 0, \quad h = \nu\left(-\nabla\phi + (u_n - 2\eta\nu^{-1}\nabla \log n) \times B\right), \tag{2.10}$$

where now all variables and vectors are two-dimensional (except $B = \hat{x}_3$) and 2-dimensional vectors are now left without being underlined. The quantity h represents the electron-ion relative velocity. We remark that $\nabla \cdot u = 0$. Therefore, we can write relation (2.9) as

$$\partial_t n + (u \cdot \nabla)n = 0. \tag{2.11}$$

As we will next see, the pressure term $\nabla \log n$ does not change the linearized stability properties of the striation model. When $\eta = 0$, (2.10) becomes

$$\nabla \cdot (nh) = 0, \qquad h = \nu\left(-\nabla\phi + u_n \times B\right). \tag{2.12}$$

In the next section, we analyze the linearized stability of this model.

3. Stability analysis of the striation model.

3.1. Introduction and phenomenology. The striation model exhibits an instability, the gradient-drift or $E \times B$ drift instability [13], [14]. In a recent work [6], local-in-time existence and uniqueness of solutions for this model have been proven and, following the methodology of [18], [20], [11], [24], smooth stationary density profiles which are linearly unstable have been shown to be nonlinearly unstable. However, it is still open whether the converse is true. In [6], a variational formulation for the instability growth rate is given. In the present work, we are aiming at a more quantitative result for certain specific classes of stationary profiles.

We restrict ourselves to two particular kinds of steady-state profiles. The first ones are smooth with an exponentially increasing density in one direction; they have already been investigated [13], [14], and we shall only summarize the results. The second ones are discontinuous density profiles; their analysis is, to the best of our knowledge, new. In passing, we shall have to show that it is meaningful to consider discontinuous solutions of the striation model.

We first give a phenomenological view of the instability of the striation model. We consider a steady state consisting of a discontinuous density $n(x) = \underline{n}$ for $x_2 < 0$ and $n = \overline{n} > \underline{n}$ for $x_2 > 0$, with $\nabla\phi = 0$ and $u_n = (0, U)$. We slightly perturb the interface, which is now represented by the graph of the function $x_2 = \varepsilon \sin(\xi x_1)$, where ε represents the magnitude of the perturbation ($\varepsilon \ll 1$) and ξ is its spatial frequency.

The term $u_n \times B$ in (2.12) creates a charge modulation along the interface which is alternately positive and negative. A nonzero electric field $-\nabla\phi$ parallel to the interface with a similar sign modulation is generated according to (2.12). Then, by (2.9), a nonzero component of the velocity u in the direction normal to the interface is created with again an alternating sign. According to the sign of u_n, this component of the velocity tends to either damp the modulation of the interface or increase it. The former case is a stable one, while the latter is an unstable one. The precise geometric configuration is depicted in Figures 3.1 and 3.2.

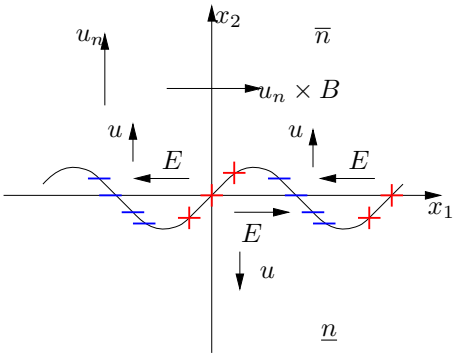


FIG. 3.1. *Stable configuration.*

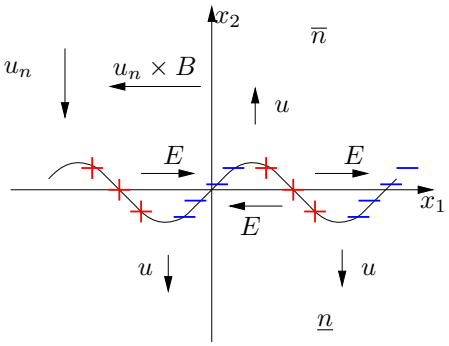


FIG. 3.2. *Unstable configuration.*

This behavior can be recovered by the linear stability analysis. We first turn to the analysis of the exponential density profile. In the remainder of this section, we assume that ν is a uniform constant and, by a convenient choice of the scaling units, we let $\nu = 1$.

3.2. Linear stability analysis: The exponential density profile. Let us denote by (n^0, u^0, h^0, ϕ^0) the unperturbed state, i.e., a time-independent solution

of the striation system (2.9), (2.12). We consider an exponential density profile in the x_2-direction, i.e., $n^0 = N\exp(x_2/\lambda)$, where $\lambda > 0$ is the gradient length, while $(u^0, h^0, \nabla\phi^0)$ are uniform constants. We suppose that the neutral wind is uniform as well and has components $u_n = (V, U)$. In this configuration, a necessary and sufficient condition for (n^0, u^0, h^0, ϕ^0) to form a steady state is that $u^0 = (V, 0)$, $h^0 = (\nu U, 0)$, $-\nabla\phi^0 = (0, V)$. In this analysis, we suppose that $\eta = 0$ unless otherwise specified.

Then, we introduce the perturbation $n = n^0(1+\varepsilon n^1 + O(\varepsilon^2))$, $(u, h, \phi) = (u^0, h^0, \phi^0) + \varepsilon(u^1, h^1, \phi^1) + O(\varepsilon^2)$, with $\varepsilon \ll 1$ in the striation model, and neglect the terms of order higher than ε. The neutral wind, being a datum, remains unperturbed. We note that $\nabla n^0 = (0, n^0/\lambda)$. An easy computation gives the linearized system governing the first order perturbation:

$$\partial_t n^1 + \lambda^{-1}\partial_{x_1}\phi^1 + V\partial_{x_1} n^1 = 0, \tag{3.1}$$

$$-\lambda^{-1}\partial_{x_2}\phi^1 - \Delta\phi^1 + U\partial_{x_1} n^1 = 0. \tag{3.2}$$

Remark 3.1. If the pressure gradient terms are retained in the model (i.e., if $\eta \neq 0$), the steady state is modified only through the expression of h^0, which should be taken as $h^0 = (U - 2\eta\lambda^{-1}, 0)$. However, the first order perturbation equations are the same as (3.1), (3.2). The details are left to the reader.

We develop the solution of (3.1), (3.2) into plane waves, i.e., (dropping the superscripts "1" for clarity) $(n, \phi) = (\bar{n}, \bar{\phi}\lambda|U|)\exp(i\lambda^{-1}(\xi_1 x_1 + \xi_2 x_2 - \omega t|U|))$, where $\xi = (\xi_1, \xi_2)$ is the (normalized) wave-vector of the perturbation and ω its frequency. Introducing this ansatz into (3.1), (3.2), we get

$$-\omega\bar{n} + \xi_1\bar{\phi} = 0, \quad i\sigma\xi_1\bar{n} + (\xi_1^2 + \xi_2^2 - i\xi_2)\bar{\phi} = 0, \tag{3.3}$$

with $\sigma = \mathrm{sign}(U) \in \{-1, 1\}$. This system has a nontrivial solution iff its determinant is nonvanishing. This condition yields the dispersion relation

$$\omega = \frac{-i\sigma\xi_1^2}{(\xi_1^2 + \xi_2^2)^2 + \xi_2^2}(\xi_1^2 + \xi_2^2 + i\xi_2). \tag{3.4}$$

We now recall the following standard definition.

DEFINITION 3.2. *The perturbation is stable if n and ϕ stay bounded for all times $t \geq 0$ and unstable in the converse situation. Therefore, a perturbation is stable iff $\Im m(\omega) \leq 0$ and unstable iff $\Im m(\omega) > 0$. A stationary state is called stable if all its perturbations are stable for all wave vectors ξ. It is unstable as soon as there exists a wave vector ξ giving rise to an unstable perturbation.*

Thanks to (3.4), we have $\mathrm{sign}(\Im m(\omega)) = -\sigma$. We then conclude the following result.

PROPOSITION 3.3. *The steady-state configuration with an exponential density profile is stable iff $U \geq 0$, i.e., if the x_2-component of the neutral wind points in the same direction as the density gradient. Furthermore, in the case $U < 0$, all wave vectors $\xi \neq 0$ are unstable, and for $\xi_2 = 0$ the growth rate is independent of ξ_1.*

As seen above, exponential density profiles allow explicit computations. However, they are fairly unrealistic, as the density tends to infinity on one side and degenerates to zero (and the elliptic problem (2.12) as well) on the other side. In order to study a more realistic situation, we extend our analysis to the case of discontinuous density profiles in the next section.

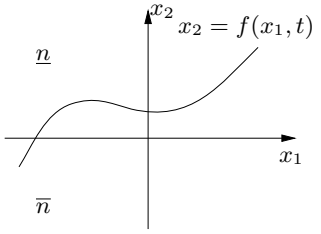


FIG. 3.3. *Discontinuity curve of n.*

3.3. Linear stability analysis: Discontinuous density profiles. We consider a density profile which is piecewise constant and discontinuous across a parametrized curve $C(t)$ given by the equation $x_2 = f(x_1, t)$, where $f \in C^1(\mathbb{R} \times [0, +\infty[)$, i.e., (see Figure 3.3)

$$n(x,t) = \begin{cases} \underline{n} & \text{for} \quad x_2 < f(x_1,t)\,, \\ \overline{n} & \text{for} \quad x_2 > f(x_1,t)\,. \end{cases} \tag{3.5}$$

First, we must give a meaning to discontinuous solutions of this kind. Toward this aim, we use the notion of a weak solution of (2.11).

DEFINITION 3.4. *Let $u \in C^1(\mathbb{R}^2 \times [0,\infty[)$. A function $n \in L^\infty_{loc}(\mathbb{R}^2 \times [0,\infty))$ is a weak solution of* (2.11) *with initial data n_0 iff n verifies*

$$\int_{\mathbb{R}^2 \times [0,\infty)} n \left(\frac{\partial \varphi}{\partial t} + \nabla \cdot (u\varphi) \right) dx\, dt + \int_{\mathbb{R}^2} n_0\, \varphi(x,0)\, dx = 0 \tag{3.6}$$

for all functions $\varphi(x,y,t) \in C^1_c(\mathbb{R}^2 \times [0,\infty))$, where C^1_c defines the space of functions of class C^1 with compact support.

The solution of (3.6) can be obtained through the method of characteristics. In particular, it satisfies the maximum principle. Therefore, if there exist two constants n_*, n^* such that $0 < n_* < n_0(x) < n^*$, this inequality is satisfied at all times: $0 < n_* < n(x,t) < n^*$.

This notion has to be extended to the case of discontinuous velocities. Suppose that $u = (u_1, u_2)$ is taken in the space $L^1_{\text{loc}}([0,\infty), H_{\text{div}})$, with $H_{\text{div}}(\mathbb{R}^2) = \{u \in L^2(\mathbb{R}^2)$, s.t. $\nabla \cdot u \in L^2(\mathbb{R}^2)\}$. Then, $\nabla \cdot (u\varphi) \in L^1_{\text{loc}}(\mathbb{R}^2 \times [0,\infty))$ for all test functions φ, and the expression (3.6)} still has a meaning. Now, in the striation model, u is a given by $u = -\nabla\phi \times B$, where ϕ is a solution of (2.12). To solve (2.12), we use the following (classical) proposition.

PROPOSITION 3.5. *Let $u_n \in L^2(\mathbb{R}^2)$ and n be such that there exist two constants n_*, n^* with $0 < n_* < n(x,t) < n^*$. Then,* (2.12)*, which can be written*

$$\nabla \cdot (n \nabla \phi) = \nabla \cdot (n u_n \times B),$$

has a solution in the space $L^1_{loc}([0,\infty), H)$, with $H = \{\phi \in \mathcal{D}'(\mathbb{R}^2),\ \nabla\phi \in L^2(\mathbb{R}^2)\}$, unique up to an additive constant.

Since u satisfies $\nabla \cdot u = 0$, this proposition guarantees that u belongs to $L^1_{\text{loc}}([0,\infty), H_{\text{div}}(\mathbb{R}^2))$. For such velocities, this allows us to define n as a weak solution of (2.11) in the sense of (3.6). Therefore, it is meaningful to look for solutions with discontinuous densities. Of course, we have not shown the actual existence of such solutions, which will be the subject of future work. Now, we recall the following classical trace property (see [16]).

LEMMA 3.6. *Let C be a regular orientable curve of $\mathbb{R}^2$. Then the mapping $\gamma_N : v \to (v \cdot N)|_C$ (with N the unit normal vector to C) defined on $\mathcal{D}(\mathbb{R}^2)$ can be extended by continuity to a linear and continuous mapping, still denoted by γ_N, from $H_{div}(\mathbb{R}^2)$ into $H^{-1/2}(C)$.*

We are now ready to determine the conditions that f must fulfill for n to be a weak solution. We have the following.

PROPOSITION 3.7. *Let u belong to $L^1_{loc}([0,\infty), H_{div}(\mathbb{R}^2))$. A function n defined by (3.5) is a weak solution to (2.11) iff f is a smooth solution to the equation*

$$\partial_t f = (u \cdot N)\left(1 + (\partial_{x_1} f)^2\right)^{1/2}, \quad (x_1, t) \in \mathbb{R} \times [0,\infty), \tag{3.7}$$

where N is the unit normal vector to the curve of discontinuity $C(t)$ (pointing towards $x_2 > 0$) and $(u \cdot N)$ is the trace along $C(t)$ as defined by Lemma 3.6. We can write $(u \cdot N)(1 + (\partial_{x_1} f)^2)^{1/2} = [u_2 - u_1 \partial_{x_1} f]\,|_C$, where the index C indicates that this quantity is the common limit of the bracket as $x_2 \to f(x_1, t)$ from above and below.

Proof. We insert the expression for n into (3.6). We get

$$0 = \underline{n} \int_{x_2 < f,\, t \geq 0} (\partial_t \varphi + \nabla \cdot (u\varphi))\, dx\, dt + \overline{n} \int_{x_2 > f,\, t \geq 0} (\partial_t \varphi + \nabla \cdot (u\varphi))\, dx\, dt\,.$$

Since φ is compactly supported, we have $\int_{\mathbb{R}^2 \times [0,\infty)} (\partial_t \varphi + \nabla \cdot (u\varphi))\, dx\, dt = 0$. We regard φ as compactly supported in $\mathbb{R}^2 \times (0,\infty)$ since the treatment of the initial condition at $t = 0$ is standard. We deduce that

$$0 = (\underline{n} - \overline{n}) \int_{x_2 < f,\, t \geq 0} (\partial_t \varphi + \nabla \cdot (u\varphi))\, dx\, dt\,. \tag{3.8}$$

In order to apply the Green formula, we use Lemma 3.6. We define the surface $\Sigma = \{(x,t) \in \mathbb{R}^2 \times [0,\infty),\, x \in C(t)\}$ and the open sets $\mathcal{O}(t) = \{x \in \mathbb{R}^2,\, x_2 < f(x_1, t)\}$ and $\Omega = \{(x,t) \in \mathbb{R}^2 \times [0,\infty),\, x \in \mathcal{O}(t)\}$. Let $\tilde{N} = (\tilde{N}_1, \tilde{N}_2, \tilde{N}_t)$ be the outgoing unit normal to Ω at (x,t) of Σ and $\tilde{N}_x = (\tilde{N}_1, \tilde{N}_2)$. Thanks to Lemma 3.6, we can apply the Green formula and get

$$\int_\Omega (\partial_t \varphi + \nabla \cdot (u\varphi))\, dx\, dt = \int_\Sigma \varphi \left(\tilde{N}_t + u \cdot \tilde{N}_x\right) d\Sigma(x,t)\,,$$

where the integrals on Σ should be understood as the duality $L^\infty([0,\infty), H^{1/2}(C(t)))$ against $L^1([0,\infty), H^{-1/2}(C(t)))$. Now, we have $\tilde{N} d\Sigma(x,t) = (-\partial_{x_1} f,\, 1,\, -\partial_t f)\, dx_1\, dt$, which implies $\tilde{N}_x d\Sigma(x,t) = (1 + (\partial_{x_1} f)^2)^{1/2} N\, dx_1\, dt$. Assuming that $\underline{n} \neq \overline{n}$, (3.8) gives

$$\int_{\mathbb{R} \times [0,\infty)} \varphi \left((u \cdot N)\left(1 + (\partial_{x_1} f)^2\right)^{1/2} - \partial_t f\right) dx_1\, dt = 0. \tag{3.9}$$

Since (3.9) has to be verified for all test functions φ, we deduce (3.7). ∎

Then, the striation model (2.9), (2.12) for weak solutions can be written

$$\partial_t f = [\partial_{x_2} \phi\, \partial_{x_1} f + \partial_{x_1} \phi]\,|_{C(t)}\,, \tag{3.10}$$

$$-\nabla \cdot ((\underline{n} \chi_f + \overline{n}(1 - \chi_f))(\nabla \phi - u_n \times B)) = 0, \tag{3.11}$$

with $\chi_f = 1$ if $x_2 < f(x_1, t)$ and $\chi_f = 0$ otherwise.

We now turn to the stability analysis of the striation model with discontinuous initial density. A steady state of this model is given by $f^0 = 0$, $\nabla\phi^0 = (0, -V)$, $u_n = (V, U)$. We define the small perturbations of order ε as $f = \varepsilon f^1$, $\phi = \phi^0 + \varepsilon\phi^1$, with $\varepsilon \ll 1$. We introduce this ansatz in (3.10), (3.11) and keep only the terms of order ε. We get

$$\left(\partial_t f^1 + V\partial_{x_1} f^1 - \partial_{x_1}\phi^1\right)|_{(x_1,0,t)} = 0, \tag{3.12}$$

$$-\nabla\cdot(n^0\nabla\phi^1) = -U\lim_{\varepsilon\to 0}\varepsilon^{-1}\partial_{x_1}(\underline{n}\chi_{(\varepsilon f^1)} + \overline{n}(1-\chi_{(\varepsilon f^1)})), \tag{3.13}$$

where $n^0 = \underline{n}\chi_0 + \overline{n}(1-\chi_0)$ is the unperturbed density profile. A simple computation leads to $\partial_{x_1}\chi_{(\varepsilon f)} = \varepsilon(\partial_{x_1} f)\delta_{x_2=0} + O(\varepsilon^2)$, where the distribution $g(x_1)\delta_{x_2=0}$ is defined through the relation $\langle g(x_1)\delta_{x_2=0}, \varphi\rangle = \int_{\mathbb{R}}\varphi(x_1,0)g(x_1)\,dx_1$ with $\varphi(x) \in C_c^\infty$. Then (3.13) reads

$$-\nabla\cdot(n^0\nabla\phi^1) = U(\overline{n}-\underline{n})\partial_{x_1} f^1\,\delta_{x_2=0}\,. \tag{3.14}$$

Like in the exponential density profile case, we develop the solution as a plane wave in the x_1 direction: $(f^1, \phi^1) = (\bar f, \bar\phi(x_2))\exp i(\xi x_1 - \omega t)$, where $\bar f$ and $\bar\phi(x_2)$ must be determined.

We introduce the plane-wave ansatz in (3.12), (3.14), and we get

$$-i\omega\bar f + iV\xi\bar f = i\xi\bar\phi(0)\,, \tag{3.15}$$

$$-\left(\partial_{x_2}\left(n^0\partial_{x_2}\bar\phi\right) - \xi^2 n^0\bar\phi\right) = U(\overline{n}-\underline{n})i\xi\bar f\delta_{x_2=0}\,. \tag{3.16}$$

If we solve (3.16) away from the point $x_2 = 0$ and look for a bounded solution when $|x_2| \to \infty$, we find $\bar\phi(x_2) = \bar\phi(0)e^{-|\xi||x_2|}$. Then, in the distributional sense on $\mathbb{R}$, we have

$$-\left(\partial_{x_2}\left(n^0\partial_{x_2}\bar\phi\right) - \xi^2 n^0\bar\phi\right) = -(\underline{n}+\overline{n})|\xi|\bar\phi(0)\delta_{x_2=0}\,. \tag{3.17}$$

We introduce (3.17) in (3.16) and find

$$i\xi\bar f\delta_{x_2=0} - \frac{\overline{n}+\underline{n}}{\overline{n}-\underline{n}}\frac{|\xi|}{U}\bar\phi(0)\delta_{x_2=0} = 0\,. \tag{3.18}$$

Solving for $\bar f$ thanks to (3.15) and inserting it into (3.18) allows us to find the dispersion relation

$$\omega = -V\xi - i\,U\frac{\overline{n}-\underline{n}}{\overline{n}+\underline{n}}\,\xi. \tag{3.19}$$

We can easily deduce the following stability result.

PROPOSITION 3.8. *Let us assume that $\overline{n} > \underline{n}$. Steady states defined by* (3.5) *with $f = 0$ are stable iff $U \geq 0$. Furthermore, if $U < 0$, all wave-vectors ξ are unstable. The growth rate of the instability is given by*

$$|\Im m(\omega)| = \frac{\overline{n}-\underline{n}}{\overline{n}+\underline{n}}\,U\,|\xi|. \tag{3.20}$$

This theorem makes the formal analysis of section 3.1 more quantitative. We note that the growth rate increases linearly as a function of ξ, while it was a constant in the exponential density case (section 3.2). This feature prevents us from extending the nonlinear instability theorem of [6] to the discontinuous case.

3.4. Conclusion of the stability analysis. The instability contributes to the development of smaller and smaller structures in the plasma. Quickly, the plasma becomes chaotic (see section 6). In practice, the instability saturates after reaching some level by the effects of physical dissipation mechanisms, which are not accounted for so far in the model. We can think of three sources of physical dissipation: (i) finite temperature effects, (ii) finite conductivity effects, (iii) turbulence effects. By Remark 3.1, we have seen that finite temperature effects alone do not change the results of the stability analysis. Therefore, we must simultaneously introduce finite temperature and finite conductivity effects; i.e., we must go back to the full dynamo model (2.1)–(2.3). In section 4, we perform the stability analysis of the dynamo model and show that large wave-vector perturbations are stable. However, the level of dissipation, i.e., the threshold wave-number for stability, is too large to match practical observations. Therefore, in section 5, we investigate the effects of fluid turbulence.

4. Stability analysis of the dynamo model (2.1)–(2.3). This analysis can be found, in parts, in [13], [14]. We consider steady states with exponential density profiles. Due to diffusion, discontinuous density profiles are not steady states of the dynamo model any longer, and there is no point in trying to analyze their stability. To simplify the analysis and to make it as close as possible to that of the striation model in section 3.2, we still consider a uniform magnetic field pointing in the x_3-direction, and we suppose that u_n is orthogonal to B. All unknowns are independent of x_3, and vectors are contained in the plane (x_1, x_2). We assume that ν_e and ν_i are constants and such that $\nu = \nu_e + \nu_i = 1$.

The steady state is given by

$$
\begin{aligned}
&n^0 = N \exp\left(\frac{x_2}{\lambda}\right), \quad u_n = (V, U), \quad h^0 = \left(U - 2\eta\lambda^{-1}\right)\hat{x}_1,\\
&u_i^0 = \left\{\kappa\nu_e(U - 2\eta\lambda^{-1}) + V\right\}\hat{x}_1, \quad u_e^0 = \left\{-\kappa\nu_i(U - 2\eta\lambda^{-1}) + V\right\}\hat{x}_1,\\
&\nabla\phi^0 = (-\kappa^2\nu_i\nu_e(U - 2\eta\lambda^{-1}), -\kappa(\nu_e - \nu_i)(U - \eta\lambda^{-1}) - V).
\end{aligned}
$$

We proceed to the linear stability analysis as in section 3.2. We introduce the perturbation $n = n^0(1 + \varepsilon n^1 + O(\varepsilon^2))$, $u_{i,e} = u_{i,e}^0 + \varepsilon u_{i,e}^1 + O(\varepsilon^2)$, and $\phi = \phi^0 + \varepsilon\phi^1 + O(\varepsilon^2)$ in the dynamo model (2.1)–(2.3). We keep only order ε terms and develop the solution as a plane wave according to the same ansatz as in section 3.2. Let us define

$$
\begin{aligned}
&\mu_-^H = \mu_i^H - \mu_e^H, \quad \mu_+^H = \mu_i^H + \mu_e^H, \quad \mu_-^P = \mu_i^P - \mu_e^P, \quad \mu_+^P = \mu_i^P + \mu_e^P,\\
&X = \xi_1\mu_i^H - \xi_2\mu_i^P, \quad Y = \xi_1\mu_-^H - \xi_2\mu_+^P, \quad Z = \xi_1\mu_+^H - \xi_2\mu_-^P,\\
&A_X = \mu_i^P|\xi|^2 + iX, \quad A_Y = \mu_+^P|\xi|^2 + iY, \quad A_Z = \mu_-^P|\xi|^2 + iZ.
\end{aligned}
$$

Then, we get the dispersion relation

$$
\omega = \frac{A_Y^*}{|U||A_Y|^2}\left(\xi_1 u_{ix}^0 A_Y - i\frac{\kappa\eta}{\lambda}A_X A_Y - \xi_1 h_x^0 A_X + i\frac{\kappa\eta}{\lambda}A_X A_Z\right), \tag{4.1}
$$

where $h = u_i - u_e$ and the star denotes the complex conjugate. The expression of $\Im m(\omega)$ may be simplified as follows:

$$
\Im m(\omega) = \frac{2\kappa\eta}{\lambda|U||A_Y|^2}\mu_i^P\mu_e^P\mu_+^P|\xi|^2 P(\xi), \qquad P(\xi) = -|\xi|^4 + (a+1)\xi_1^2 + 2c\xi_1\xi_2 - \xi_2^2, \tag{4.2}
$$

with $a = -U\lambda\,(\,2\kappa\eta\nu_e\nu_i\mu_+^P\,)^{-1}$, $c = \mu_-^H(\mu_+^P)^{-1}$.

If $\kappa \to 0$, the dynamo model reduces to the striation model with nonzero temperature, which has the same dispersion relation (3.4) as the striation model with zero temperature (see Remark 3.1). We can verify this property on the dispersion relation (4.2). Indeed, as $\kappa \to 0$, we have $a \sim -U\lambda\,(\,2\eta\nu_e\nu_i\,)^{-1}\kappa^{-2}$, $c \sim \kappa(\nu_e - \nu_i)$. Therefore, $P(\xi) \sim -U\lambda\,(\,2\eta\nu_e\nu_i\,)^{-1}\xi_1^2\kappa^{-2}$, and we have $P(\xi) < 0$ if $U > 0$ (resp., $P(\xi) > 0$ if $U < 0$). Thus, we recover the results of section 3.2 for the striation model.

On the other hand, if we let the temperature go to zero (i.e., $\eta \to 0$) in (4.2) while keeping κ finite, we find $a = O(\eta^{-1})$ while $c = O(1)$. Therefore, $P(\xi) \sim -U\lambda(2\kappa\nu_i\nu_e\mu_P^+\eta)^{-1}$, and again the stability conditions for this model are the same as those of the striation model; i.e., the model is unstable for all wave-vectors if $U < 0$ and stable otherwise. Therefore, for the model to exhibit a stable range of wave-vectors, we need at the same time a finite conductivity and a finite temperature. We now show that this is indeed the case.

PROPOSITION 4.1. *Suppose that $\eta > 0$ and $\kappa > 0$. Then,*

(i) *the dynamo model* (2.1)–(2.3)*, linearized about the above-defined steady states, is stable iff $U\lambda > 2\eta\kappa^2\nu_e\nu_i$;*

(ii) *if $U\lambda < 2\eta\kappa^2\nu_e\nu_i$, there exists $R_0(\eta,\kappa) > 0$ such that if ξ is an unstable wave-vector, then $|\xi| < R_0$. Furthermore, $R_0 = O((\sqrt{\eta}\kappa)^{-1})$ as $\sqrt{\eta}\kappa \to 0$.*

We note that the stability criterion for the dynamo model is more restrictive than that of the striation model. The quantity $U\lambda$ needs to be not only positive, but also large enough. However, in the unstable case, when $U\lambda$ is not large enough, the instability region is a bounded domain in wave-vector space (by contrast with the case of the striation model, in which the instability domain is unbounded). The instability region grows as η or κ decreases to 0, and ultimately fills the entire wave-vector space in the limit.

Proof. We introduce polar coordinates $\xi_1 = r\cos\theta$ and $\xi_2 = r\sin\theta$. We can write $P = -r^2Q$ with $Q = r^2 - F(\theta)$ with

$$F(\theta) = (a+1)\cos^2\theta + 2c\cos\theta\sin\theta - \sin^2\theta = \delta\cos(2\theta - \alpha) + \left(\frac{a}{2}\right)$$

and $\delta = (((a/2)+1)^2+c^2)^{1/2}$, $\alpha = \tan^{-1}(a/(a+2))$. Therefore, $Q = 0$ iff $\cos(2\theta-\alpha) = \delta^{-1}(r^2 - (a/2))$. For this equation to have roots, we need that $-1 \le \delta^{-1}(r^2 - (a/2)) \le 1$. Therefore, if $-a/(2\delta) > 1$, this equation cannot have any root, for any value of $r > 0$. Conversely, if $-a/(2\delta) \le 1$, this equation has roots as long as $(a/2) - \delta \le r^2 \le (a/2) + \delta$. Therefore, the case $-(a/(2\delta)) > 1$ characterizes the stable cases. This condition is equivalent to $a + 1 + c^2 < 0$, or, after some easy computations, to $U\lambda > 2\eta\kappa^2\nu_e\nu_i$. In the unstable case, the instability domain, i.e., the set of wave-vectors ξ leading to unstable modes, is contained in the ball centered at 0 and of radius R_0 with $R_0^2 = (a/2) + \delta$. As κ or η tend to 0, we notice that $R_0^2 \sim |U\lambda|(2\nu_e\nu_i)^{-1}(\eta\kappa^2)^{-1}$, which ends the proof of Proposition 4.1. ∎

The fact that the model is stable apart from a bounded region of wave-vectors can be seen as a favorable feature. Indeed, in such a case, small wave-vector (i.e., long wave-length) perturbations first grow exponentially due to the instability, but also undergo a mode cascade towards higher wave-numbers due to nonlinearity. Once the wave-vectors are large enough to reach the stability region, they are damped by the dissipation. We therefore expect that only structures of typical size R_0^{-1} will remain for long times.

However, the values of the physical parameters in the dynamo model are too small to ensure a viable stabilization process. Indeed, we see that $R_0 = O((\kappa^2\eta)^{-1/2})$

when κ or $\eta \to 0$. This is too large compared with the observations (see, e.g., [13], [14]). Therefore, another dissipation mechanism must be present. In this paper, we postulate that the turbulence of the plasma induced by the instability modifies the dissipation constants in a way similar to what happens in fluid mechanics (see, e.g., [32] and references therein). To make this assumption more quantitative, in the next section we develop a statistical approach to turbulence adapted to the striation model.

5. A "turbulent" striation model.

5.1. Derivation of the "turbulent" striation model. To produce this new model, we follow the statistical approach to turbulence [32] (see also [28] for an application to MHD). We suppose that the unknowns (n, u, ϕ, h) in the striation model (2.9), (2.12) are random variables representing the possible realizations of the flow. Any of these quantities a can be decomposed according to $a = \bar{a} + a'$, where $\bar{a}$ is its mean value and a' is a random fluctuation about this average. Since the randomness concerns the realization of the flow, the mean value operator commutes with the space and time derivatives. Therefore, we have $\overline{(\bar{a})} = \bar{a}$, $\overline{a'} = 0$, $\nabla a = \nabla \bar{a} + \nabla a'$, $\partial_t a = \partial_t \bar{a} + \partial_t a'$. If b is a nonfluctuating quantity, we have $\overline{ba} = b\bar{a}$, and for two random quantities a and b, $\overline{ab} \neq \bar{a}\bar{b}$ unless they are statistically independent. However, we note that $\overline{a\bar{b}} = \bar{a}\bar{b}$.

We assume that u_n and ν are nonfluctuating quantities. Under this assumption, by averaging the striation model (2.9)–(2.12), we obtain

$$\partial_t \bar{n} + \nabla \cdot (\overline{nu}) = 0, \quad \bar{u} = -\nabla \bar{\phi}, \tag{5.1}$$

$$\nabla \cdot (\overline{nh}) = 0, \quad \bar{h} = -\nu(\bar{u} - u_n) \times B. \tag{5.2}$$

We can write $\overline{nu} = \bar{n}\bar{u} + \overline{n'u'}$ with $\overline{n'u'} \neq 0$, since n' and u' are in general not independent random variables. In a same way, we have $\overline{nh} = -\nu(\bar{n}\bar{u} + \overline{n'u'} - \bar{n}u_n) \times B \neq \bar{n}\bar{h}$.

To close the model, we need a prescription for the correlation $\overline{n'u'}$ as a function of the mean quantities. As in fluid turbulence (see, e.g., [32]), we model this correlation by means of a diffusion term acting on the density, i.e., $\overline{n'u'} = -D\nabla\bar{n}$, where $D > 0$ is a diffusion coefficient. The use of this ansatz can be formally justified by invoking Kesten–Papanicolaou's theorem [27] (see also [32] for a review and [35] for a related result). For simplicity, we assume that D is a constant. Under this assumption, and noting that $\nabla \cdot (\overline{n'u'} \times B) = -D\nabla \cdot (\nabla n \times B) = 0$, system (5.1)–(5.2) reduces to the following (turbulent striation) model (dropping the bars):

$$\partial_t n + \nabla \cdot (nu) - \nabla \cdot (D\nabla n) = 0, \quad u = -\nabla\phi \times B, \tag{5.3}$$

$$\nabla \cdot (nh) = 0, \quad h = \nu\left(-\nabla\phi + u_n \times B\right). \tag{5.4}$$

The difficulty is now to find the correct value for the diffusion coefficient D. For this purpose, we again proceed to a stability analysis, in a similar fashion as what was done in sections 3.2 and 4.

5.2. Stability analysis of the turbulent striation model. We again choose a steady state characterized by an exponential density profile and uniform neutral wind $u_n = (V, U)$ and electric field. The unperturbed state is defined by $n_0 = Ne^{x_2/\lambda}$, $u_0 = (V, 1/\lambda) = (-\partial_{x_2}\phi_0, \partial_{x_1}\phi_0)$. We introduce $\bar{D}$ such that $D = |U|\lambda\bar{D}$. We proceed as in sections 3.2 and 4 and we get the following imaginary part of the dispersion

relation (with $\sigma = \text{sign}(U)$):

$$\Im m(\omega) = \frac{\mathcal{N}}{\mathcal{D}}, \quad \mathcal{N} = -\bar{D}|\xi|^2(|\xi|^4 + \xi_2^2) - (\sigma - \bar{D})\xi_1^2|\xi|^2, \quad \mathcal{D} = |\xi|^4 + \xi_2^2. \tag{5.5}$$

Since $\mathcal{D} \geq 0$, we just have to discuss the sign of $\mathcal{N}$. We introduce the polar coordinates $\xi_1 = r\cos\theta$ and $\xi_2 = r\sin\theta$. Then, $\mathcal{N} = -\bar{D}r^4(r^2 + \sin^2\theta - (1 - \sigma\bar{D}^{-1})\cos^2\theta)$. The domain $\mathcal{I}$ defined in polar coordinates by $r \leq r(\theta) := \{(1 - \sigma\bar{D}^{-1})\cos^2\theta - \sin^2\theta\}^{1/2}$ (for all θ such that the expression inside the square root is positive) characterizes the (bounded) instability domain. We can summarize the results in the following.

PROPOSITION 5.1. (i) *If $\sigma = 1$ (stable case for the original striation model) and $\bar{D} < 1$, the turbulent striation model* (5.3), (5.4) *linearized about the above-defined stationary states is stable.*

(ii) *If ($\sigma = 1$ and $\bar{D} > 1$) or $\sigma = -1$, the turbulent striation model is unstable for wave-vectors lying in the instability domain $\mathcal{I}$. $\mathcal{I}$ is bounded and contained in the ball centered at the origin and of radius $(1 - \sigma\bar{D}^{-1})^{1/2}$.*

We note this strange feature that adding too large a diffusion can destabilize the striation model in the case where the unperturbed striation model is stable (case $\sigma = 1$ and $\bar{D} > 1$).

Thanks to this stability result, we can return to the problem of finding the value for D. Suppose that we know (from experimental observations, for instance) that no structures finer than a certain scale ℓ can persist. This means that all perturbations with a wavelength less than ℓ are stable (i.e., are damped by dissipation), or equivalently, that all wave-vectors ξ larger than $1/\ell$ lie in the stability domain. To ensure this property, it is enough to have $1/\ell > (1 + \bar{D}^{-1})^{1/2}$. (We take $\sigma = -1$ because, in practice, there are always regions where the density gradient and the neutral wind have configurations which trigger the instability; see, e.g., the numerical results in section 6.) This condition translates into $\bar{D} \geq \ell^2(1 - \ell^2)^{-1}$. In practice, it is legitimate to assume that $\ell \ll 1$ (because the typical size of the ultimate permanent structures is far smaller than the typical size of the observation domain). Going back to the unscaled value of the diffusion constant D, we get

$$D \gtrapprox \ell^2\lambda|U|\,. \tag{5.6}$$

In the next section, we present numerical simulations which display the relation between the typical size of the persisting structures triggered by the instability and the value of this diffusion coefficient.

6. Numerical experiments. In this section, we present some numerical simulations of the striation model (2.9), (2.12) and of the turbulent striation model (5.3), (5.4). The elliptic equation (2.12) or (5.4) is discretized by a conservative finite difference method. The plasma velocity is computed by means of finite differences applied to the second equation of (2.9) or (5.3) on staggered grids. The transport equation (first equation of (2.9) or (5.3)) is discretized thanks to a classical TVD-scheme [17], [42], [30]. In order to deal with steep density gradients, the diffusion operator in (5.3) is implicitly discretized, and we make use of a Strang splitting for the overall time discretization of this equation. A preconditioned gradient method [40] is applied to solve the linear systems resulting from the discretization of the elliptic equation (2.12) or (5.4) and from the implicit discretization of the diffusion equation (5.3).

Our first test problem is intended to mimic that of [44]. The initial density is a random perturbation of a uniform density in the x_1-direction with a Gaussian profile

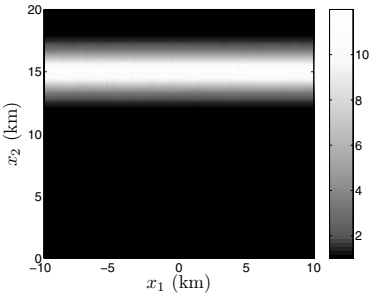
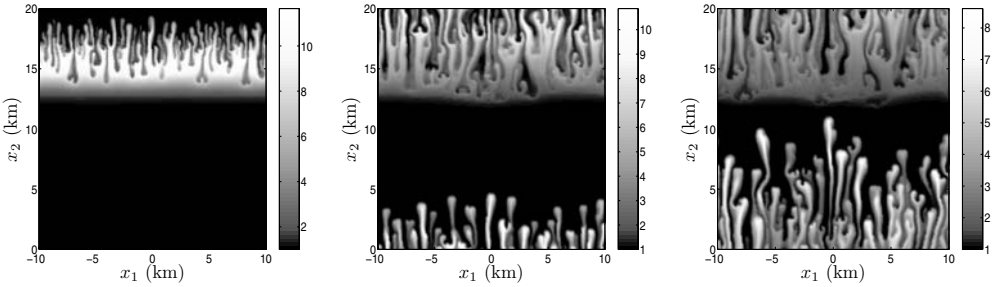


FIG. 6.1. *Initial plasma density (m^{-3}).*

TABLE 6.1
Number of cells and mesh sizes.

	Nb. of cells	Δ_x, Δ_y (m)
Mesh 1	200×200	$0.1 \quad 10^3$
Mesh 2	400×400	$0.05 \quad 10^3$
Mesh 3	800×800	$0.025 \quad 10^3$

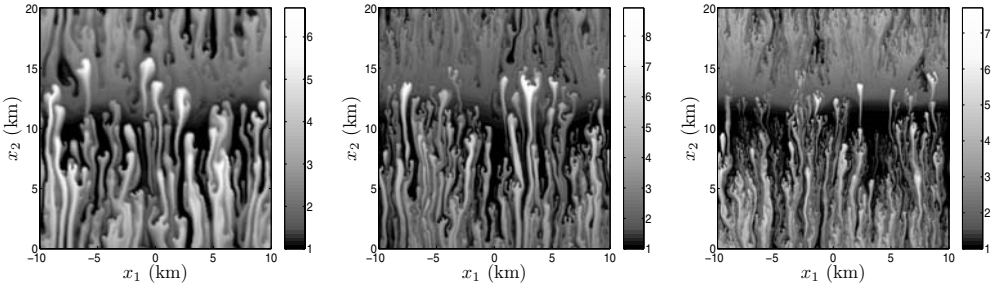


(a) t = 196 s. (b) t = 392 s. (c) t = 588 s.

FIG. 6.2. *Plasma density at various times given by the striation model discretized on mesh* 1.

in the x_2-direction (cf. Figure 6.1). The neutral wind u_n is directed along the x_2-axis and has a value of 45 ms^{-1}. Different mesh sizes listed in Table 6.1 are considered. When the turbulent striation model is considered, the diffusion length ℓ (which sets the value of the diffusion coefficient through (5.6)) is equal to 0.1×10^3 m (a scale resolved by all mesh steps). In practice, its value should be prescribed by comparing with experimental measurements (see, for instance, [8]). However, our purpose here is towards qualitative rather than quantitative results.

We first consider the original striation model (2.9), (2.12). Figure 6.2 displays the time evolution of the plasma density as a function of the two-dimensional coordinate x. Periodic boundary conditions are used. We see that the upper side (with respect to the orientation of the figure) of the density gradient is unstable, while the lower side is stable. The instability produces finger-like structures which rise in the positive x_2-direction and eventually (by periodicity) appear as originating from the lower boundary. In Figure 6.3 we represent the plasma density computed on the different meshes (see Table 6.1) at time $t = 804$ s. The mesh-size is divided by a factor 2 from panel (a) to panel (b) and from panel (b) to panel (c). One can notice that the



(a) Mesh 1 (200 × 200). (b) Mesh 2 (400 × 400). (c) Mesh 3 (800 × 800).

FIG. 6.3. *Plasma density at $t = 804$ s given by the striation model discretized on the three different meshes of Table* 6.1.

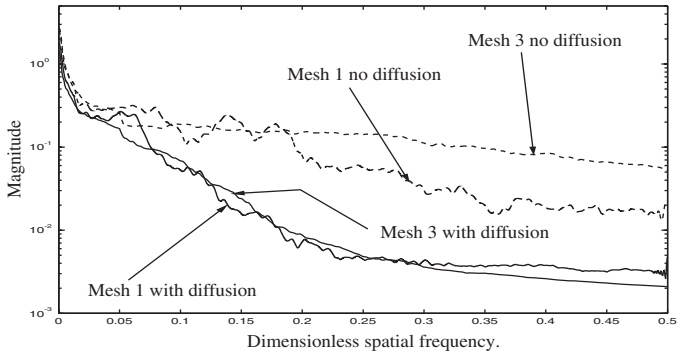
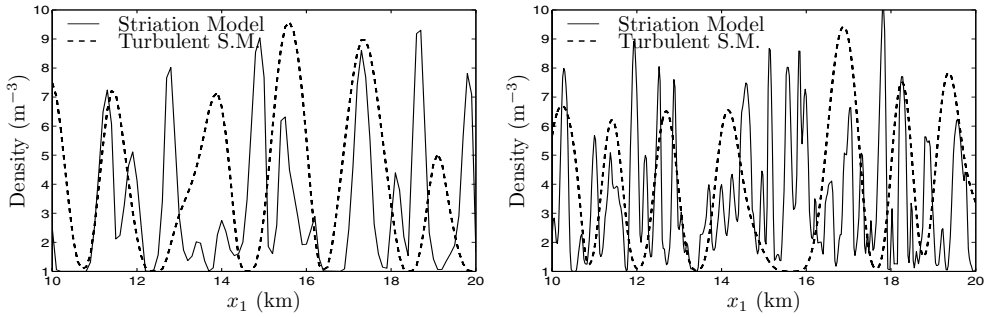


FIG. 6.4. *Spectral density of the plasma density n computed on the coarsest and the finest meshes of Table* 6.1 *(mode magnitude versus dimensionless spatial frequency). The spatial frequency scale is set to ℓ^{-1}, where ℓ, given by* (5.6)*, is equal to* $0.1\,10^3$ *m.*

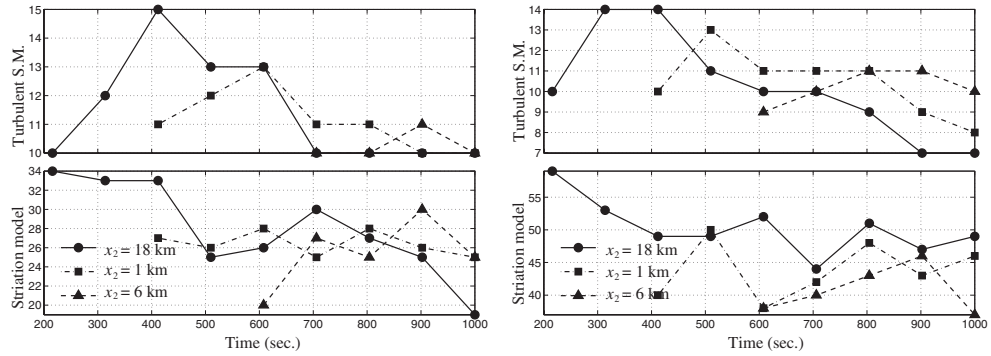
number of persisting structures grows with the number of cells, while their typical size decreases with the mesh-size. This remark can be made more quantitative as in Figure 6.4, where the spectral density (i.e., the modulus of the Fourier transform) of the plasma density is displayed for the coarsest and finest meshes (dashed curves). We can see that high frequency modes (corresponding to space scales ranging from 2 to 5 times the value of ℓ) have a significantly larger contribution when the finest mesh is used. Correspondingly, in Figure 6.3, we notice that the structures are all the tinier as the mesh becomes finer.

This behavior can be related to the instability of the model. Indeed, numerical diffusion is the only damping mechanism, and the numerical diffusivity is proportional to the mesh size [17], [42], [30]. According to the stability analysis in section 5.2, the diffusive striation model becomes stable for wave-vectors of the order of $1/\sqrt{D}$, which is proportional to $1/\sqrt{\Delta x}$. Therefore, the size of the typical persisting structures must be divided by a factor $\sqrt{2}$ each time the mesh-size is divided by 2. This is, roughly speaking, what we observe in Figure 6.5, where cuts of the density along lines $x_2 =$ constant are plotted. The calculation carried out on the coarsest mesh (solid line of Figure 6.5(a)) exhibits five to six main structures (areas where the density varies significantly) in the last quarter of the x_1 range. In the same interval approximately

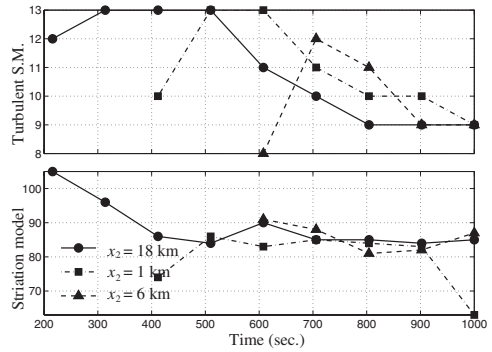


(a) Mesh 1 (second half of the x_1 range). (b) Mesh 3 (second half of the x_1 range).

FIG. 6.5. *Plasma density profiles along the line $x_2 = 2$ km at time $t = 392$ s.*



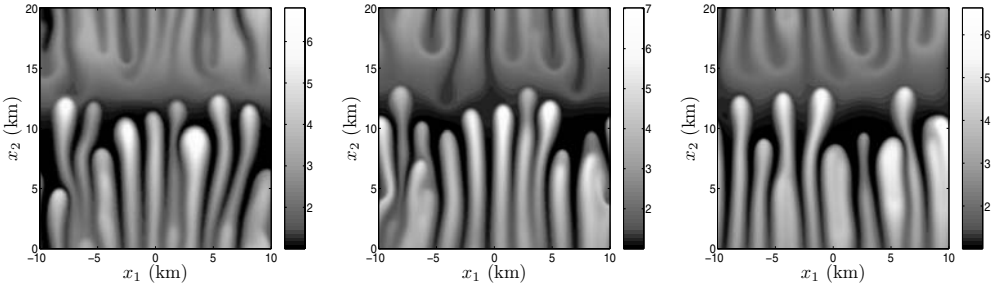
(a) Mesh 1 (200 × 200). (b) Mesh 2 (400 × 400).



(c) Mesh 3 (800 × 800).

FIG. 6.6. *Number of local maxima along the lines $x_2 = 18, 1, 6$ km (circles, squares, and triangles, respectively) as a function of time for the turbulent striation model* (a), (b) *and the classical one* (c) *computed on the three meshes of Table* 6.1. *The number of local maxima is computed as half the number of sign changes in the density derivative.*

14 main structures are counted for the density profile computed with the finest grid (solid line of Figure 6.5(b)). Note that small patterns can exist in addition to the persisting structures. Indeed there are seven and more than twenty local maxima for meshes 1 and 3, respectively. The same ratio is observed in Figure 6.6, where the



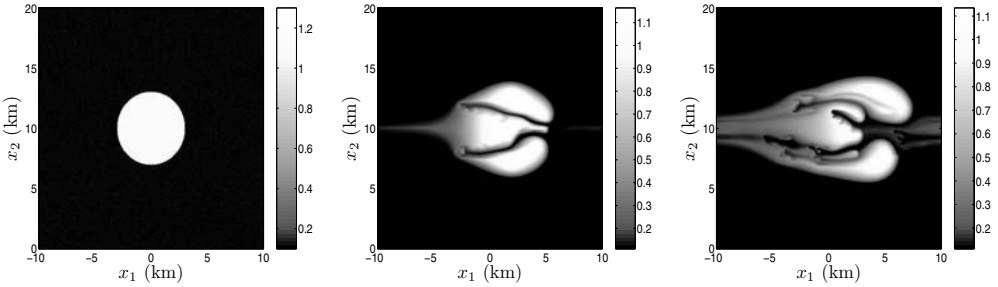
(a) Mesh 1 (200 × 200). (b) Mesh 2 (400 × 400). (c) Mesh 3 (800 × 800).

FIG. 6.7. *Plasma density at* $t = 804$ *s, given by the turbulent striation model discretized on the three different meshes of Table* 6.1.

time evolution of the number of local maxima is displayed. For long time evolution (greater than 600 (s.)) the number of local maxima can be estimated as 25 for the coarsest mesh and 85 for the finest one. These results confirm the mesh-dependence of the simulations.

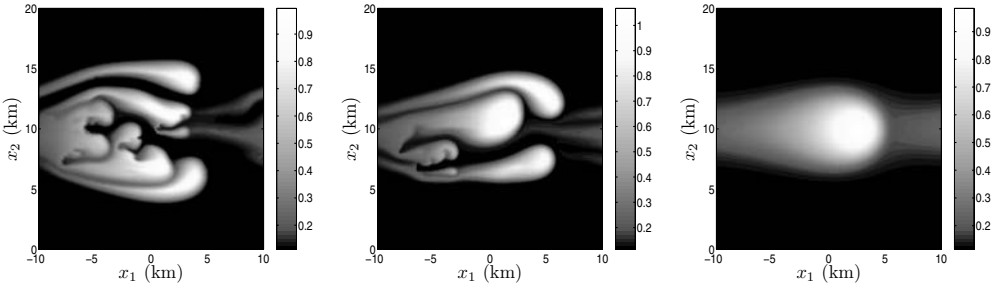
We next consider the turbulent striation model (5.3), (5.4). Figure 6.7 demonstrates the stability brought by the diffusion: the number and size of the structures remain almost the same when the mesh resolution increases. The dashed curves of Figure 6.5 display cuts of the density on a line $x_2 =$ constant for the coarsest and finest meshes, respectively. The small patterns which could be observed on the results computed with the classical striation model (solid curves) have disappeared. Moreover, the number of local maxima (six for the coarsest mesh, seven for the finest one) observed in Figure 6.5 are now quite independent of the grid resolution. This invariance of the number of local maxima with respect to the grid resolution can be observed in the course of the time evolution (see Figure 6.6). Indeed, this number remains almost constant in time, and equal to 9 when the mesh-size varies. These results therefore show a significant difference between the turbulent striation model and the original one. The spectral densities computed with the turbulent striation model are displayed in Figure 6.4. The diffusion damps out the high frequency modes, and the curves computed with the two different mesh-sizes are very similar, whatever space scales are considered, in contrast with the behavior of the original striation model. Note that the characteristic size of the striations observed in Figures 6.7 and 6.5 amounts to a few kilometers, which fits well with experimental observations.

The second simulation is aimed at illustrating the results of the stability analysis developed in the discontinuous density profile framework (see section 3.3). To this purpose, we consider a set of simulation parameters similar to those above, except for the initial density and the neutral wind. The initial density consists of a plasma bubble (density equal to one) in a quasi-vacuum medium (very small density). This initial data is perturbed by a random noise. The neutral wind is oriented along the x_1-axis; its speed is set to 100 $\mathrm{m} \cdot \mathrm{s}^{-1}$. Simulations performed on mesh 2 (Table 6.1) with the classical striation model are displayed in Figure 6.8(a), 6.8(b), and 6.8(c), respectively, at time t = 0, 281.4, and 562.8 seconds. The plasma bubble is set into motion by the neutral wind, and since periodic boundary conditions are used, the bubble seems to go out of the domain through the right boundary of the frame and to re-enter the computational box through the left boundary. The instability develops



(a) Initial density. (b) t = 281.4 s. (c) t = 562.8 s.

FIG. 6.8. *Plasma density at various times given by the classical striation model discretized on mesh* 2 *in the case of a discontinuous initial density.*



(a) Striation model on mesh 1. (b) Turbulent striation model with low diffusion. (c) Turbulent striation model with high diffusion.

FIG. 6.9. *Plasma density at* $t = 562.8$ *s given by the classical striation model on mesh* 1 (a)*, the turbulent striation model with a moderate diffusion on mesh* 2 (b)*, and the turbulent striation model with a large diffusion on mesh* 2 (c)*, with the discontinuous initial density of Figure* 6.8(a)*.*

along the right edge of the bubble, the other edge being unaffected.

The same simulation run on mesh 1 produces the result displayed on Figure 6.9(a) and shows the sensitivity of the instability pattern with respect to the grid resolution. The last two panels of the figure, (b) and (c) show calculations performed with the turbulent striation model on mesh 2. The diffusion parameter used in Figure 6.9(c) is four times as big as the one considered for 6.9(b). When comparing Figures 6.8(c) (without any diffusion) and 6.9(b), we get the same conclusion as before: diffusion brings stability for small space scales, since the tiniest patterns have disappeared from Figure 6.9(b). More diffusion can also bring stability for all space scales and prevent the growth of the instability, as demonstrated by the results of Figure 6.9(c).

7. Conclusion. In this paper, we have been concerned with the modeling of ionospheric plasma instabilities. The first main point of this work was to remark that the "striation model" allows for discontinuous solutions, and that discontinuous steady states may be unstable in a similar way as are smooth ones. The second point was to propose that the turbulence induced by the instability may actually produce diffusion, in a way similar to what occurs in fluid mechanics, and that this diffusion may actually contribute to stabilizing large wave-number perturbations. Following the statistical

approach to turbulence, we have derived and analyzed a "turbulent striation model." Numerical simulations have been produced in support of our analysis.

REFERENCES

[1] W. G. Baker and D. F. Martyn, *Electric currents in the ionosphere,* I. *The conductivity*, Phil. Trans. Roy. Soc. London, A246 (1953), pp. 295–305.

[2] J. J. Berthelier, *L'ionosphère*, in Environnement spatial: Prévention des risques liés aux phénomènes de charge, J. P. Catani and M. Romero, eds., Cépaduès éditions, Toulouse, 1996.

[3] C. Besse, J. Claudel, P. Degond, F. Deluzet, G. Gallice, and C. Tessieras, *A model hierarchy for ionospheric plasma modeling*, Math. Models Methods Appl. Sci., 14 (2004), pp. 393–415.

[4] C. Besse, J. Claudel, P. Degond, F. Deluzet, G. Gallice, and C. Tessieras, *Numerical simulations of the ionospheric dynamo model in a nonuniform magnetic field*, J. Comput. Phys., submitted.

[5] C. Besse, J. Claudel, P. Degond, F. Deluzet, G. Gallice, and C. Tessieras, *Ionospheric plasmas: Model derivation, stability analysis, and numerical simulations*, in Numerical Method for Hyperbolic and Kinetic Problems, Th. Goudon and E. Sonnendrücker, eds., de Gruyter, New York, Berlin, to appear.

[6] C. Besse, P. Degond, H-J. Hwang, and R. Poncet, *Nonlinear instability of the two-dimensional striation model about smooth steady states*, Comm. Partial Differential Equations, to appear.

[7] D. Bilitza, *International reference ionosphere* 2000, Radio Science, 36 (2001), pp. 261–275.

[8] E. Blanc, B. Mercandelli, and E. Houngninou, *Kilometric irregularities in E and F regions of the daytime equatorial ionosphere observed by a high resolution HF radar*, Geophys. Res. Lett., 23 (1996), pp. 645–648.

[9] J. W. Chamberlain and D. W. Hunter, *Theory of Planetary Atmospheres*, Academic Press, New York, 1987.

[10] S. Chandrasekhar, *Hydrodynamic and Hydromagnetic Stability*, Dover, New York, 1981.

[11] B. Desjardins and E. Grenier, *Linear instability implies nonlinear instability for various types of viscous boundary layers*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 87–106.

[12] J. H. Doles, III, N. J. Zabuski, and F. W. Perkins, *Deformation and striation of plasma clouds in the ionosphere.* 3. *Numerical simulations of a multilevel model with recombination chemistry*, J. Geophys. Res., 81 (1976), pp. 5987–6004.

[13] D. T. Farley, *Theory of equatorial eletrojet plasma waves, new developments and current status*, J. Atmospheric Terrestrial Phys., 47 (1985), pp. 729–744.

[14] B. G. Fejer and M. C. Kelley, *Ionospheric irregularities*, Rev. Geophys. Space Phys., 18 (1980), pp. 401–454.

[15] X. Garbet, *Instabilités, Turbulence et Transport dans un Plasma Magnétisé*, Habilitation dissertation, University of Marseille, France, 2001.

[16] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier Stokes Equations, Theory and Algorithms*, Springer, New York, 1986.

[17] E. Godlewski and P. A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Springer, New York, 1996.

[18] E. Grenier, *On the nonlinear instability of Euler and Prandtl equations*, Comm. Pure Appl. Math., 53 (2000), pp. 1067–1091.

[19] C. Grimault, *Caractérisation des canaux de propagation satellite-Terre SHF et EHF en présence de plasma post-nucléaire*, Ph.D. dissertation, University of Rennes, France, 1995.

[20] Y. Guo and W. Strauss, *Nonlinear instability of double-humped equilibria*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 12 (1995), pp. 339–352.

[21] J. K. Hargreaves, *The Solar-Terrestrial Environment*, Cambridge University Press, Cambridge, UK, 1992.

[22] A. M. Hamza and J. P. St. Maurice, *A fully self-consistent fluid theory of anomalous transport in Farley–Buneman turbulence*, J. Geophys. Res., 100 (1995), pp. 9653–9668.

[23] A. M. Hamza and J. P. St. Maurice, *A turbulent theoretical framework for the study of current-driven E region irregularities at high latitudes: Basic derivation and application to gradient-free situations*, J. Geophys. Res., 98 (1993), pp. 11587–11599.

[24] H. J. Hwang and Y. Guo, *On the dynamical Rayleigh–Taylor instability*, Arch. Ration. Mech. Anal., 167 (2003), pp. 235–253.

[25] M. J. Keskinen, *Nonlinear theory of the $E \times B$ instability with an inhomogeneous electric field*, J. Geophys. Res., 89 (1984), pp. 3913–3920.
[26] M. J. Keskinen, S. L. Ossakow, and B. G. Fejer, *Three-dimensional nonlinear evolution of equatorial ionospheric spread-F bubbles*, Geophys. Res. Lett., 30 (2003), pp. 1855–1858.
[27] H. Kesten and G. C. Papanicolaou, *A limit theorem for stochastic acceleration*, Comm. Math. Phys., 78 (1980), pp. 19–63.
[28] F. Krause and K. H. Rädler, *Mean-field magnetohydrodynamics and dynamo theory*, Pergamon Press, Elmsford, NY, 1980.
[29] M. Lesieur, *Turbulence in Fluids: Stochastic and Numerical Modeling*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.
[30] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhaüser-Verlag, Basel, Switzerland, 1992.
[31] S. Matsushita, *On artificial geomagnetic and ionospheric storms associated with high-altitude explosions*, J. Geophys. Res., 64 (1959), pp. 1149–1161.
[32] B. Mohammadi and O. Pironneau, *Analysis of the K-Epsilon Turbulence Model*, Wiley, New York, 1994.
[33] M. Oppenheim, N. Otani, and C. Ronchi, *Saturation of the Farley–Buneman instability via nonlinear $E \times B$ drifts*, J. Geophys. Res., 101 (1996), pp. 17273–17286.
[34] S. L. Ossakow and P. K. Chatuverdi, *Morphological studies of rising equatorial spread F bubbles*, J. Geophys. Res., 83 (1978), pp. 2085–2090.
[35] F. Poupaud and A. Vasseur, *Classical and quantum transport in random media*, J. Math. Pures Appl., 82 (2003), pp. 711–748
[36] G. C. Reid, *The formation of small-scale irregularities in the ionosphere*, J. Geophys. Res., 73 (1968), pp. 1627–1640.
[37] H. Rishbeth and O. K. Garriott, *Introduction to Ionospheric Physics*, Academic Press, New York, 1969.
[38] C. Ronchi, R. N. Sudan, and D. T. Farley, *Numerical simulations of large-scale plasma turbulence in the daytime equatorial electrojet*, J. Geophys. Res., 96 (1991), pp. 21263–21279.
[39] C. Ronchi, R. N. Sudan, and P. L. Similon, *Effect of short-scale turbulence on kilometer wavelength irregularities in the equatorial electrojet*, J. Geophys. Res., 95 (1990), pp. 189–200.
[40] Y. Saad, *SPARSKIT: A Basic Tool Kit for Sparse Matrix Computations—Version* 2, Technical Report, Computer Science Department, University of Minnesota, Minneapolis, MN, 1994.
[41] R. N. Sudan, J. Akinrimisi, and D. T. Farley, *Generation of small scale irregularities in the equatorial electrojet*, J. Geophys. Res., 78 (1973), pp. 240–248.
[42] E. F. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamics*, Springer, New York, 1999.
[43] W. White, *An overview of high-altitude nuclear weapons phenomena*, Heart conference short course, preprint, 1998.
[44] S. T. Zalesak and S. L. Ossakow, *Nonlinear equatorial spread F: The effect of neutral winds and background Pedersen conductivity*, J. Geophys. Res., 87 (1982), pp. 151–166.